Collaborative Language Documentation: the Construction of Repositories of the Huastec Language

Anuschka van ´t Hooft, José Luis González Compeán
Autonomous University of San Luis Potosí
Av. Industrias 101-A, Col. Talleres, 78399 San Luis Potosí, SLP, México
Technological Institute Cd. Valles
Av. Carr. al Ingenio Plan de Ayala km.2, Col. Vista Hermosa, 79010 Cd. Valles, SLP, México
E-mail: avanthooft@uaslp.mx, joseluig@yahoo.com

Abstract

In this paper, we describe the design and functioning of a web-based platform called Nenek, which aims to be an on-going language documentation project for the Huastec language. In Nenek, speakers, linguistic associations, government instances and researchers work together to construct a centralized repository of materials about the Huastec language. Nenek not only organizes different types of contents in repositories, it also uses this information to create online tools such as a searchable database with documents on Huastec language and culture, E-dictionaries and spell checkers. Nenek is also a monolingual social network in which users discuss contents on the platform. Until now, the speakers have created a monolingual E-dictionary and we have initiated an on-going process of the construction of a repository of written texts in the Huastec language. In this context, we have been able to localize and digitally archive documents in other formats (audios, videos, images), yet the retrieval, creation, storage, and documentation of this type of materials is still in a preliminary phase. In this presentation, we want to present the general methodology of the project.

Keywords: collaborative research, language documentation, online repositories.

1. Introduction

The Huastec language is a Mayan language spoken in the Mexican Gulf Coast region, in an area known as The Huasteca. This language has at least 215.500 speakers (INEGI 2010) and is the only Mayan language isolated geographically from the others, which are spoken in the southeastern part of Mexico, in Belize and Guatemala. Huastec language can be roughly divided into a western, eastern and southeastern variant that are present in the states of San Luis Potosí and Veracruz, respectively.

Until now, the generation of Huastec dictionaries and other written materials has resulted in a somewhat slow, disjointed, and static maintenance process for the Huastec language. Also, these sources are completely dispersed and contain local publications with a very limited distribution as well as a few written texts that are available on the internet. Access to both sources is rather difficult. In the Huastec case, the collecting of materials and the construction of dictionaries has almost exclusively depended on individual researchers, who usually perform this task through long-term fieldwork periods. This kind of methodology produces repositories that are mainly based on transcriptions, which commonly have static formats such as compact disks, tapes, books and articles. These repositories are rarely used by the speakers because they are either private or shared with other researchers only.[1]

We developed a collaborative strategy to construct the Huastec corpus through the use of the web, which may store an unlimited source of linguistic data including massive amounts of complete electronic texts that are usually in the public domain (Sinclair 2002). The key factors in the success of this strategy are the constant generation of contents (especially written texts and posts) and the availability of those contents online. However, at this point, the construction of the Huastec corpus through the retrieval of sources on the internet alone cannot be successful: there are still not enough online Huastec materials available, and the variability of their formats and contents do not favor the building of a solid linguistic repository.

This is why we constructed a web-based platform with which to develop a collaborative language documentation project and create, archive and analyze "a comprehensive record of the linguistic practices characteristic of a given speech community" (Himmelmann 1998:166). The platform is called Nenek (www.nenek.mx), which is a colloquial form of greeting in Huastec. Nenek combines digital archiving with language description tasks carried out by native speakers, linguistic associations, government instances and researchers. The way in which we promote the project is through an online monolingual social network in which speakers exchange ideas about their language and culture. At present, more than 1,800 Huastec speakers are actively involved in the project. Their internet activity generates materials in the Huastec language and enables us to retrieve and document different types of sources. At the same time, Nenek aspires to improve the weak situation and position of this language, and aims to strengthen its maintenance and revitalization process. We hope it will also be helpful for students and researchers who want to study themes related to the fields of linguistics or linguistic anthropology on Huastec, in particular to specialists in the natural language processing (NLP) of this lesser-resourced Amerindian language.

[1] We could discuss additional problems that arise when compiling the available Huastec sources, all of which are in tune with the ones described by Bird and Simons (2003) concerning language documentation and description projects.

In this paper we want to describe the design and functioning of the Nenek platform. In particular, we present the collaborative methodology of the project through which speakers, together with the Nenek staff, build different types of repositories.

2. The Nenek Platform

The Nenek platform creates virtual communities of indigenous languages that provide the speakers with a monolingual social network online. The social network includes functionalities such as profiles, work groups and contents management, as well as tools that allow the speakers to create web pages and blogs in which they can contribute to the repository building by sharing and discussing texts, audios, images and videos.

Each registered participant in Nenek has a personal account, with both a private and a public window to access the virtual community. In this private window, the speakers can store materials such as written texts, images, audios and videos. The private window includes both the monolingual social network and a set of linguistic collaborative applications called *Nenek-joined*. We created these specialized computer tools in order to encourage the virtual community to use the language to a major extent, starting with a lexicography tool for the making of E-dictionaries[2] and then constructed a spell checker[3] for the Huastec language. The *Nenek-joined* tools are also used by work groups that are in charge of specific tasks in the language documentation project such as the construction of E-dictionaries, spell checker validation and content evaluation.

Nenek´s public window is for all speakers and those interested in Huastec language and culture. Here one can find published repositories, a dynamic monolingual searchable dictionary, a spell checker and some other materials about Huastec language and culture. This means that the results of the tasks developed by both the work groups and Nenek staff are published here and are freely available. The public window of the virtual community is also interconnected with traditional social networks, such as Facebook, YouTube and Twitter for collecting sentences or small written texts from the speakers (NenekFacebook, 2014).

3. The Collaborative Methodology of Nenek

The language documentation activities are developed in the private window of each user. In this private window, the speakers can store materials such as written texts, images, audios and videos. Here, the user decides whether his or her materials can be consulted publicly, are private or may go into the repository.

Speakers who are interested in participating in Nenek, may choose between two different roles[4]:

- Nenek-User, who is a registered user who has access to the monolingual social network with his or her private account. These persons are mostly students, young workers or teachers who live in the Huasteca region, but a significant segment of participants are migrants who live in different cities in Mexico or the USA. Most of them are between 15 and 40 years old (78% of this age group still speaks the language). They are receptive to the written expression of their language and have internet (HD, 2013);
- Collaborator-User, who is a registered users who participates in a specific language documentation task and has access to both the social network, the private account and the linguistic tools. These users are commonly local linguists, academics and researchers. Like Nenek-users, these participants are registered in the monolingual social network, yet they also have access to Nenek-joined (that is, to the linguistic tools) in order to validate the materials deposited by Nenek-users and other Collaborator-users.

When generating materials (written texts, audio recordings, videos, photos, vocabulary entries), both Nenek-Users and Collaborator-Users decide among three options where to store these items:

- *Japidh:* This Huastec adjective (which means “open” or “disclosed”) represents the public content category. When a speaker introduces a content in the virtual community by choosing this category, the platform stores this content in the Nenek-social repository and it sends an e-mail alert to all participants who have accepted to receive it. This content is now open for viewing, but it is not included in the heritage repository.
- *Mapudh:* This Huastec adjective (which means “closed” or “enclosed”) defines the private content category. When a speaker introduces contents in the virtual community by choosing this category, the platform does not send alerts to the community.

[2] The E-Lexicography tool builds dictionaries attending the demands expressed in the literature about Internet dictionaries (Almind, 2005), since our pilot dictionary is easy to find, the search field is the center of attention, and it gives instant and simple results, which is limited to nine entries per page. Also, it has an autocomplete search function that predicts a word or phrase when the user is trying to type in and it gives alternatives and displays results. Nenek allows several workgroups to develop different dictionaries at the same time.

[3] Huastec speakers are not necessarily familiar with writing in their language. Moreover, there is no standardized alphabet or standardized spelling available for Huastec. In order to provide the speakers with a reference framework to write texts in Huastec, we developed CoTenek. This checker detects new spelling forms and gives multiple writing options for each term, so a speaker can choose whether he or she wants to use one of the options given or not. CoTenek is available for some of the most popular text editors, such as Microsoft Word, OpenOffice, LibreOffice (CoTenek 2014), and a Firefox version has been developed by Kevin Scannell from Saint Louis University by using CoTenek lexicon (CoTenekFirefox 2014).

[4] People who are only interested in consulting Nenek´s public window are called Public-Users. They are not registered and do not participate in any of Nenek´s activities. They can only consult and retrieve the information that is publicly available on the platform.

- *Wejladh K'anilab:* This category (which refers to something "chosen and stored") indicates that a speaker who is participating in a specific task donates content to the community heritage. The speaker offers his or her material to the repository, where it is stored. This time, the platform sends an e-mail notification to all the Collaborator-Users and automatically starts a consensus polling procedure to decide whether that content is valid for repository or not. The results of this evaluation process are reported to Nenek-Users who receive alerts, and thus start a second consensus polling procedure among the Nenek-Users. The basic idea is to emulate the meetings and the members' participation in the decision-making process of real communities. Only when accepted by the community the contents go into the heritage repository, where it is publicly available for all the speakers.

It should be said that while storing the materials, the user has to provide the metadata of each item, so that NLP researchers could make use of it.

Thus, in Nenek, the documentation activities are carried out collaboratively and in a cyclic process that starts when the speakers propose a task for a work group and store their materials in the *wejladh k´anilab* category, the heritage repository. Then, either the speakers' communication or input of materials returns to the virtual community after a categorization and consensus polling procedure (validation process) carried out by speakers, linguists or native linguistic associations who are Collaborator-Users.

Until now, the workgroups have been working on the construction of an E-dictionary of the Huastec language, which includes almost 2,000 entries and is the first dynamic Huastec dictionary online that is constructed in a collaborative manner by the speakers. During the working process, Nenek staff often leads the discussion and poses questions to the virtual Huastec community, for example by sending an image and asking about the forms to describe the item on the image. It then collects as much as ten different proposals, all of which are debated among the members of the work group. Moreover, the community also describes the specific region in which each of the expressions is used. All participants deliver their opinions to our web page by reacting to our question in a public manner, which represents a situation similar to when a researcher obtains terms during fieldwork.

Also, we have initiated the work on the construction of repositories of written texts in the Huastec language. The retrieved materials were obtained from three different sources: on the internet (based on crawler software designed for collecting Huastec texts), through donations of published and unpublished materials by their authors, and through the retrieval of written texts from the Nenek social network. Here too, work groups are created that stimulate speakers to hand in specific types of materials, such as essays, tales, anecdotes, or other written reports.

Nenek handles the contents as a digital library that includes dictionaries and repositories that are automatically categorized according to the type of the role used by the speaker who donates contents. Thus, the materials are stored into three different repositories:

- Nenek-social: This repository includes written texts donated by speakers (Nenek-users) through public blogs of the monolingual social network (NenekBlogs 2014).
- Nenek-academic: This repository includes documents and written texts donated by speakers who are Collaborator-Users. This repository also includes texts written in Huastec that were collected from a special edition of an academic journal coordinated for this project (JournalTenek 2013).
- Nenek-published: Nenek-published is the heritage repository of the virtual community. This repository includes published materials from two different sources. The first source includes books that are automatically recovered from public sites on the internet by using crawler software. The second source includes books that were donated to the project by both government instances and indigenous associations. The last source includes a set of publications that were digitized by the Huastec speakers of the Nenek staff.

As the written texts of the Nenek-published repository have passed through a full reviewing and editorial process, Nenek automatically uses them as valid for the corpus building. The materials from the other two repositories (Nenek-social and Nenek-academic), however, require verification and consensus from the virtual community before considering them for the corpus. It is important to note that all the repositories handled by Nenek are stored in a fault-tolerant cloud including sites in Spain and Mexico to guarantee the contents availability in failure scenarios (González et al. 2012; González et al. 2013).

As a result, and even though there are online sources that offer materials in Huastec (AILLA 2014; OLAC 2014; SOAS 2014; CAILLA 2014), Nenek-published is currently the largest repository of written texts in this language on the internet. It contains materials that belong to the fields of law, education and local oral traditions. Thus, our collaborative approach allowed us to cover a wide social profile for language usage.

It should also be mentioned that these other repositories are based on a depositor scheme in which the volume of contents depends on the activity of few researchers (sources). In addition, the most of the sources of these repositories have defined to deny the access to the contents. Contrastingly, the number of different sources in Nenek is significantly higher because the speakers, associations and government instances are collaborating in the content collection process. Besides, all of its contents is freely accessible online.

4. Conclusions

Before Nenek, Huastec speakers preferred Spanish as their language of communication on the internet because there were no cybernetic spaces where to use their

mother tongue. Now, we have young people joining the project and using the platform to search for friends and discuss various issues in their mother tongue. These speakers are participating in linguistic tasks or debates about Huastec sentences and are gradually creating their own initiatives and debates about their language. This means that they are writing their language –some of them for the first time- and that they do so in a new media, the internet. Consequently, Nenek has been able to expand the use of the language to spheres in which this language was not present before and has contributed to some extent to its revitalization.

The first stage of the language documentation process conducted through collaborative research allowed us to create a searchable pool of information that reflects part of the living language. Nenek concentrated this heritage in a centralized site that can reconstruct contents in failure scenarios. Since this is a collaborative platform, we did not only achieve to store the greatest quantity of materials, but also the most varied ones (at least in regard to written texts in Huastec). Nenek has proved to be an important tool in the language documentation process of the Huastec language.

In the following stage of the project we want to focus on the documentation of other materials, such as audio or video recordings, and images. These materials will give a better view on the living language in its social and cultural context (Gippert, Himmelmann & Mosel, 2006). We are constantly making improvements on the platform (for example, creating mobile applications) in order to make the collaborative work more profitable. Thus, we think, Nenek fosters the empowerment of native peoples in taking care of their linguistic and cultural heritage, making it a project for and by native speakers.

Nenek´s more inclusive process of repository building concentrates efforts and improves the collection results. The collaboration process of a growing social collective as well as the use of the crawler collector software appear to be more time effective than the deposits scheme used by traditional repositories. We believe that researchers who are interested in generating materials for other languages through collaborative approaches may take advantage of the described strategy.

5. Acknowledgements

The Nenek project is sponsored through a grant from the Mexican Secretary of Public Education and the National Council of Science and Technology (SEP-Conacyt research grant CB-2012-180863).

6. References

AILLA (2014). AILLA: The archive of the indigenous languages of Latin America, http://www.ailla.utexas.org/site/welcome.html

Almind, R. (2005). Designing Internet Dictionaries In I. Barz, H. Bergenholtz & J. Korhonen (eds). *Schreiben, Verstehen, Übersetzen, Lernen. Zu ein- und zweisprachigen Wörterbüchern mit Deutsch*. Frankfurt am Main: Peter Lang, pp.103-119.

Bird, S. & Simons, G. (2003). Seven dimensions of portability for language documentation and description. *Language* 79 (3), pp. 557-582.

CAILLA (2014). Chicago Archive of Indigenous Literatures of Latin America. http://cailla.uchicago.edu/?page=browse_by_language;family=4;language=48

CoTenek (2014). Co-Tenek, Huastec Spell checker, http://www.nenek.mx/huasteco.dic

CoTenekFirefox (2014). https://addons.mozilla.org/addon/huastec-spell-checker/

Gippert, J., Himmelmann, N.P. & Mosel, U. (2006) *Essentials of Language Documentation*. Berlin, New York: Mouton de Gruyter.

González, J.L. et al. (2013). González Compeán, J.L., Carretero Pérez, J. Sosa-Sosa, V. Rodríguez Cardoso, J.F. & Marcelín-Jiménez, R. An approach for constructing private storage services as a unified fault-tolerant system. *Journal of Systems and Software* 86 (7), pp. 1907–1922.

González et al. (2012). González Compeán, J.L., Sosa-Sosa, V., Bergua Guerra, B, Sánchez, L.M. & Carretero Pérez, J. Fault-Tolerant Middleware Based on Multistream Pipeline for Private Storage Services. In *Proceedings of the International conference for internet technology and secured transactions*. London, 10-12 Dec. 2012, pp. 548 – 555.

HD (2013). *Habilidades Digitales para todos*. http://www.hdt.gob.mx/hdt/

Himmelmann, N.P. (1998). Documentary and descriptive linguistics, *Linguistics*, 36, pp. 161-195.

INEGI (2010). *Censo de Población y Vivienda 2010*. Instituto Nacional de Estadística, Geografía e Informática. Aguascalientes (Mexico): INEGI. http://www.inegi.org.mx/est/contenidos/proyectos/ccpv/cpv2010/

JournalTenek (2013). Special Edition for Huastec Speakers, Teczapic ITV Journal. http://www.nenek.mx/esTeczapicITV/index

NenekBlogs (2014). Nenek Speakers In http://www.nenek.mx/prof.php.

NenekFacebook (2014). Nenek in Facebook. https://www.facebook.com/NenekMexico .

OLAC (2014). OLAC resources in and about the Huastec language. http://www.language-archives.org/language/hus

Sinclair, J. (2002). Intuition and annotation - the discussion continues. In K. Aijmer & B. Altenberg (eds.) *Advances in corpus linguistics*. Amsterdam: Rodopi, pp. 39-59.

SOAS (2014). The Endangered Languages Archive at SOAS, London. http://elar.soas.ac.uk/