Collaborative corpus building for minorized languages using wiki-technology.

Documenting the Asturian language

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

Roser Saurí
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 minorized 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. That will benefit both Eslema, allowing to enlarge 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:**
  In Spain:
  - Most of the Principality of Asturies
  - The provinces of Lleón and Zamora

  In Portugal:
  - The district of Miranda do Douro
- **Asturian is also known as:**
  - Asturian-Leonese
  - Asturleonese
  - Bable
  - Mirandese (only in Portugal)
- **Speakers number** (estimated):
  300,000 speakers
  (over a population of 1,000,000 people)

Language health:

- Generational transmission is seriously threatened.
- Tendency towards losing Asturian competence among young generations.
  Asturian at (public) schools: (year 2002) primary 86% (15,227 students)
  secondary 20% (2,171 students)
  However, Asturian is conceived as a marginal subject, often taught at inconvenient times, and forcing students to choose between Asturian and other essential subjects (e.g., English, computer programming).
- 20% loss of native speakers, during last decade (Llera Ramo, 2002).

Legal Status:

- Co-official only in **Portuguese** area, since 1998.
- In **Asturias**: The Asturias Autonomy Statute (1981) recognizes it as a language. However, it does not bestow oficial status upon it, thus not recognizing its speakers rights.
- Its official status in the reformation of the Autonomy Statute in 1998 was blocked by the mainstream parties.
- Local politicians enacted the ``Law for the protection of Asturian'', which includes some linguistic rights (e.g., street signs and administrative documentation in Asturian). So far, this law has not been significantly applied.

Eslema. A corpus for the Asturian language

- Eslema is the first corpus created for Asturian (**http://www.uniovi.es/eslema**)
- Its goals are:
  1. Documenting the linguistic tradition of Asturian and its historic evolution.
  2. Contributing to the processes of codification and normalization of this language, and therefore aiming at reversing the current diglossic situation.
- It has been conceived as a representative corpus, hence covering texts of a varied typology in terms of: domain and genre, controlled vs. spontaneous production, written vs. oral, etc.
- It consists of 3 subcorpora, reflecting the main historical periods in the evolution of Asturian. Each of them is defined according different parameters:

	Medieval Astur-Leonese	Classical Literature	Present-Day Asturian	
Period	13th-15th centuries	from 1639 to 1950	from 1970 onwards	
			Oral	Written
Size	200 docs	200 docs	60 sessions	1700 docs
Linguistic Coverage	Legal and administrative text	Literary texts (mainly poetry)	• Formal and controlled language – radio programs – news reports – etc. • Colloquial language	All kinds of genres: • Literary • Journalistic • Expert domain – Scientific – Administrative
Issues	• Selection criteria: border between Latin and Romance texts. • Hybridation of Asturian wrt Castilian	• Selection criteria: filtering out texts of minor quality: – Poor Asturian – Often times from Spanish	• Bottleneck in digitalization and transcription	• Bottleneck in linguistic processing

- 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.

Using wiki-technology for collaboratively building Eslema

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:
  – (2003) 21.38%
  – (2006) 41.4% (Spanish average: 39%)

Favorable social conditions

- As speakers of a minority language, they feel fairly commited 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: [data from Feb, 2009]
  – #78, in terms of #articles (12,092 arts.)
  – #90 in terms of #contributors (3,717 contributors)

Wiki as friendly technology

- A website focusing on fast co-construction and sharing of free-form textual documents (wikipages)
- Philosophy: together we make it better than individually
- A simple interface and markup language enables experts and non tech-savvy 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 Alterman, 2008).
- So far, WDP-based wikis have been deployed to support various collaborative learning activities. Hence, Eslema explores its capabilities within a research-oriented setting.

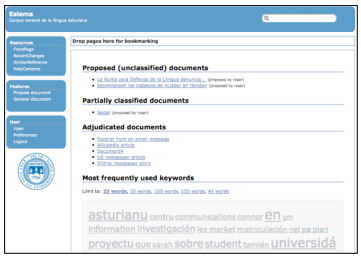

Fig. 1: Home

Fig. 2: Document discussion

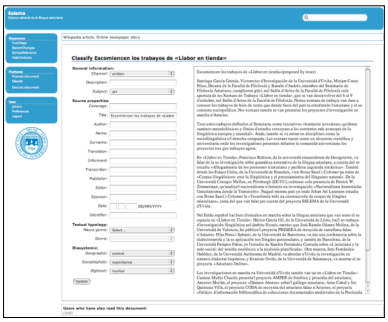

Fig. 3: Document classification

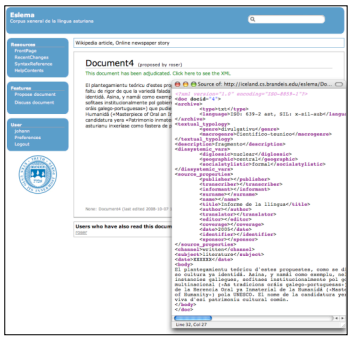

Fig. 4: Classified document

The Eslema-Wiki

Adequate for:

1. Introducing new documents. In some cases, they can be of great value; e.g.,
   – Personal letters
   – Family documentation
   – Old publications of local domain
2. Classifying documents; i.e., assigning metadata tags.

Main functionalities:

1. **Home page** (fig. 1), which provides:
   - A list of documents and their status (central area of the page), which can be:
     – Pending classification.
     – Partly classified: one or more contributors have already worked on them.
     – Adjudicated: once there is enough agreement among contributors, an Eslema researcher aproves and stores the document into the DB.
   - A word cloud (lower area of the page), pointing to the most frequent words.
   - A document dock (below the banner), where documents in progress can be bookmarked.
2. **Discussion page** (fig. 2), promoting collaboration among participants.
3. **Doc classification page** (fig. 3), providing a general template to create the metadata for each document.
4. **Classified document** (fig. 4), with its metadata encoded in XML.

References

Larusson, J. and R. Alterman (2008) Supporting and Tracking Collective Cognition in Wikis. In *Proceedings of ICLS 2008: International Conference for the Learning Sciences*: Vol. 3: 330-337. The International Society of the Learning Sciences.

Llera Ramo, F. (2002) II Estudiu sociollingüístico d'Asturies. Avance. In *Lletres Asturianes*, 89, 181-197.

Viejo, X., R. Saurí and A. Neira (2008) Eslema. Towards a corpus for Asturian. *Collaboration: Interoperability between people in the creation of language resources for less-resourced languages. A SALTMIL workshop*. LREC 2008. Marrakesh.