

Eslema. Towards a Corpus for Asturian

Xulio Viejo[‡], Roser Saurí*, Ángel Neira[†]

[‡]Departamento de Filología Española

[†]Computer Science Department

Universidad de Oviedo

{jviejo, neira}@uniovi.es

*Computer Science Department

Brandeis University

rosier@cs.brandeis.edu

Abstract

We present *Eslema*, the first project devoted to building a corpus for Asturian, which is carried out at Oviedo University. *Eslema* receives minor funding from the Spanish government, which is fundamental for basic issues such as equipment acquisition. However, it is insufficient for hiring researchers for a reasonable period of time. The scarcity of funding prompted us to look for much needed resources in entities with no institutional relation to the project, such as publishing companies and radio stations. In addition, we have started collaborations with external research groups. We are for example initiating a project devoted to developing a wiki-based platform, to be used by the community of Asturian speakers, for loading and annotating texts in *Eslema*. That will benefit both our project, allowing to enlarge the corpus at a minimum cost, and the Asturian community, causing a stronger presence of Asturian in information technologies 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.

1. Introduction

We present *Eslema*, the first project devoted to building a corpus for Asturian, which is carried out by the Research Group on Asturian Philology (*Seminariu de Filoloxía Asturiana*) within the Spanish Philology Department at Oviedo University (<http://www.uniovi.es/eslema>). Its goal is compiling a corpus representative of this Romance language, hence containing documents of a varied typology, including both written and oral discourse, and texts from the different historical periods of the Asturian language, starting from the medieval ages.

The name of the project, *Eslema*, is an Asturian word not documented in any dictionary but in a collection of oral Asturian folk tales (Suárez López, 2000). *Eslema* refers to an impression drawn in a stone indicating the existence of a hidden treasure. Aside from its obvious poetic force, this word also expresses the great value that a complete collection of documents, adequately structured and organized, has for research in linguistic-related disciplines.

The paper is organized as follows: The next section offers a small introduction to Asturian and its current situation, and section 3 presents the project in some detail. Then, section 4 comments on the resources (both material and human) that make *Eslema* possible. And finally, section 5 touches upon the social contribution of the project.

2. The Asturian language

Asturian (or, equivalently, Asturian-Leonese, Asturleonese, or Bable) is the autochthonous language of most of the territory in the Principality of Asturias, the provinces of Leon and Zamora, in Spain, and the district of Miranda do Douro (Portugal). Nowadays it is most actively used in the Asturian territory proper and in the north of Portugal, where the local version, Mirandese, achieved co-official status in

1998. The most reliable estimates of the status and vitality of Asturian nowadays calculate the community of speakers corresponds to approximately a third of the population. That is, there are about 300,000 speakers of Asturian in a community of 1,000,000 people, including about 50,000 speakers of Galego-Asturian concentrated in the westernmost part of Asturias. These figures bode ill for the future of the language since Asturian competence is notably reduced among young people, in such a way that the generational transmission of the language is seriously threatened. In fact, there has been a 20% loss of native speakers during the last decade (Llera Ramo, 2002).

The legal status of Asturian is very different from that of other languages in Spain, or other European minority languages such as Romansh, Irish, and Frisian. The Asturias Autonomy Statute (1981) recognizes the Asturian linguistic specificity in its fourth article, and establishes that “Bable will enjoy protection. Its use, diffusion in the mass media, and teaching will be promoted, respecting in every case the local variants and the voluntary nature of its learning.” However, it does not bestow official status upon Asturian, or in other words, it does not recognize the civil rights of its speakers as such. Increasing social demand for linguistic normalization triggered an intense political debate about the inclusion of the official status of the language in the reformation of the Autonomy Statute in 1998 (blocked by the mainstream parties). It also led local politicians to enact the so-called “Law for the protection of Bable/Asturian”, which explicitly includes some linguistic rights (e.g., street signs in Asturian, publication of administrative documentation and official advertising brochures in Asturian, etc.) and was unanimously approved by all political parties. So far, however, this law’s normative intentions have not been significantly developed.

One of the main linguistic demands since the 1970s has been the possibility of teaching Asturian at schools. So socially deep-rooted is that demand that, as a matter of fact, it is specifically contained in the fourth article of the Autonomy Statute, as mentioned above. The first school-teaching initiatives took place in 1984. Nowadays, the teaching of Asturian has attained a remarkable expansion in primary schools (86% of the public schools, with 15,227 students in 2002), and a much more modest one in secondary school (20% of the institutions, with 2,171 students in that same year). Nonetheless, Asturian is conceived as a marginal subject and not as part of the core school curriculum, and it is often taught at inconvenient or unpopular times. In fact, its optional status implies that students frequently have to choose between studying Asturian or other subjects essential for their background, such as a foreign second language or computer programming. Viejo Fernández (2004) offers a detailed view of Asturian socio-linguistic situation.

Currently, Asturias is undergoing a new reformation of its statute, during which the issue of the Asturian legal status will probably be debated again. In spite of some slight advancements concerning the normalization of Asturian in the past few years, its officialization still appears to be a remote possibility given the opposing stance of the two main political parties. This situation is indeed of absolute relevance to Eslema. Not having a satisfactory political and legal framework seriously restricts the possibilities of an adequate linguistic normalization and, consequently, of any potential research on Asturian aiming at effective and practical outcomes. On the other hand, this precarious situation makes it even more necessary developing new strategies for ensuring the use of Asturian in everyday situations, as well as for boosting its prestige as a modern language, capable of coping with the technology-based communication needs of present-day society.

3. Eslema. A General Corpus for Asturian

3.1. Project goal

Eslema was conceived as a framework to develop several subcorpora—hence the title General Corpus for the Asturian Language (*Corpus xeneral de la llingua asturiana*), which is part of the project's official name. Our initial goal is the construction of three of these subcorpora: the Corpus of Medieval Asturian-Leonese (including documents from the 13th-15th centuries), the Corpus of Classical Literature (between 1639 and 1950), and the Corpus of Present-day Asturian.

Being the corpus of a minorized language, the main objective of Eslema is helping set the foundation for fully normalizing Asturian as the language of use in any possible social context. At first impression, medieval and classical Asturian texts do not appear to be able to contribute to this goal, even less so considering the remarkable differences that exist between the historical and the current language (Viejo Fernández, 2003). We nevertheless decided to compile language samples from those periods. Aside from the fact that they are inherently interesting for linguistic and philological research in general, this decision follows two strategic reasons and thus aspires to have some degree of normalizing effect as well.

Asturian is generally seen through a Castilian perspective and considered a mere dialect, an approach that undervalues, or simply ignores, the long and solid written tradition that exists in that language. Sometimes its historical texts have been presented (with scarce scientific rigor) as either samples of "Old Spanish" or "dialectal literature" epiphenomena. As a matter of fact, most historical corpora for Spanish include medieval Asturian-Leonese texts, adhering to no linguistic criteria other than a unitarizing discourse of clear ideological and nationalistic objective, often times fostered by the Spanish philological tradition itself. In that sense, documenting in a coherent way the linguistic tradition of Asturian and its corresponding historic evolution is a priority issue.

There is yet an additional reason for compiling a historical corpus. Current Asturian, mainly relegated to oral linguistic exchanges, is affected by the strong pressure of Spanish, with which it tends to blend in an ongoing process of dissolution. Amidst this adverse situation, Asturian is currently undergoing the codification process that healthier languages (e.g., French, Spanish, English) experienced at an earlier time, mostly during the Renaissance. And for the success of that process, the existence of historical text has proven to be crucial. As it happens, oftentimes establishing the autochthonous vocabulary is approached from the mere intuition of intellectuals and writers, leading to unnatural or unpopular solutions which end up being rejected by the speakers. Therefore, having historical text dating previous to the Castilianization suffered in the past half century will substantially help codify and stabilize the language. What is more, medieval and classical Asturian texts provide consolidated models of specific domains and their corresponding terminology, such as administrative and juridic registers, which are fundamental when normalizing a language used mainly in oral contexts.

3.2. Subcorpora

3.2.1. Corpus of Medieval Asturian-Leonese

It contains several hundreds of texts in medieval Asturian and Leonese (two dialectal variations of the same language), all of them of legal typology. These documents are not available online yet, given that they are still pending annotation. Currently, we are working on the tag set design.

Aside from specific issues related to medieval legal language (e.g., what is an adequate textual typology, etc.), we need to address two additional problems. First, determining what kind of texts we consider as representative of the medieval Asturian-Leonese language; that is, setting a border between Latin and Romance, an issue that affects a considerable number of documents that date back to a time earlier than 1260. And second, addressing the process of hybridization of Asturian with respect to Spanish, which took place from the end of the 14th century and throughout the whole 15th century.

3.2.2. Corpus of Classical Literature

It is a wide collection of literary texts (aside from few exceptions) belonging to the genre of poetry. These are selected according to criteria concerning their linguistic qual-

ity given that, together with pieces of work of absolute linguistic quality, since the 19th century there exists a profusion of folk texts of minor quality, characterized by the use of very poor Asturian and often times written from Castilian hence, just imitative of the Asturian voice. We considered that this type of texts is not representative of the actual Asturian used at that period. However, determining the samples that are truly representative based on criteria that go beyond our mere subjective intuitions as speakers and philologists, is a challenging task. The most representative part of this corpus is already in digital format, although it is not yet annotated and cannot be accessed online.

3.2.3. Corpus of Present-Day Asturian

This corpus collects documents in Asturian that have been produced during the last three decades. Currently, it is the most visible part of the project, although some great amount of it is still under development.

The Corpus of Present-Day Asturian consists of two subcorpora, a written and an oral one, each containing documents of a varied nature. The documents in the written subcorpus are available in text format. As for those in the oral subcorpus, we plan to have them available in audio format and accompanied with their corresponding transcription.

Documents in both subcorpora are of different types. There are mainly literary texts (from all genres), although we also have a significant sample of journalistic language (news reports, op-eds, etc.) and, to a lesser extent, of expert domain language (i.e., scientific, technical, administrative, etc.). Similarly, we aim at gathering a balanced sample of documents covering different geographical and social dialects although most of the text collected so far is in normative central Asturian.

Written documents. At the linguistic level, we are currently working on the annotation at the different relevant layers –mainly, morphology, part-of-speech, and basic syntax. This work is becoming particularly slow, given that we are unable to hire professional computer scientists (or computational linguists) who can work full time in the project. The only advancement in this area has been done thanks to the collaboration of computer science students. However, we foresee that we will be able to report initial results along this direction within the few coming months.

Oral documents. At present, Eslema has a wide collection of audio files, which will be available online momentarily. Some of them are representative of a more formal (or controlled) use of Asturian –basically, radio programs, broadcasting of sport events, news reports, etc. Some others, on the other hand, represent the colloquial (and spontaneous) use of Asturian. The progress made for this part of the corpus is slow mainly due to two reasons. First, the need for digitalizing the original recordings, which are in cassette support (they started being recorded in the 80s). And second, the fact that our goal is not only to make this material available online in audio format, but to provide their corresponding transcription as well.

In addition to these materials, which were generated previously to (and independently from) the beginning of Eslema, we are currently working on compiling further new

oral documents, mainly dialogue, by means of surveys designed to that purpose. Throughout this past year, we have managed to record several hours of spoken language of high quality, from different dialectal areas in Asturias.

3.3. Document annotation

3.3.1. Metadata

Each document is described at the metadata level with XML tags codifying a subset of the 15 basic elements in the Dublin Core scheme, as customized by the Open Language Archives Community (OLAC). OLAC Metadata Set extends the Dublin Core in order to express community-specific needs, namely, those concerning archiving resources that document or describe languages –e.g., dictionaries and grammars (Bird & Simons, 2001).

We decided to adopt OLAC Metadata Set in order to adhere to an international standard that can potentially facilitate the future sharing of our data with other groups. Eslema does not consist of resources describing a language, but resources that exemplify and actualize it instead. In that respect, its contents do not correspond to the type of documents targeted by OLAC. However, because Eslema's focus is on representing language as well, the set of basic elements and attributes considered by OLAC are fully adequate for describing its documents.

The following list provides the name of the OLAC elements used in Eslema to encode document metadata, with a brief description if deemed necessary:

- **Contributor.** The entity that provided the document.
- **Creator.** The person (or institution) responsible for the content of the document (e.g., author, translator).
- **Date.** The document creation time.
- **Description.** Free description of the document content.
- **Identifier.** Unambiguous reference to the resource (e.g., ISBN)
- **Language.** Asturian.
- **Publisher.**
- **Relation.** A reference to a related resource, if applicable.
- **Subject.** Keywords describing the topic of the content.
- **Title.**
- **Type.** Sound (for recordings of oral discourse) and text (for written documents).

In addition, documents are characterized according to a textual typology we have created, which classifies them according to their genre and the geographical and social dialect they represent. Through this classification, we can control the representativity and balance of the corpus, facilitate the process of information search, and contribute to future research projects concerning the Asturian language. The following distinctions are made:

- **Channel.** The original physical medium of the document contents: oral or written.
- **Textual typology.** With possible values: *colloquial*, *literary*, *journalistic*, *political*, *advertising*, *scientific-professional*, *legal-administrative*, and *ritual*.
- **Diasystemic properties.** Identifying the dialectal variety of the text.

3.3.2. Linguistic information

The encoding of linguistic information is currently one of the main bottlenecks of our project, due to the scarcity of funding and the lack of (volunteering or poorly paid) personal with an adequate formation for developing and customizing linguistic processing tools.

We are currently working on the POS and morphological annotation. So far, we have already delimited the POS tagset to be used, and are exploring what is the best solution for text tagging given existing tools for similar languages. The tagset is mostly based in the one for Spanish adopted in the FreeLing platform of linguistic analyzers (Carreras et al., 2004),¹ inherited from the EAGLES initiative for annotating morphological and syntactic information in lexicon and corpora, in all European languages.

Furthermore, the specificity of certain grammatical elements in Asturian made it necessary to minimally extend the Spanish POS tagset. For example, the gender attribute in nouns, adjectives, and determinants needs to account for the possibility of the neutral value as well.

3.4. Eslema search tool

Eslema is provided with a basic search tool, which has access to a collection of around 8,000,000 words from the written part of the Corpus of Present-Day Asturian. It allows for two different types of search, simple and complex, which so far can only be formulated in terms of the attributes encoding information at the document metadata level. This tool, however, is about to be substituted by a more efficient index-based tool, more in accord with standard approaches to search engines. In addition to the already available search features of the current tool, the new one will allow for context- and linguistic-based searches, as it is expected from any corpus query tool. Figure 1 illustrates the interface of the of the current version (for simple queries).

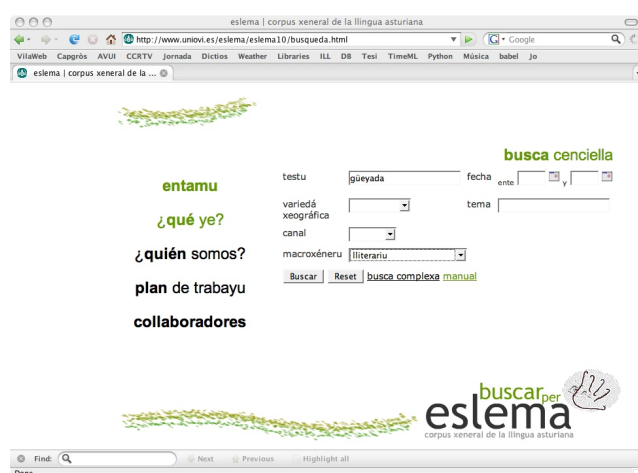


Figure 1: Current Eslema's search tool

4. The few resources of a less-resourced language

4.1. Funding sources

Eslema was born in the Research Group on Asturian Philology (*Seminariu de Filoloxía Asturiana*), within the Department of Spanish Philology at Oviedo University, and was initially supported by specific grants from the Culture Ministry of the Asturian Government, through the Office for Linguistic Planning. Since 2006, it is benefiting from funding from the National Plan for Research and Development of the Spanish Ministry of Culture and Education (period 2006-2008). Such funding is fundamental for basic issues such as equipment acquisition, or for covering for specific activities related to the project. However, it is fairly limited in the sense that it is insufficient for hiring researchers or project assistants for a reasonable period of time.

This situation has very negative consequences on the pace of the project, and has prevented us from making the progress we had initially planned. It is for that reason that, currently, more than a corpus appropriately annotated and provided with an adequate set of tools for accessing and managing its information, Eslema is simply a repository of documents with a basic search tool associated to it.

At present, we plan to apply for the yearly research grants offered by the Asturian Government for work on the autochthonous language, which would allow us to pay students and hire collaborators who can then work on Eslema full time, even if it is at a yearly basis.

4.2. Eslema's human face

In spite of the economic situation of the project, we consider that the amount of work accomplished up to now is remarkable. This has been possible thanks to the collaboration of professors and researchers from different departments at Oviedo University (namely, Spanish Philology and Computer Science), the assistance from researchers at Brandeis University and Georgetown University, in the United States of America, the specific contribution of graduated philologists as well as computer science students, and most remarkably, the unconditional dedication of very faithful supporters of the project, who cannot always be paid for their work but whose contribution is truly essential.

Furthermore, we benefit from the interest shown by some students in the undergraduate degree of Philology. In exchange for working on enlarging the corpus (mainly, collecting oral discourse and generating its corresponding transcription), we let them use Eslema as the framework within which to develop research projects funded by specific student scholarships.

Integrating all these collaborators in a project with such limited amount of resources is a challenging and arduous task. Even more so given the fact that, in the Asturian scientific scene, there was an absolute lack of experience in corpus linguistics, not to mention computational linguistics –cf. Saurí Colomer (2004). Furthermore, most of the people showing some interest in the project were coming from fairly unrelated backgrounds. They were either computer scientists with no knowledge whatsoever on linguistics, or

¹FreeLing is presented at <http://garraf.epsevg.upc.es/freeling/>. The specifics of the Spanish tagset can be found in: <http://garraf.epsevg.upc.es/freeling/doc/userman/parole-es.html>.

philologists with no computational background who, in addition, were specialized in very different areas of the field –e.g., language history, diachronic linguistics, literature, grammar.

Because of this difficult balance among so varied backgrounds, we are presently looking for additional funding for supporting initiatives to further educate our community of researchers, students, and collaborators on aspects that are of central relevance within the scope of the project, mainly by organizing short term courses that can address these topics.

4.3. External data contributors

The scarcity of funding prompted us to look for much needed resources to entities with no institutional relation to the project, but which nevertheless got involved and contributed in a very remarkable way. In particular, we are referring to the publishing companies Trabe and Ámbitu, which provided us with an important collection of digitalized text archives, and to the broadcasting station Radio Sele, which contributed a set of audio archives containing discourse of a genre not very common for the Asturian language –broadcastings of sport events, news reports, interviews, etc. Aside from their obvious value in terms of material resources, these contributions have been an important source of encouragement to us, and have helped project our work outwards at the social scene.

4.4. Collaborating with other research groups

Research funding in general is fairly neglected by our local administration. Because of that, we have decided to move to a more practical phase, focused on developing tools and products that can be of interest to private companies but also, and very specifically, to the public administration. For that purpose, we are currently in contact with the Culture Ministry of the Asturian Government concerning the creation of a Spanish-Asturian/Asturian-Spanish translator, a tool that would be inspired on the Opentrad translator between Spanish and the co-official languages under the Spanish administration (i.e., Catalan, Euskera, and Galician) (Alegría Loinaz et al., 2006), and which would effectively contribute to the much needed normalization of Asturian.

Given the limited material and economic conditions of our project, this constitutes a clear challenge. For that reason, we initiated collaboration contacts with groups with similar research interests. In particular, Xavier Guinovart, the head of the Research Group on Computational Linguistics (*Seminario de Lingüística Informática*) at Vigo University, has shown great interest in our translation project, to which his group could contribute the expertise it acquired during the development of the Galician part of Opentrad.

In parallel with that, we are also in collaboration with two research groups at Brandeis University (the Lab for Linguistics and Computation, and the Group for Research On Usage and Pragmatics), as explained in the next section.

4.5. Opening Eslema to the Asturian community. Wiki technology

Beyond the inherent interest Eslema offers to researchers on Asturian, the corpus was conceived with a true vocation of helping in both codifying the language and normalizing its use, or in other words, aiming at assisting Asturian speakers to be granted their linguistic rights.

As it happens, however, not only Eslema can benefit the Asturian community, but benefit from it as well. As is the case with many minorized languages in Western societies, a good number of Asturian speakers feels a fair degree of commitment to Asturian and its survival, an attitude that certainly plays in our favor.

Added to that, there is the fact that the role of information technology in Asturias has increased notably in the past few years. According to the latest report by Telefónica on the use of this type of technology within the borders of the Spanish administration, internet access in Asturian homes increased from 21.38% in 2003 to 41.4% in 2006, placing this community slightly above the Spanish average (39%).² The two necessary ingredients for involving Asturian society to the project are therefore at place: a motivated and committed attitude of speakers towards their language, and the basic technology for them to collaborate. Furthermore, that there is an interest on the population side towards making Asturian present in modern technologies can be attested by, for example, its incidence in the popular Wikipedia. Out of the 255 languages covered there, Asturian is ranked as the 74th in terms of number of articles (11,130), the 75th in terms of collaborators (833), and the 62th in terms of editing activity (242,204).³

Encouraged by these favorable conditions, we are currently collaborating with the Lab for Linguistics and Computation and the Group for Research On Usage and Pragmatics, both at Brandeis University, in developing a wiki-based platform for loading and annotating texts in Eslema. The relevance of wiki technology for facilitating online collaboration among the members of community of practice is already well-known. Wikis are easy to use, flexible in nature, and provide with a set of core components which can promote interaction among collaborators in various ways. Previous research developed at Brandeis has focused on the use of wikis for different educational and research tasks (Larusson & Alterman, 2007). A product resulting from that research is a wiki environment called WikiDesignPlatform (WDP) (Larusson & Alterman, 2008), which is conceived as an online meeting space for members in the same community of practice and, at the same time, a central repository and workspace. In addition, the WDP has been designed so that it can be tailored for other kind of collaboration activities.

So far, the emphasis of that research has been put on using wiki technology to support class work. Our collaboration will therefore be beneficial to both sides. Eslema will become the testbed for expanding WDP to different domain

²Refer to *La sociedad de la información en España 2007*, available at: http://sie07.telefonica.es/aplicacion_sie.html.

³http://meta.wikimedia.org/wiki/List_of_Wikipedias#10.2B_articles. Data obtained on March 30th, 2008.

needs and, at the same time, the resulting WDP will benefit our project both at the material level, allowing to enlarge the corpus at a minimum cost, and socially, making Asturian more active within information technologies and, consequently, contributing to slow down the serious process of substitution it is currently undergoing.

5. Contribution of the project

5.1. Increasing linguistic resources in Asturian

The availability of Eslema as a digitalized and structured collection of documents in Asturian is of great significance for linguistic research on this language. Eslema has opened venues for advancing in the knowledge of Asturian lexicon and grammar at the theoretical level. Within the Eslema framework, for example, a programmatic work on Asturian syntax has just been published (Viejo Fernández, 2008). This kind of work does not represent any novelty in more stable languages, but it is certainly a new contribution to the Asturian linguistic community. Along similar lines, there is a doctoral dissertation currently under development devoted to the analysis of prepositions, oblique complements, and verbal periphrasis in Asturian (Hinojal Díaz, frth). And finally, it is about to be published a wide anthology of historical texts (from Medieval Ages until nowadays) written in the Occidental Asturian dialect (Cueto & Viejo, frth). Furthermore, Eslema also plays a central role in the task of setting the autochthonous vocabulary and terminology, thus contributing to its codification. The adverse socio-political conditions of Asturian during the past centuries and its resulting diglossic situation caused all lexicographic enterprise to fail (Arias Cabal, 1996), at the time when other languages underwent the process of codification that guaranteed their nowadays healthy status; e.g., English, French, or Spanish. The sample of medieval and classical Asturian texts in Eslema, which date previous to the strong Castilianization suffered in the past half century, is of great use for recovering Asturian lexicon. As a matter of fact, the language from those periods provides consolidated models of specific domains and their corresponding terminology, such as administrative and juridic registers.

5.2. Social projection: boosting the confidence in our language

An incipient social projection of Eslema has manifested, for instance, through references to our work in *Radio Sele*, or through an article published in the Asturian weekly magazine *Les Noticies*, which caused an increase in the number of online visits to our corpus. In terms of the small social scale that defines the Asturian community, such public projection has generated some degree of expectation at two different levels. On the one hand, at the administration level, situation that can be of help to us when looking for funding in order to pursue new lines of research. On the other, among the general public, which tends to appreciate Asturian from a diglossic bias and now has the chance to start looking at it as an adequate and modern language because it is capable of benefiting from information technology. This boosting of confidence in Asturian will hopefully contribute to slow down, or even reverse, the serious process of substitution it is currently undergoing.

References

- Alegría Loinaz, I., Arantxabal, I., Forcada, M. L., Gómez Guinovart, X., Padró, L., Pichel Campos, J. R., & Walino, J. (2006). OpenTrad: Traducción automática de código abierto para las lenguas del estado español. *Procesamiento del Lenguaje Natural*, 37, 357–358.
- Arias Cabal, A. (1996). La lexicografía asturiana. Cronología de doscientos años d'intentos. *Lletres Asturianes*, 6, 41–63.
- Bird, S. & Simons, G. (2001). The OLAC Metadata Set and Controlled Vocabularies. In *Proceedings of the ACL Workshop on Sharing Tools and Resources for Research and Education*, (pp. 7–18)., Toulouse. ACL.
- Carreras, X., Chao, I., Padró, L., & Padró, M. (2004). FreeLing: An open-source suite of language analyzers. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, Lisboa.
- Cueto, M. & Viejo, X. (frth). *Antoloxía de textos n'asturianu occidental*. Asociación de Escritores en Asturiano Occidental. Forthcoming.
- Hinojal Díaz, R. (frth). *Complejos verbales y regimen preposicional de los verbos asturianos*. PhD thesis, Universidad de Oviedo. Forthcoming.
- Larsson, J. A. & Alterman, R. (2007). Tracking online collaborative work as representational practice: Analysis and tool. In Steinfield, C., Pentland, B., Ackerman, M., & Contractor, N. (Eds.), *Communities and Technologies 2007: Proceedings of the Third Communities and Technologies Conference*, (pp. 245–264)., Michigan.
- Larsson, J. A. & Alterman, R. (2008). Wiki technology for collaborative learning. In *Supporting and Tracking Collective Cognition in Wikis Symposium at ICLS 2008*, Utrecht. Forthcoming.
- Llera Ramo, F. (2002). II Estudiu siciollingüísticu d'Asturies. Avance de datos. *Lletres Asturianes*, 89, 181–197.
- Saurí Colomer, R. (2004). Un corpus pal asturiano. les tecnoloxíes llingüístiques na consolidación de les llingües minorizaes. *Revista de Filoloxía Asturiana*, 3/4, 135–174. Spanish version: http://www.cs.brandeis.edu/roser/pubs/rfa_corpus_sp.pdf.
- Suárez López, J. (2000). *Tesoros, ayalgas y chalgueiros. La fiebre del oro en Asturias*. Muséu del Pueblu d'Asturies.
- Viejo Fernández, X. (2003). *La formación histórica de la llingua asturiana*. Uviéu: Trabe.
- Viejo Fernández, X. (2004). Asturian: Resurgence and impending demise of a minority language in the Iberian Peninsula. *International Journal of the Sociology of Language*, 170, 169–190.
- Viejo Fernández, X. (2008). *Pensar asturiano. Ensayos programáticos de sintaxis asturiana*. Uviéu: Trabe.