Challenges for Language Technologies in Critically Endangered Languages

Jhonnatan Rangel

Challenges for Language Technologies in Critically Endangered Languages

Jhonnatan Rangel
INALCO, SeDyL
7 Rue Guy Moquet, 94800 Villejuif
jhonnatan.rangelmurueta@inalco.fr

Abstract

There are currently 577 critically endangered languages in the world, making up almost 10% of all languages. These languages are also technologically low-resourced and are only spoken by a few elder speakers. As such, critically endangered languages pose various fundamental challenges, such as the annotation bottleneck, that seriously hinder future perspectives of language documentation, preservation, reclamation, revitalization and utilization in language technologies. This paper addresses the challenges critically endangered languages face in implementing language technologies.

Keywords: low-resourced, critically endangered, language technology

Resumen

En el mundo hay 577 lenguas en muy alto riesgo de desaparición que representan casi el 10% de todas las lenguas. Estas lenguas, que además cuentan con pocos recursos tecnológicos, las hablan únicamente unos cuantos adultos mayores. Las lenguas en muy alto riesgo de desaparición plantean retos fundamentales, como el cuello de botella de anotación, que limitan enormemente las perspectivas de documentación, mantenimiento, recuperación, revitalización y uso de tecnologías del lenguaje. Este artículo aborda los retos que enfrentan las lenguas en muy alto riesgo de desaparición en cuanto a la implementación de tecnologías del lenguaje.

1. Critically endangered languages

There are currently 577 critically endangered languages in the world, making up almost 10% of the world's 6,000+ languages (Moseley, 2010).

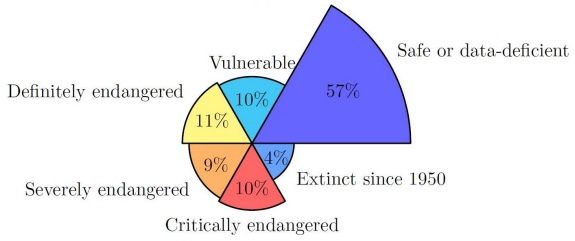

Figure 1: Vitality of the world's languages

These languages present the following characteristics that directly impact their vitality in the short term (Rangel, 2019):

- Youngest speakers are grandparents and older
- Small proportion of speakers in relation to larger community
- Inter-generational (or traditional) transmission of the language interrupted for decades
- Infrequent use of the language in language practices of remaining speakers
- Lack of written tradition and literacy in the language

Critically endangered languages can intersect with limited documentation and/or limited studies, and in most cases they are also indigenous or autochthonous languages. All of these characteristics can coincide such as in the case of Ayapa Zoque, Ayapaneco or *numde 'oode* (autonym), an indigenous, under-documented, under-studied and critically endangered language spoken in southern Mexico (Rangel, 2017; Rangel, 2019).

Because language endangerment is a global phenomenon, there are critically endangered languages on every continent of the world (Moseley, 2010). As these languages are at the highest level of endangerment, their disappearance could occur at any time in the next decade. Consequently, concrete and multifaceted measures must be taken immediately to reverse or at least slow down language endangerment in critically endangered languages such as Ayapaneco.

2. Low-resourced languages

Low-resourced languages (LRL) do not have the extensive resources required (annotated and parallel corpora) for the implementation of Language Technologies and techniques such as Machine Translation or Machine Learning. It is estimated that out of the world's 6,000+ languages, only about 20 of them have the resources to be considered high-resourced languages (HRL) while an additional 60 have some sort of resources available to be considered medium-resourced languages (MRL) (Duong, 2017).

HRL	MRL	LRL
0.4%	1%	98.6%

Table 1: Resource distribution of world languages

This means that Language Technologies are only applied to about 1.4% of the world's languages, leaving the vast majority of them unattended. Examples of HRL include English, Spanish or French while MRL include Hebrew,

Indi or Czech. These languages are spoken by millions of people of multiple generations in the world and therefore are not necessarily at immediate risk of disappearing.

On the other hand, some LRL are spoken by millions of people and are not at risk of immediate disappearance, such as Swahili. However, LRL can be critically endangered languages such as Ayapaneco. While there is a tendency for critically endangered languages to also be LRL, the opposite correlation does not always hold (not all LRL tend to be critically endangered languages). Indeed, a language's number of speakers does not necessarily determine the resources available. In fact, the development of language resources is strongly influenced by social, political, and financial factors. For instance, languages that are considered international and spoken predominantly in Western, Educated, Industrialized, Rich and Democratic (WEIRD) societies (Henrich et al., 2010) have an abundance of resources while minority languages spoken in non-WEIRD societies significantly lack resources (King, 2015).

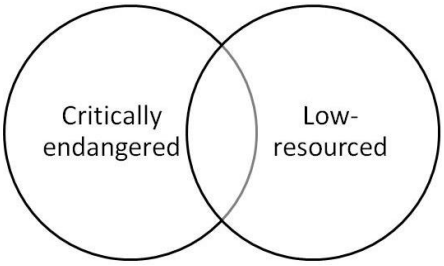

Figure 2: Linguistic and technological characteristics

I will address the challenges faced by languages situated in the overlapping zone between low-resourced and critically endangered that accounts for almost 10% of the world's linguistic diversity.

3. Challenges

Low-resourced critically endangered languages face multiple and multifaceted challenges for implementing language technologies. I will address 6 major challenges.

3.1. Annotation bottleneck

The annotation bottleneck corresponds to the gap between the large amount of data that we are capable of gathering with current technology and the limited amount of data we are capable of annotating (transcribing, translating and glossing). Estimates suggest that for every hour of recording, between 40 and 100 hours are required to annotate it (Seifart et al., 2018).

Although the annotation bottleneck is a challenge that can impact every language in the world, it is amplified in the context of critically endangered languages. The reason for this, as mentioned above, is that critically endangered languages are also LRL. Consequently, applying Natural Language Processing (NLP), Machine Learning or Artificial Intelligence tools and techniques that could help us to open up this annotation bottleneck is convoluted. By contrast, these tools and techniques can be implemented more easily in HRL to reduce the gap between data available and annotated data.

Although promising advances have recently been made in implementing Language Technologies for LRL (Ćavar et al., 2016), the annotation of critically endangered languages continues to be a primarily manual task carried out by a few researchers and community members, contributing in turn to the existing annotation bottleneck. This brings us to the second group of challenges, human resources.

3.2. Human resources

Human resources tend to be very scarce in the context of critically endangered languages. From the amount of existing speakers to the potential manual annotators, human resources are limited.

The potential universe of annotators of these languages is reduced to a language expert (which could be a linguist, an anthropologist or a teacher) and in best case scenarios, a few speakers. As mentioned, in the context of critically endangered languages, the youngest speakers are grandparents and older. Consequently, the remaining speakers of these languages cannot always contribute to the annotation process because as they get older, they suffer from physical conditions such as vision and hearing problems that prevent them from participating in the annotation of their language, not to mention that they also have very limited digital literacy.

A possible solution to this challenge is recruiting young community members who could help annotate data. The caveat is that in many critically endangered languages, the younger community members do not know the language well enough to perform this task by themselves and they need the help of elder speakers. As the number of remaining speakers is limited and their physical conditions are not optimal, this task becomes very slow and cumbersome.

I usually spend 60-120 hours annotating for every hour of Ayapaneco recorded using assisted annotation tools such as ELAN (Wittenburg et al., 2006). One younger community member with limited digital literacy assists with the annotation. However, neither of us know the language well enough to perform this task by ourselves and still require the speakers' input, resulting in very slow progress.

Recently, some tools based exclusively on oral annotations such as SayMore (Moeller, 2014) have emerged as a solution to opening up the annotation bottleneck. Although this can be a promising option for some world languages, it can be complicated to implement it in critically endangered languages because speakers are still required to contribute to the oral annotations, and speakers are scarce in these languages.

3.3. Capacity

As mentioned above, critically endangered languages are in most cases indigenous or autochthonous languages and

are also minority languages spoken in societies with poor economics. This directly impacts the capacity and the infrastructure available for these communities.

A very common capacity challenge among communities in which critically endangered languages are spoken is the scarce access to computers. Contrary to WEIRD societies, computer access can be very limited in the context of critically endangered languages as these communities most likely face poverty and marginalization. Without computers, it is difficult to introduce Language Technologies in these communities.

The recent global democratization of cellphones could facilitate the introduction of Language Technologies in these communities provided that access to internet is guaranteed. Unfortunately, this is not always the case. In Ayapa, the village where Ayapaneco is spoken, very few people own a computer but cellphone availability has dramatically increased in recent years with many community members owning one.

A second capacity challenge is related to literacy. In order for critically endangered languages to have some sort of online presence (ex: social media), they need to be written. However, as the case of Ayapaneco illustrates, not all world languages have a writing system, and when it comes to critically endangered languages, this seems to be the norm rather than the exception. Indeed, orthography development in Ayapaneco is a recent endeavour, and the writing system is not yet functional. Currently only two people are familiar with the orthography and therefore Ayapaneco is not yet used online.

Most speakers of critically endangered languages tend to be bilingual in their minority language and the majority language of the wider society. Nevertheless, when speakers of critically endangered languages are literate, they are only so in the majority language. Consequently, there is a lack of written tradition and literacy in these languages, complicating the task of applying Language Technologies.

3.4. Infrastructure

Closely related to the previous point, internet access tends to be a common challenge among endangered language communities. When internet access is available, it can be expensive to access given the poverty and marginalization discussed above. Critically endangered languages are commonly spoken in rural areas with limited or unreliable internet access. In Ayapa, community members access the internet mainly via mobile internet with a Smartphone. That said, internet access is cost-prohibitive for most community members given their economic status.

Without proper internet access, these communities will continue to struggle to bring their languages online, resulting in a circular dynamic regarding the lack of resources in critically endangered languages.

3.5. Documentation and study

When critically endangered languages face limited language documentation and studies, the perspectives for Languages Technologies are tortuous to say the least. Is not a coincidence that HRL and MRL are among the best documented and most studied languages, while a good amount of LRL are currently under-documented and under-studied like Ayapaneco.

When documentation and studies of a language increase, so do the chances of implementing Language Technologies. Consequently, it is fundamental to improve the documentation and study of critically endangered languages, especially those that are among the least documented and studied.

3.6. Linguistics

Language variation can pose a challenge for Language Technologies in critically endangered languages. Variation is an intrinsic characteristic of human language, and it follows the orderly heterogeneity premise (Weinreich et al., 1968) under which language variation can be conditioned by either linguistic or social factors, or the interaction of both. While variation is widely attested in all world languages, the study of critically endangered languages has recently called into question the orderly heterogeneity premise. Indeed, critically endangered languages exhibit a large proportion of unstructured variation that cannot be linked to social or linguistic factors (Dorian, 2010). My recent research on Ayapaneco shows that this language also exhibits a high proportion of unstructured variation, thus confirming the trend found in other critically endangered languages (Rangel, 2019).

Unstructured variation has been left out of Language Technologies. This is understandable considering that until recently, it has also been overlooked by the fields traditionally concerned with variation in general. Furthermore, as unstructured variation is widely present in critically endangered languages that also happen to be LRL, this contributes to the existing blind spot in modeling, processing and analyzing this type of variation and hinders the documentation, description, and revitalization of these languages as well as the implementation of Language Technologies.

4. Conclusion

The implementation of Language Technologies in critically endangered languages poses complex and multifaceted challenges such as the annotation bottleneck, heavy limitations in human resources and capacity, scarce infrastructure, limited documentation and study, as well as under-studied linguistic particularities.

These multifaceted challenges seriously hinder future perspectives not only for Language Technologies but also for the documentation, preservation, reclamation, and revitalization of critically endangered languages. Consequently, it is imperative to think outside of the box to apply these technologies as they could help maximize the limited

time we have left to engage with critically endangered languages. As these languages represent almost 10% of the world's total, and could disappear at any time in the next decade, time is of the essence and this task should be prioritized.

It is no longer enough to have just a few isolated experts and community members working to document and study these languages as the annotation bottleneck severely limits language reclamation and revitalization efforts as well as the deployment of Language Technologies. On the contrary, concrete and multifaceted measures must be taken immediately involving new multidisciplinary approaches while creating synergies among varied actors (academia, governments, NGOs, communities and civil society) to better address these challenges and support capacity building in the long term for these communities. Simultaneously, and more fundamentally, we must address the root causes of these challenges such as inequality, poverty and discrimination as is not a coincidence that critically endangered languages are also low-resourced. Thus, the lack of resources and Language Technologies replicates those social, political, and economic inequalities existing in the world. The next few years will be decisive in attempting to break this circle before a significant proportion of languages disappear from the face of the Earth.

5. Bibliographical References

Ćavar, M., Ćavar, D., and Cruz, H. (2016). Endangered language documentation: Bootstrapping a chatino speech corpus, forced aligner, ASR. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4004–4011, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Dorian, N. (2010). *Investigating variation: the effects of social organization and social setting*. Oxford studies in sociolinguistics. Oxford University Press, Oxford-New York.

Duong, L. (2017). *Natural Language Processing for Resource-Poor Languages*. PhD dissertation, University of Melbourne.

Henrich, J., Heine, S. J., and Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2-3):61–83.

King, B. P. (2015). *Practical Natural Language Processing for Low-Resource Languages*. PhD dissertation, University of Michigan.

Moseley, C. (2010). UNESCO Atlas of the World's Languages in Danger. Online version: `ttp://www.unesco.org/culture/en/endangeredlanguages/atlas`.

Rangel, J. (2017). Les derniers locuteurs : au croisement des typologies des locuteurs de langues en danger. *Histoire Epistémologie Langage*, 39(1):107–133.

Rangel, J. (2019). *Variations linguistiques et langue en danger. Le cas du numte oote ou zoque ayapaneco dans l'état de Tabasco, Mexique.* Thèse de doctorat, INALCO.

Seifart, F., Evans, N., Hammarström, H., and Levinson, S. (2018). Language documentation twenty-five years on. *Language, Journal of the Linguistic Society of America*, 94(4).

Weinreich, U., Labov, W., and Herzog, M. (1968). *Empirical foundations for a theory of language change*. University of Texas Press, Austin, Texas.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. (2006). ELAN: a professional framework for multimodality research. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May. European Language Resources Association (ELRA).