

# Nenek: Digital Self-documentation for Minority and Under-resourced Languages

**Anuschka van 't Hooft, José Luis González**

Universidad Autónoma de San Luis Potosí, Cinvestav-Tamaulipas  
 Av. Industrias 101-A, Fracc. Talleres, 78399 San Luis Potosí, S.L.P., Mexico;  
 Km. 5.5 carretera Cd. Victoria-Soto la Marina, 87130 Cd. Victoria, Tamps, Mexico  
 avanthooft@uaslp.mx, jgonzalez@tamps.cinvestav.mx

## Abstract

The Nenek platform was designed to support speech communities in their language documentation efforts. Its features and linguistic tools enable speakers to work together in virtual communities and create, document, archive, and mobilize language resources. These, in turn, can be used to manufacture more complex multimedia language resources, all of which are available under the “open archive” principle. The four-stage cyclical management model and the use of virtual communities throughout the documentation process facilitate the self-documentation of language and culture and provide a space for speakers to discuss about and in their language. Also, it makes the minority language visible on the Internet.

**Keywords:** self-documentation, language resources management, virtual communities

## Résumé

An ts'ejkadh mulkux kaw Nenek k'wajat t'ajadh abal ka tolmiyat an kwenchal tin tsáb abal kin jila' dhuchadh in kawintal. In walkixtal ani in eyextalabh tin kawintal in t'ajal ka t'ojon ti jun yanel an kwenchalchik axi k'wajat ti al an tsu'udh kaw abal kin tsalpay, dhucha', dhaya' ani ka wat'baxin patal in kawintal. Patal axe' xi tolmixtalab k'wajat japidh abal jita' kin le'na' ba jun i tsakam muke'. Axe' xi tsalpaxtalab abal kin mulkuw an tolmixtalab in t'ajal ádhik an kwetem-dhuchnel kawintalab ani an biyal t'ajnel, in t'ajal jun i tamkuntalab abal ka t'ilmaxin an ebchalabchik tin kwetem kawintal, ani abal ka tejwamej axe' xi kawintalab ti al an buk'ux kaw.

### 1. Technological support for language documentation and revitalization efforts

Language documentation (Himmelman, 2006) is one of the main answers to language endangerment, as it aims to compile and preserve linguistic primary data of communicative events and creates interfaces to bring about the study of these data. One of the two leading paradigms of documentary practices is active documentation or documentation oriented to the community (Flores Farfán and Ramallo, 2010). Active documentation is often community-based and linked to efforts to maintain and revitalize endangered languages.

We developed an online collaborative strategy for speech communities to engage in self-documentation projects (Quatra, 2011) of their language and culture. In these projects, speakers of minority languages and under-resourced languages are the main investigators, compilers and users of language resources. These language resources are either audios, texts, videos or images, and can be combined and worked on to create multimedia materials.

Our project is called Nenek, which is also the name of the online platform that operates as a virtual community and was designed to support the self-documentation activities of the Huastec<sup>1</sup> speech community (van 't Hooft and González, 2014). In Huastec, “nenek” is an informal greeting, which aims to be inviting, plus it provides the platform with a recognizable local identity that make speakers feel part of a community. In this paper, we present the features of the Nenek platform (Nenek, 2019) and discuss its qualities to support self-documentation for minority language speakers.

### 2. The Nenek Platform: Language Resources Management Model

Nenek supports the creation of virtual communities of minority language speakers on the Internet. This platform includes a set of tools that enables users to work collaboratively on language documentation tasks, build lexicographic assets and produce new language resources (van 't Hooft and González, 2014; González, *et.al*, 2017a, 2017b). Here, it shifts from the common five-stage process of language documentation -recording, capture, analysis, archiving and mobilization- (Austin, 2006) to a four-stage cyclic management model to control the acquisition of existing materials in the target language and the manufacturing and archiving of new language resources, as well as their distribution within the virtual community and to the general public.

(1) In the acquisition stage, already existing language materials are either automatically extracted from the web by a crawler or received through donations from users who participate in a monolingual virtual community.

(2) In the manufacturing stage (merging recording, capture, and analysis), the speakers collaboratively document these acquired language materials, creating metadata and annotations. They also manufacture and document new language resources, either exclusively with their own means (e.g. filming a ceremony or recording a story) or using the acquired language resources (e.g. combining grandmother's knowledge on local gastronomy with photos from the virtual community to create a recipe book). All resources are discussed and validated by the virtual community before entering the repositories and corpora.

<sup>1</sup> Huastec is a Mayan language spoken across communities and towns in the subtropical Gulf region of Mexico. This language, known by its speakers as Tének, Teenek or Tenec (due to its three

writing systems), has three linguistic varieties (INALI, 2008) and at least 170.000 speakers (INPI, 2019).

(3) Meanwhile, Nenek creates repositories and computerized corpora of the language, which are exploited to generate linguistic tools such as spell checkers and e-dictionaries. Likewise, these tools enable the building of new language resources.

(4) The acquired and manufactured resources are published in the mobilization stage, either within the virtual community or publicly. All these language resources (in formats such as audios, videos, texts, photos and multimedia) are available on the Nenek platform as an e-library and can re-enter the manufacturing stage when used again to create more complex language resources.

A life cycle mapping scheme registers the transformations of the language resources at each of the stages of the language resources management cycle. This scheme also traces the utilization and diffusion of each resource that is produced by the virtual community.

### 3. The Nenek Platform: Virtual Communities

Speakers of minority languages are increasingly exposed to forced migrations, which are often the result of economic pressures and inequalities, but also of environmental disasters and social and political conflicts. These migration processes contribute to language loss. Our proposal to work through virtual communities brings these migrants -mostly young people or heritage speakers- into the documentation project, giving them spaces to use their language as well as a task in the conservation of their language and culture.

In a virtual community (VC), the participants are involved in social or emotional interpersonal interactions and have access to information. What makes users gather and operate as a community is both the type of relations they establish as well as the characteristics of that network. In a VC, users are interdependent: resources are transferred from the links existing among members and are based on their common interests (Gupta and Kim, 2004).

We identified the most effective strategies to encourage speakers' participation in the native language on the Internet (González Compeán *et.al*, 2017a). After creating pages on the most popular social networks (Facebook, Twitter), the life cycle of the VC was steered and monitored through the postings of three types of publications (announcements, debates, and pep talks) with a clear graphical identity, in a specific order and with regular time intervals. At the same time, speakers were invited to join the Nenek platform, which is also a VC, and collaborate in the acquisition, manufacturing and mobilization stages of the documentation process.

On the Nenek platform, the VC 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. 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 one of the repositories. Then, either the speakers' communication or input of materials returns to the VC after a categorization and consensus polling 187

procedure (validation process) carried out by the collaborators.

## 4. Discussion

Nenek was developed to support speakers' engagement in documentation activities while discussing among themselves about their language. It is a unique design that can be replicated in other documentation efforts of minority languages, since it can be localized and installed for minority languages in a relatively short time (Manuals are available online to facilitate its use). Among the special features of our project are its focus on active documentation, its cyclical resources management, and the use of virtual communities throughout the documentation process.

### 4.1 Active documentation

As an active documentation project, Nenek was set off by members of the Huastec speech community, who designed the platform and started the documentation activities. The platform can only operate in a close collaboration with the speech community, including their wishes and needs in each project, and involving them directly in all stages of the documentation process. Nenek addresses differentiation and heterogeneity in the speech community (Grinevald and Bert, 2011), provides training for collaborators, and guides the documentation activities aiming for speakers to become custodians of their linguistic and cultural heritage (Czaykowska-Higgins, 2009; Furbee, 2010). Thus, Nenek creates language resources that are esteemed to be necessary or valuable by the speech community and also agree with academic standards of language documentation, which enables their use by the scientific community.

The Nenek platform is flexible and can be adapted to the needs of speakers and their language. Not all features have to be used, and Nenek can operate next to other documentation or revitalization tools and practices. Also, it is versatile, in that it focuses on the documentation of language and culture and not only on linguistic documentation, as speakers often see language as being intricately interwoven with their cultural heritage (Franchetto, 2006). Accordingly, it supports the documentation of materials that deal with more cultural aspects of the speech community, such as dances and music, in formats that do not necessarily contain speech, such as photos.

### 4.2 Resources management

Nenek's resources management scheme incorporates an acquisition stage to collect and digitize donations and already existing materials in the target language on the Internet. For under-resourced languages, these materials are often dispersed and difficult to find (González Compeán *et.al*, 2017b). The inclusion of this stage in the documentation process provides a first e-library with language resources, which may also contain sources with information about the language and culture, as well as some linguistic tools (e-dictionaries, spell checkers), and is vital to attract speakers to the documentation project and initiate the virtual community (Garber, 2004; Iriberry and Leroy, 2009).

The management scheme is cyclical, in that manufactured language resources that are validated by the speakers are automatically mobilized and can be used again to create more complex ones. This mobilization through the VC enhances the use of language resources, fuels discussions among speakers about the contents and about their language, and strengthens the VC.

Our resources management scheme allows the building of linguistic tools that support speakers in their documentation activities, as they require a degree of literacy (for annotations, metadata, messages to the CV). More often than not, minority language speakers are short of experience with reading and writing in their mother tongue in any medium and they develop literacies during the project, which must be carefully discussed (Lüpke, 2011).

### 4.3 Virtual communities

Nenek uses virtual communities in all stages of the project, which is challenging but feasible. Young speakers of minority languages are present on the Internet and are participating in social networks. Some of them already are involved in digital language activism projects (Llanes-Ortiz, 2016) to promote their languages. Migrant speakers are especially keen to collaborate, since it connects them with people who share the same cultural background. When creating new language resources, their activities usually engage older members of their families and home villages, so that the whole speech community becomes involved and a broad range of verbal expressions can be addressed.

The key factors in the success of a VC are the constant generation of contents and the availability of those contents online. Our strategy to improve the life cycle of the VC requires a continuous mobilization of the acquired and manufactured language resources (González Compeán *et.al*, 2017a). In order to be successful, the contents on the platform, in the format of multimedia materials, should address, in the first place, community needs and interests in the language and culture.

With the use of VCs as a means to develop language documentation projects, language resources are acquired, manufactured, mobilized and discussed on the Internet, which brings about a greater visibility of these minority languages. For young people, it can be stimulating to see that their language is consistent with a modern global network society like the Internet and that the minority language can be used to express oneself in this new system of communication. Therefore, it also enhances processes of digital self-determination (McMahon, 2013).

## 5. Ethical issues

Active documentation recognizes that the situation and position of the world's languages are an expression of historical processes in which some languages have had economic, cultural and political advantages over others. Different from the type of documentation in which solely scientific questions are addressed, it conceives of documentation as a means to assist speech communities in their efforts to maintain and revitalize their language (Flores Farfán and Ramallo, 2010), and as a way to help guarantee the use of the mother tongue as a fundamental human right (Unesco, 2003). Here, several ethical issues have to be addressed, of which we can only mention a few.

We aspire to improve the weak situation and position of minority languages on the Internet and address the linguistic rights of people to participate on the Internet on their own terms, thus contributing to revitalization processes. For one, the Nenek platform is localized in the minority language which is the focus of the project. Also, the members of the virtual community communicate in their mother tongue while collaborating online and not in the majority language. This monolingual immersion is a logical consequence of our aim to create more horizontal relations among all participants in the project, at the same time expanding the domains of the use of the minority language and creating more visibility of this language on the Internet.

Additionally, we recognize the demand of speakers to have access to the outcome of the documentation project. Nenek is an open archive with multiple repositories, which are available to speakers. Automatic URLs guarantee the rights of the authors and participants of each language resource.

Furthermore, Nenek recognizes the oral expression as an essential component in the transmission of linguistic and cultural knowledge and as a fundamental element to express and protect the linguistic and cultural diversity (Maffi, 2003), including all varieties of the language, and favoring the compilation and mobilization of oral language resources. The written expression is used in new media, social networks and other internet-based resources, accepting all existing written norms for languages that are not standardized. Standardization of the language is not held to be a prerequisite to promote literacy processes.

This way, the Internet can become a strong ally in preserving and revitalizing minority languages, as speakers find here linguistic situations and materials that allow and even vindicate the use of the language, and they widen the spaces for expression in their mother tongues. However, in order to be effective, these virtual initiatives should be accompanied by other efforts to preserve the languages (according to the sociolinguistic diagnosis of each speech community), as well as by appropriate linguistic policies to ensure and promote their autonomous development.

## 6. Conclusions

Nenek's technological support fosters the empowerment of indigenous peoples in taking care of their linguistic and cultural heritage, making it a project for and by native speakers. At present, more than 3,300 Huastec speakers are actively involved in the Nenek project and decide together on the activities they want to develop to document and promote their language. Their Internet activity generates materials in the Huastec language and enables us to retrieve and document different types of sources.

In Nenek, the complete resource life cycle happens on the Internet, which represents a source to get new resources (acquisition stage), a space to produce (manufacturing stage) and store (archiving stage) new resources and a site to publish acquired and manufactured language resources (mobilization stage).

To the members, contributing in a VC through collaborative action in the production of digital contents is an opportunity to articulate, transmit and discuss knowledge and traditional practices. By placing their language on the Internet, speakers participate to expand

their linguistic heritage into new domains of usage and build a forum to discuss issues about their language and culture. This participation makes the language and culture visible in a virtual medium like the Internet. Nenek is an example of the important role the Internet plays in the current repositioning of indigenous and minority languages as opposed to the dominant languages.

## 7. Acknowledgements

The Nenek project was sponsored by CONACYT, the Mexican Council for Science and Technology, through grant CB-2012-180863. A technology transfer to INALI, the National Institute of Indigenous Languages, makes it possible to support speech communities in Mexico who want to use our platform and start their own language documentation project. We thank Alejandra Santiago Bautista for translating the abstract of this paper into the Huastec language.

## 8. References

- Austin, P. K. (2006). Data and Language Documentation. In J. Gippert, N. P. Himmelmann and U. Mosel (eds.), *Essentials of Language Documentation*, pages 87-112. Berlin / New York: Mouton de Gruyter.
- Czaykowska-Higgins, E. (2009). Research Models, Community Engagement, and Linguistic Fieldwork: Reflections on Working with Canadian Indigenous Communities. *Language Documentation and Conservation* 3(1):15-50.
- Flores Farfán, J. A. and Ramallo, F. (2010). Exploring Links Between Documentation, Sociolinguistics and Language Revitalization: An Introduction. In J. A. Flores Farfán and F. Ramallo (eds.), *New Perspectives on Endangered Languages. Bridging Gaps Between Sociolinguistics, Documentation, and Language Revitalization*, pages 1-12. Amsterdam / Philadelphia: John Benjamins Publishing Company.
- Franchetto, B. (2006). Ethnography in language documentation. In J. Gippert, N. P. Himmelmann and U. Mosel (eds.), *Essentials of Language Documentation*, pages 183-211. Berlin / New York: Mouton de Gruyter.
- Furbee, N. L. (2010). Language Documentation. Theory and Practice. In L. A. Grenoble and N. L. Furbee (eds.), *Language Documentation. Practices and Values*, pages 3-24. Amsterdam / Philadelphia: John Benjamins Publishing Company.
- Garber, D. (2004). Growing Virtual Communities. *International Review of Research in Open and Distance Learning* 5(2):1-7.
- González Compeán, J. L., van 't Hooft, A., Carretero Pérez, J., and Flores Martínez, L. (2017a). La introducción de la lengua huasteca a internet. Una estrategia para crear comunidades virtuales en lenguas amerindias. *Comunicación y Sociedad* 28:131-153. <http://www.comunicacionsociedad.cucsh.udg.mx/index.php/comsoc/article/view/6399> (Accessed: 21 November 2019).
- González, J. L., van 't Hooft, A., Carretero Pérez, J. and Sosa Sosa, V.J. (2017b). Nenek: a cloud-based collaboration platform for the management of Amerindian language resources. *Language Resources and Evaluation*, 51(4):897-925. doi: 10.1007/s10579-016-9361-8
- Grinevald, C. and Bert, M. (2011). Speakers and Communities. In P. Austin and J. Sallabank (eds.), *The Cambridge Handbook of Endangered Languages*, pages 45-65. New York: Cambridge University Press.
- Gupta, S. & Kim, H. W. (2004). Virtual community: Concepts, implications, and future research directions. *Proceedings of the Tenth Americas Conference on Information Systems*, pages 2679-2687. New York.
- Himmelmann, N. P. (2006). Language Documentation: What is it and What is it Good for? In J. Gippert, N. P. Himmelmann and U. Mosel (eds.), *Essentials of Language Documentation*, pages 1-30. Berlin / New York: Mouton de Gruyter.
- Hooft, A. van 't and González Compeán, J. L. (2014). Collaborative Language Documentation: The Construction of the Huastec Corpus. In L. Pretorius, C. Soria and P. Baroni (eds.), *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 67-70. Reykjavik, Iceland, European Language Resources Association (ELRA).
- INALI (2008). *Catálogo de las Lenguas Indígenas Nacionales: Variantes Lingüísticas de México con sus autodenominaciones y referencias geoestadísticas*. Mexico, Instituto Nacional de Lenguas Indígenas.
- INPI (2019). *Atlas de los pueblos indígenas de México*. <http://atlas.cdi.gob.mx/> Mexico, Instituto Nacional de Pueblos Indígenas (Accessed: 7 October 2019)
- Iriberry, A. and Leroy, G. (2009). A Life-Cycle Perspective on Online Community Success. *ACM Comput. Surv.* 41(2), article 11. <https://dl.acm.org/citation.cfm?id=1459356> (Accessed: 21 November 2019)
- Llanes-Ortiz, G. (2016). Primeros pasos de estudio participativo: lenguas indígenas y medios digitales. <https://rising.globalvoices.org/lenguas/2016/01/12/estudio-addli-primeros-pasos/> (Accessed: 21 November 2019)
- Lüpke, F. (2011). Orthography development. In P. K. Austin and J. Sallabank (eds.) *The Cambridge handbook of endangered languages*, pages 312-336. New York: Cambridge University Press.
- Maffi, L. (2003). The 'Business' of Language Endangerment: Saving Languages or Helping People Keep Them Alive. In H. Tonkin and T. Reagan (eds.), *Language in the 21st Century*, pages 67-86. Amsterdam / Philadelphia: John Benjamins Publishing Company.
- McMahon, R. (2013). *Digital Self-determination: Aboriginal Peoples and the Network Society in Canada*. Ph.D.Thesis. Vancouver, Simon Fraser University.
- Nenek (2019). Nenek. Platform for collaborative self-documentation of language and culture. <http://nenek.inali.gob.mx> (Accessed: 21 November 2019)
- Quatra, M. M. (2011). 'Auto-documentación Lingüística': La Experiencia de una Comunidad Jodí en la Guayana Venezolana. *Language Documentation & Conservation*, 5:134-156.
- UNESCO (2003). *Language Vitality and Endangerment*. UNESCO Ad Hoc Expert Group on Endangered Languages, 23 pp.