linguapax review 9

2021

Language Technologies and Language Diversity

Tecnologies de la llengua i diversitat lingüística

Linguapax

Linguapax Review 2021

Language Technologies and Language Diversity

Tecnologies de la llengua i diversitat lingüística

Editat per:

Linguapax International
In official partnership with UNESCO (consultative status)

Amb el suport de:

Generalitat de Catalunya
Departament de Cultura

Generalitat de Catalunya
Departament d'Acció Exterior i Govern Obert

Coordinació editorial: **Maite Melero Nogués**
Disseny i maquetació: **Grafia, serveis gràfics**
Traduccions: **Sara Blackshire**

CONTENTS
CONTINGUTS

FOREWORD

András Kornai[1]
SZTAKI Institute of Computer Science

With this special issue of Linguapax Review on Language Technologies and Language Diversity, readers have an extraordinarily rich and very timely resource at hand. To quote Juan Carlos Moreno Cabrera's article **Xenography: the impact of linguistic diversity on the evolution of writing**: "Linguistic diversity is one of the most conspicuous manifestations of the impressive adaptability and creativity of this curious animal we call *human being*".

In spite of all the known difficulties of drawing the line between language and dialect, the diversity is enormous: the current (24th) edition of the Ethnologue (Eberhard, Simons, and Fennig, 2021) lists 7,139 living languages. But this diversity hides great inequality: if all languages would be the first language of same number of people, we would have over 1.1m speaker/language. Yet at the endangered end, about a fifth of the languages have less than a thousand speakers, and at the vital end, the top 20 languages reach over 70% of the world population.

It is not at all trivial to measure this, and **Internet and linguistic diversity: the cybergeography of languages with the largest number of speakers** by Daniel Pimienta is a *tour de force* even though the broadest swath of languages considered is the 330 "richest" languages with over a million L1 speakers. As Bromham et at. (2021) note, the greatest predictor of language endangerment is the number of first-language (L1) speakers, though there are other important factors. It is worth emphasizing that many of these factors cannot be controlled by policy:

> A language is more likely to be endangered if a higher proportion of languages in the region are also endangered, suggesting that, in addition to language-specific threats, there are also widespread factors that influence language vitality across a region. (Bromham 2021)

Other factors, such as road density or average level of schooling, can only be controlled by regressive policies that encourage a brutal 'let's keep them on the reservation' form of colonialism. This is best countered by Steven Bird's paper **Creating a World that Sustains its Languages**, which offers a humanistic program of "concrete actions that you can take to help create (...) a world that sustains its languages", ranging from the simplest everyday actions of greeting people in their own language to more complex language games.

Besides outright colonialism (and its greatest driver, racism) there is also a form of misplaced romanticism, already noted by Russell (1937), that sometimes serves as a justification for a reservationist policy. This is the deeply flawed assumption that the 'purest' and 'most authentic' form of any language is manifested in the best isolated 'most backward' dialect (Kornai 2019).

1.- kornai@sztaki.hu

Altogether, the Bromham et al. (2021) results justify areal approaches such as described in **The Indigenous Languages Technology (ILT) Project at the National Research Council of Canada, and its Context** by Roland Kuhn and **Language technology and language revitalization (Rising Voices)** by Eddie Avila, which identifies eight strategies to organize digital activism:

1 Facilitating digital communication in Indigenous languages
2 Multiplying Indigenous language content online
3 Normalizing the use of Indigenous languages online
4 Educating in and teaching Indigenous languages online
5 Reclaiming Indigenous languages and knowledges
6 Imagining and creating new media in Indigenous languages
7 Defending spaces for Indigenous languages and linguistic rights
8 Protecting linguistic heritage and communities

While not organized along the exact same lines, Bird's paper gives specific examples of most of these strategies from the viewpoint of the *individual activist:* "(...)concrete actions that you can take to help create this future world (...) You don't need to join a campaign".

For the technologist the key issue, asked clearly in **Linguistic Diversity in the Age of Big Data** by Tunde Adegbola is whether "modern Information Communication technologies will promote or work against linguistic diversity". There is no clear-cut answer, as Adegbola notes, this will "depend on the way they are employed". In the digital arena, Kornai (2013) argued that the best predictor of vitality is the size of the Wikipedia, and to get there requires standardized orthography. This is a subtle matter, as discussed in Cabrera's contribution, but the cause and effect mechanism is very clear: to get the kind of tools discussed by Kuhn and by **OpenSpeaks: Transforming learning of tech innovations in low-resource language documentations to Open Educational Resources** by Subhashish Panigrahi, we need corpora, which come about only by communities using their speech and language in digitally mediated contexts.

Altogether, all main viewpoints, from the individual to the collective, from the areal to the global, from the native speaker to the language-agnostic technologist, are well represented in this collection, and anyone interested in diversity will find that this special issue will repay handsomely the attention one must pay to the papers collected here.

References

Bromham, Lindell et al. (2021). "Global predictors of language endanger- ment and the future of linguistic diversity". In: Nature Ecology & Evolution. issn: 2397-334X. doi: 10.1038/s41559-021-01604-y. url: https://doi.org/10.1038/s41559-021-01604-y

Eberhard, David M., Gary F. Simons, and Charles D. Fennig, eds. (2021). Ethnologue: Languages of the World. 24th ed. SIL International. url: http://www.ethnologue.com

Kornai, András (2013). "Digital language death". In: PloS ONE 8.10, DOI 10.1371/journal.pone.0077056. url: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0077056

— (2019). The Ni'ihau Paradox: isolation or assimilation. url: https://kornai.com/Slide/lt4all.pdf

Russell, Bertrand (1937). "The Superior Virtue of the Oppressed". In: *The Nation* 144.

INTRODUCCIÓ

András Kornai[1]
SZTAKI Institut de Ciències de la Computació i Control

Amb aquesta edició especial de Linguapax Review sobre les tecnologies de la llengua i la diversitat lingüística, els lectors disposen d'un recurs extraordinàriament ric i molt oportú. Citant l'article de Juan Carlos Moreno Cabrera **Xenografia: l'impacte de la diversitat lingüística en l'evolució de l'escriptura**: "La diversitat lingüística és una de les manifestacions més conspícues de l'impressionant adaptabilitat i creativitat d'aquest curiós animal que anomenem ésser humà."

Malgrat totes les dificultats que tenim per distingir entre llengua i dialecte, la diversitat és enorme: l'edició actual (24ª) d'Ethnologue (Eberhard, Simons, and Fennig, 2021) inclou 7.139 llengües vives. Aquesta diversitat, però, amaga una gran desigualtat: si totes les llengües fossin la llengua materna del mateix nombre de persones, tindríem més d'1,1 milions de parlants per llengua. Tot i això, a l'extrem en perill, aproximadament una cinquena part de les llengües compten amb menys de mil parlants, mentre que a l'extrem vital, les 20 primeres llengües corresponen a més del 70% de la població mundial.

No resulta trivial mesurar això, i **Internet i diversitat lingüística: cibergeografia de les llengües amb el nombre més gran de locutors** per Daniel Pimienta és una obra magistral encara que la franja més àmplia de llengües considerades inclou les 330 llengües "més riques" amb més d'un milió de parlants L1. Com cita Bromham et at. (2021), el principal indicador de que una llengua està en perill és el nombre de parlants com a llengua materna (L1), tot i que hi ha d'altres factors importants. Val la pena recalcar que molts d'aquests factors queden fora del control de la política:

> És més probable que una llengua estigui en perill si també ho està una proporció més alta de llengües a la regió, la qual cosa suggereix que, a més d'amenaces específiques per cada llengua, també existeixen factors generalitzats que afecten la vitalitat lingüística a tota una regió. (Bromham 2021)

D'altres factors, com la densitat de carreteres o la mitjana d'escolarització, només es poden controlar mitjançant polítiques regressives que fomenten una forma de colonialisme inhumana basada en "mantenir-los a la reserva". La millor resposta a això és l'article de Steven Bird **Crear un món que sostingui les seves llengües**, que ofereix un programa més humà "d'accions concretes per ajudar a crear (...) un món que sostingui les seves llengües", que inclou des d'accions quotidianes senzilles com saludar a la gent en la seva pròpia llengua fins a jocs lingüístics més complexos.

Apart de l'indiscutible colonialisme (i el seu principal impulsor, el racisme), també existeix una mena de romanticisme inapropiat, ja observat per Russell (1937), que a vegades serveix per justificar polítiques a favor de les reserves. Es basa en la suposició summament errònia que la forma "més pura" i "més autèntica" de qualsevol llengua es manifesta en el dialecte més aïllat i "més endarrerit" (Kornai 2019).

1.- kornai@sztaki.hu

Conjuntament, els resultats de Bromham et al. (2021) justifiquen perspectives d'àrea com les descrites a **El Projecte Indigenous Languages Technology (ILT) del Consell Nacional d'Investigació de Canadà, i el seu context** de Roland Kuhn i **Tecnologies de la llengua i revitalització lingüística (Rising Voices)** d'Eddie Avila, que identifica vuit estratègies per organitzar l'activisme digital:

1 Facilitar la comunicació digital en llengües indígenes
2 Multiplicar els continguts en llengües indígenes a Internet
3 Normalitzar l'ús de llengües indígenes a Internet
4 Educar en i ensenyar llengües indígenes en línia
5 Recuperar coneixements i llengües indígenes
6 Idear i crear nous mitjans de comunicació en llengües indígenes
7 Defensar els espais per les llengües indígenes i els drets lingüístics
8 Protegir el patrimoni lingüístic i les comunitats

Tot i que no està organitzat seguint la mateixa estructura, l'article de Bird dona exemples concrets de la majoria d'aquestes estratègies des de la perspectiva de l'*activista individual*: "(...) accions concretes per ajudar a crear aquest món del futur (...) No cal unir-se a una campanya".

Pels tecnòlegs, la qüestió clau, tractada de manera palesa a **La diversitat lingüística a l'era de les dades massives** per Tunde Adegbola, és si les "tecnologies de la informació i la comunicació modernes aniran a favor o en contra de la diversitat lingüística". Com destaca Adegbola, no existeix una resposta òbvia, dependrà de "com s'utilitzin". A l'escenari digital, Kornai (2013) argumentava que el millor indicador de vitalitat és la dimensió de la Viquipèdia, i per arribar a tenir-la cal una ortografia normalitzada. És un assumpte subtil, com comenta Cabrera a la seva contribució, però el mecanisme de causa i efecte és evident: per obtenir el tipus d'eines que comenten Kuhn i **OpenSpeaks: Transformar les lliçons d'innovacions tecnològiques a documentacions de llengües amb pocs recursos en Recursos Educatius Oberts** per Subhashish Panigrahi, calen corpus, que només s'aconsegueixen quan les comunitats utilitzen la seva parla i llengua dins de contextos digitals.

En conjunt, aquesta col·lecció recull totes les principals perspectives, des de l'individual a la col·lectiva, des de la de l'àrea a la global, des de la del parlant nadiu a la del tecnòleg lingüísticament agnòstic, i qualsevol persona interessada en la diversitat trobarà una gran recompensa en aquesta edició especial per l'atenció prestada a aquests articles.

Referències

Bromham, Lindell et al. (2021). "Global predictors of language endangerment and the future of linguistic diversity". Nature Ecology & Evolution. issn: 2397-334X. doi: 10.1038/s41559-021-01604-y. url: https://doi.org/10.1038/s41559-021-01604-y

Eberhard, David M., Gary F. Simons, and Charles D. Fennig, eds. (2021). Ethnologue: Languages of the World. 24th ed. SIL International. url: http://www.ethnologue.com

Kornai, András (2013). "Digital language death". PloS ONE 8.10, DOI 10.1371/journal.pone.0077056. url: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0077056

— (2019). The Ni'ihau Paradox: isolation or assimilation. url: https://kornai.com/Slide/lt4all.pdf

Russell, Bertrand (1937). "The Superior Virtue of the Oppressed". *The Nation* 144.

INTERNET AND LINGUISTIC DIVERSITY: THE CYBERGEOGRAPHY OF LANGUAGES WITH THE LARGEST NUMBER OF SPEAKERS

Daniel Pimienta

The *Observatory of Linguistic and Cultural Diversity in the Internet* has recently published the results from its latest study[1], updating and improving on prior work from 2017 aimed at producing indicators of the presence of languages with more than 5 million L1 speakers[2]. In this article, we suggest analysing the data obtained and interpreting its meaning in terms of the **cyber geography of languages**.

First, an explanation regarding the languages selected: choosing to focus on the languages with the largest number of speakers and putting others aside, especially those classified as indigenous languages, within this period that UNESCO has designated, in 2019, as the *year of indigenous languages* (https://en.iyil2019.org/), followed by, the *2022-2032 decade of indigenous languages*[3], is not a political choice. Choosing languages with the largest number of speakers is simply a restriction resulting from the methodology adopted, whereby the bias analysis leads us to conclude that biases would be too high for languages with less than one million L1 speakers.

Before discussing the results, this leads us to briefly present the methodology, main sources and biases that can affect the data produced for the languages considered.

The demo-linguistic source for this new edition is the Ethnologue "Global Dataset #24", from March 2021, without a doubt the most complete source in terms of languages, as well as the most up-to-date and reliable, although we must stress that perfection does not exist in this field and experts may object to some figures. We have also rearranged, according to Ethnologue, some languages into macro-languages[4]. There are 130 languages with L1>5M and the list can be consulted at Pimienta (2021). It should be noted that a new study is under way to eliminate or diminish the remaining biases and extend coverage towards L1>1M, which represents 330 languages, a figure that is closer to the estimated 500 languages present in the Internet. We will use the intermediate results of this last study that have been shared[5].

The detailed methodology, sources, and associated biases, as well as the results are completely documented in Pimienta (2019, 2021). This is an indirect approximation to measuring the presence of languages in the Internet, based on a large number of data sources on languages or countries in the Internet. The data per country have been turned into data per languages using a

1.- Check http://funredes.org/lc2021. This study has focused especially on Portuguese and has been possible thanks to the *Department of Culture and Education of the Ministry of Foreign Affairs of Brazil* within the framework of the International Portuguese Language Institute coordinated by the UNESCO Chair Language Policies for Multilingualism.
2.- We use the term L1 to refer to the native language and L2 for the second language(s).
3.- https://en.unesco.org/news/unesco-launches-global-task-force-making-decade-action-indigenous-languages
4.- Macro-languages are indicated in cursive.
5.- http://funredes.org/lc2021/Results1M.xlsx

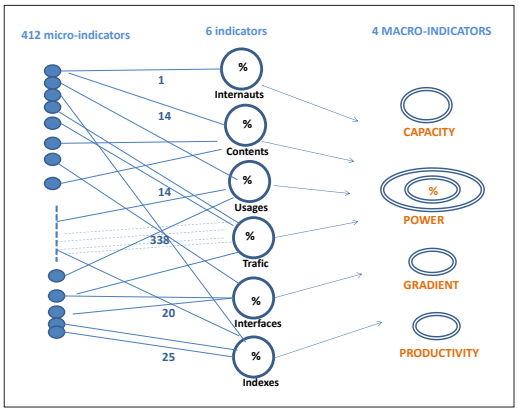
weighting technique with the demo-linguistic data; this is one of the method's biggest innovations[6].

Presence is measured using statistical calculations based on primary sources, in terms of global percentage of L1+L2 population, according to 6 indicators[7]; 4 macro-indicators are calculated based on these 6 indicators:

- **Power**: the average of the 6 indicators (absolute weight of the language in the Internet, which obviously favours languages with more speakers)
- **Capacity**: power divided by the number of speakers (a relative weight that allows us to measure the strength of languages regardless of the number of speakers)
- **Gradient**: power divided by the number of connected speakers (a value that measures the drive of connected speakers when producing content, generating traffic, subscribing to platforms...)
- **Content productivity**: content indicator divided by number of speakers

All data are processed as percentages of the **worldwide number of L1+L2 speakers**, a value that is obviously greater than the worldwide population[8]. According to the latest version of Ethnologue, the global data are the following:

- World population (worldwide total of L1 speakers): 7,231,699,136
- Worldwide total of L1 + L2 speakers: 10,361,716,756
- The "multilingualism global rate" is, therefore, 10,361,716,756 / 7,231,699,136 = 1.4328
(in other words, 43% of the world population is, at least, bilingual).



No statistical method is free from bias, and it is important to identify and analyse these biases and their effects on the results. The distribution of L2 speakers per country that Ethnologue offers now has allowed us to eliminate the method's largest bias, which consisted of extrapolating the results in terms of L1 for L2 (a bias that favoured languages such as English and French with a high L2 population in countries with a low connectivity rate). A significant bias remains in the method, whereby the percentage of people connected to the Internet within a country is considered identical for all existing languages. So, the percentage of Catalan speakers connected to the Internet in Spain is calculated identically to the percentage of Spanish or Arabic speakers, whether L1 or L2, even if the reality is probably different, with some languages with rates above the national av-

6.- Credit to Daniel Prado, who created this concept in 2012.

7.- The 6 indicators are:
Internauts: global percentage of connected speakers
Traffic: percentage of traffic in considered language
Uses: participation percentage on platforms or connectivity resources
Content: percentage of content in considered language
Interfaces: presence of the language in application interfaces or as an online translation language
Indexes: transformation in terms of languages within the range of countries in several parameter classifications related to the information society

8.- In this count, the same person is counted once for their native language and as many times more for each second language.

INDICATOR	SCORE	BIASES
INTERNET USERS	19>16	The main source is the ITU[10]. In 2017, it was the source with the best score (19/20), but in this version, its score falls to 16 because the ITU has stopped providing its own estimate when the country does not produce official data.
INDEX	15>18	This indicator derives from a combination of 25 micro-indicators (in 2017 there was a sole source with 5 parameters). Sources are international organisations, NGOs, or universities.
CONTENT	5>8	There are only 13 micro-indicators to build this indicator, and 11 of them come from excellent Wikimedia statistics. However, Wikimedia does not reflect the real diversity of the Web since it is biased towards a western vision. A weighting system has been implemented to reduce this dependence slightly. It is quite a sensitive bias.
TRAFFIC	13>11	This indicator comes from measuring the traffic per country using Alexa.com on a selection of 338 websites. In 2017, the analysis showed that Alexa was negatively biased against Asian countries and Brazil. In 2021, new biases which are now detrimental to European countries have been detected.
INTERFACES	19	These are objective data. However, it is still a "radical indicator" which omits most of the languages in the world and is focused on an extremely limited subset.
USES	12	This indicator is mainly based on subscription data per country to the most well-known social media (Facebook, Twitter, LinkedIn, etc.) which implies a bias against non-western countries where alternative applications exist.

erage and others below. This bias has been considered acceptable as long as we do not intend to compare languages within a country and languages with a small number of speakers are not processed; in any case, it is important to be aware of it when interpreting the results.

The remaining biases come from the selection of sources to calculate the indicators and are summarised in the table above, scoring each indicator from 0 (unacceptable biases) to 20 (completely free from biases) and showing the changes between 2017 and 2021.

The ongoing study is focused on French, with the support of OIF[10], and aims to reduce the aforementioned biases, seeking alternative ways to reflect the applications from Asian countries that offer similar services to those by Wikimedia or the most well-known social media. Meanwhile, the referenced documents offer results with biases corrected "manually" (but applied only to a small part of the results).

The reader may consider somewhat boring the introduction presenting, before the results, the methodology, sources, and biases. Nevertheless, we consider it our ethical duty to provide the elements to evaluate biases before offering results to avoid the common and terrible practice of using data found in the Internet as truth, without the precautions that analysing its production methodology and biases should imply.

What do the results of the 2021 study tell us compared to the 2017 and previous results[11]?

Let us see the more *powerful* languages first.

9.- The International Telecommunication Union (http://itu.int), the United Nations agency that provides statistics on telecommunications, including the percentage of people connected to the Internet per country.

10.- http://francophonie.org

11.- The Observatory conducted measurement campaigns, using another methodology, between 1998 and 2007, and was forced to interrupt them because the development of search engines (loosing reliability as a scientific tool) has rendered the method obsolete. The results can still be consulted at http://funredes.org/lc.

	POWER	CONTENT
English	25%	30%
Chinese	15%	10%
Spanish	8%	6%
French	3.8%	4.5%
Hindi	3.8%	3%
Portuguese	3.5%	2.8%

Portuguese is closely followed by **Russian** and **Arabic**, in turn followed by **German**, **Japanese**, **Malay**, **Turkish**, **Korean,** and **Bengali**, in this order. The 3 factors that will determine future development are, in order of importance: *demography*, overcoming the *digital divide* and the ability to create *content*. Demography favours Hindi, which will probably cruise past French and could even surpass Spanish in the medium term. Demography will end up favouring Arabic over other nearby languages, including Portuguese, which would consolidate its position in front of Russian.

We can clearly observe that the Internet's centre of gravity is quickly moving from the western world, where it was born and prospered initially, towards Asian languages and Arabic. Demography should favour African languages in the long term, as well as European languages spoken after the colonisation of the African continent, but the African digital divide is still noticeable and slow to decrease compared to the global growth of the Internet. This table, made with the latest results from 330 languages, clearly shows this situation:

The table reads as follows: the study includes 138 African languages; on average, 28.6% of their speakers are connected to the Internet, the group represents 9.15% of the worldwide total of L1 + L2 speakers, however, together they only represent 2.55% of the total weight of the Internet and 5.18% of the connected L1 + L2 population.

Languages of European origin continue to lead the Internet, above the average, but the recent pressure of Asian languages places them, already, in a better position in terms of connected people, and their weight in the Internet is quickly growing. Keeping in mind that the measuring instrument is currently considerably biased against Asian languages, it is likely that the difference in terms of power is significantly less than indicated. In any case, as the connectivity rate of Asian countries rises, and comes closer to the exceptional average rate of more than 80% of European languages, they will also gain first place in terms of power.

The *power* macro-indicator, favours languages with a larger number of speakers by definition. So, let us observe the languages that lead the *capacity* and *gradient* macro-indicators, as well as the most connected languages to have an indication that does not take into account the number of speakers.

	Languages of Africa	Languages of the Americas	Arabic as a macro-language	Languages of Asia	Languages of Europe	Rest of languages
% of Internet users	28.6%	59.7%	60.2%	46.6%	81.1%	54.2%
Power	2.55%	0.19%	3.11%	35.71%	54.93%	3.47%
Capacity	0.24	0.63	0.88	0.57	1.61	0.45
Gradient	0.45	0.59	0.80	0.65	1.07	0.55
L1 + L2 Pop.	9.15%	0.31%	3.53%	48.21%	30.91%	7.81%
Connected population	5.18%	0.32%	3.89%	44.60%	39.51%	6.45%
Number of languages with L1>1M	**138**	**8**	**1**	**135**	**47**	**0**

	CAPACITY
Norwegian	4.65
Hebrew	4.40
Estonian	*3.93*
Finnish	3.49
Serbo-Croatian	3.18
Swedish	2.64
Dutch	2.28
Danish	2.21
Catalan	2.14
Italian	2.11
German	2.11
Macedonian	2.08
Japanese	2.08

The bias resulting from the use of Wikimedia statistics considerably favours languages that have invested in this space (this is the case of Hebrew or Swedish, for example). It is important to consider the list globally, without taking too much notice of the order, and realising that it clearly points towards national languages from countries and regions recognised for **their leadership in the field of information society**.

The most connected languages are the following:

	% OF CONNECTED SPEAKERS
Norwegian	97.87%
Danish	97.82%
Swedish	93.49%
Japanese	92.62%
Dutch	92.03%
Limburgish	91.90%
Swiss German	91.56%
Catalan	90.47%
West Flemish	90.43%
Finnish	89.67%
Estonian	89.10%
Latvian	88.95%
Galician	88.61%
Upper Saxon German	88.13%
German	87.68%
Bavarian	87.68%

And, finally, the languages with the largest gradient:

	GRADIENT
Hebrew	2.82
Norwegian	2.60
Estonian	2.41
Serbo-Croatian	2.24
Finnish	2.13
Malagasy	2.13
English	1.73
Swedish	1.54
Italian	1.53
Macedonian	1.37
Dutch	1.36
German	1.31
Slovenian	1.30
Catalan	1.30
Polish	1.27
Spanish	1.25

The presence of **Malagasy**, a language with an extremely small number of connected people (less than 10%), in this table raises questions about the method. What has happened is that Malagasy appears disproportionately regarding its presence in the Internet in some Wikimedia[12] tables and, due to the averages, the disproportion is so huge that it manages to impact the results. It is one of the symptoms of the content indicator biases that we continue to work on.

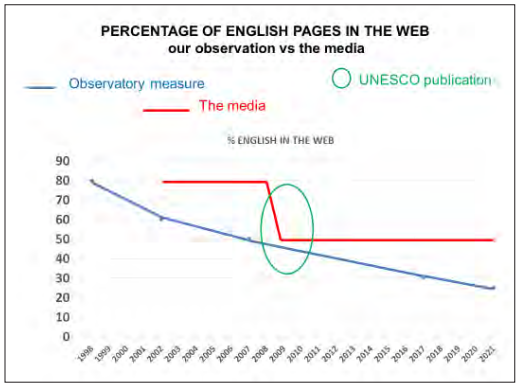
The new languages that rank highly for *capacity* and *gradient* in the edition with 330 languages confirm the 2017 diagnosis: **Norwegian**, **Slovenian**, **Estonian** and **Catalan** are associated to valued countries or regions in terms of information society parameters.

Let us focus on the first languages of the Internet, starting with **English**. The most used, and for a long time the only source of data

12.- Especially the "Wiktionary" (https://en.wiktionary.org/wiki/Wiktionary:Main_Page) where it accounts for 18% of entries.

regarding the presence of languages on the Web, is W3Techs[13].

How is it possible that this source indicates a content percentage in English of 62.5% in 2021 when the Observatory suggests 30%? By measuring the websites directly with a language recognition algorithm, it should not make mistakes. *Well, that's multilingualism, stupid!* If we are allowed to rehash that famous and provocative expression... and address a historical problem of misinformation about the space English occupies in the Internet, illustrated with these two curves:



Until 2009, the media published reports on the presence of English on the Web placing it at 80% with no changes during a decade, meanwhile our measurements indicated a progressive decline towards 50%. The media was supported by 3 publications which suggested, with the same methodology, the same results in 1997, 1999 and 2002. The methodology was not really biased but scientifically invalid[14], see Pimienta (2009) for more details. After UNESCO's publications on the matter (Pimienta 2006, 2009), the media[15] progressively adopted this new 50% value. Then W3Techs appeared as the sole source, whose results maintained a value between 50% and 60% since 2011[16].

How does W3Techs work, and which problem related to multilingualism occurs?

W3Techs selects the 10 million most visited sites on the Web, according to the Alexa digital marketing application (http://alexa.com). Let us brush aside the biases in favour of English of language recognition algorithms and the fact that selecting the 10 million most visited sites (out of the close to 1.2 billion existing websites[17], that is, less than 1%) favours websites in English. Today, analysing the entire Web seems an insurmountable task and we do not aim to disregard the commendable work of W3Techs, which provides many useful data. We focus particularly on how it manages multilingualism. W3Techs applies its algorithm to the **homepage** of these 10 million sites daily. The decision to restrict itself to the homepage, without somehow compensating for this, is part of the problem. Of course, the languages on the Web should be counted at page level and not site level, since a website limited to a single page cannot be counted in the same manner as another with thousands of pages. To this we must add that this website with thousands of pages may also include pages in different languages, even though the homepage may be mainly in English, thus increasing the error threshold. Nowadays, a large part of the

13.- https://w3techs.com/technologies/overview/content_language
14.- A language recognition algorithm was applied a single time to the homepage of 3000 randomly chosen websites, based on IP numbers, and the percentages were calculated. The statistical method to validate this approach would be to repeat the method many times and then analyse the random variable with statistical tools (average, variance, etc.). A single arrow shot at a target does not usually report on the archer's abilities!
15.- And also, unfortunately Wikipedia (https://en.wikipedia.org/wiki/Languages_used_on_the_Internet), from whom we expect more caution.
16.- See https://w3techs.com/technologies/history_overview/content_language/ms/y).
17.- Source: https://news.netcraft.com/archives/category/web-server-survey/

most visited sites (such as Facebook, for example) offer scores of linguistic versions from the homepage; counting the homepage in English is brushing aside all those versions. Finally, it is extremely common for the homepage of a site in a language other than English to have some words in English (for example, keywords or copyright); counting it as an English page, which is probably what happens with the W3Techs algorithm, means leaving out scores of pages in other languages.

One does not need to be a statistics expert to understand that the method, due to not considering the reality of multilingualism, can be hugely mistaken... What could the W3Techs algorithm do to improve its products, without abandoning a pragmatic approach, that is, without facing the challenge of analysing all the pages of all websites?

- Analyse the language options offered on the homepage and count each option as well as the English version.
- Find a method to obtain an approximate estimate of the number of website pages and multiply each linguistic version by that number to count pages instead of sites.
- When the algorithm reports more than one language on the homepage, do not count it as English.

Other factors that should alert us to the improbability of the W3Techs data and draw attention to some symptomatic statistical anomaly of a gross error:

- it makes no sense that the amount of content in English has remained stable for the last 10 years while Asian and Arab countries have invaded the Web during the same period and a set of non-European languages[18] now takes up close to a third;
- the presence of English-speaking Internet users (L1+L2) has gone from 32% in 2017 to 13% today;
- showing Chinese with just 1.3% of content and Hindi with 0.1% when both these languages represent 17.5% and 4.2% of connected people, respectively.

To close this chapter, the fact that the number of websites in English decreases in no way means that the presence of English in absolute terms diminishes, nor that it has stopped growing; it just means that new languages are taking up more and more space, which reduces the proportion of English. Of course, English continues to be the leading language in the Internet, whose estimated amount of content (30%) surpasses the number of Internet users (15%) by a factor of 2.

In Pimienta (2017), we have discussed the biases in different projects and how the lack of consideration for multilingualism can lead to blatant errors. The most typical frequent occurrence is calculating the elements based on L1+L2, divided by the world population, which causes magnitude errors, hidden within the values of the rest of the languages. The number of L1+L2 speakers is much higher than the world population, we had estimated the proportion of multilingual people in 2017 at 25%, in that updated version, Ethnologue offers us a more accurate figure of 43%.

Let us see the languages that follow English now. **Chinese** is in second position in terms of *power* and *content*, but already takes first place in terms of *connected people* in the world and, unlike western countries, where many are above 90%, there is plenty of room to grow. The following table shows data with minimal biases obtained from the ITU data of people connected to

18.- Chinese, Hindi, Arabic, Turkish, Bengali, Vietnamese, Urdu, Persian and Marathi.

the Internet and Ethnologue's demo-linguistic data, weighing the first with the second[19]. The presence of Asian languages and Arabic is noteworthy.

	% OF WORLDWIDE INTERNAUTS	% OF WORLDWIDE SPEAKERS	% OF CONNECTED SPEAKERS
Chinese	17.5%	14.6%	65.59%
English	15.2%	12.9%	64.35%
Spanish	7.0%	5.2%	73.04%
Hindi	4.2%	5.8%	40.18%
Arabic	3.9%	3.5%	60.25%
Russian	3.5%	2.5%	77.21%
Portuguese	3.0%	2.5%	66.96%
French	3.0%	2.6%	63.33%
German	2.2%	1.4%	87.68%
Malay	2.2%	2.3%	51.01%
Japanese	2.0%	1.2%	92.62%
Turkish	1.3%	0.9%	77.95%
Bengali	1.1%	2.6%	24.16%
Urdu	1.0%	2.2%	24.13%
Persian	0.9%	0.8%	63.99%
Vietnamese	0.9%	0.7%	69.00%
Korean	0.9%	0.8%	64.73%
Italian	0.9%	0.7%	75.66%

We suggest this historical analysis of languages in the Internet.

PERIOD	FEATURES
1970-1990	The Internet was born in the West, clearly marked by English in its initial phase, for both technological reasons (the language of network professionals[20]) and due to the nature of its first users (the research world), where a high proportion used English as L2 even if it was not their native language. English dominated during this period.
1990-2010	This period corresponds to the birth of the Web (1992): European languages invested in the Internet, which turned into a privileged space for those languages, and the command of English decreased from 80% to 50% because of the drive of other European languages.
2010-2020	The Internet was both the driving force and the subject of globalisation and the amount of Internet users with English as L1 or L2 rapidly diminished to come closer to the real worldwide percentage, less than 20%. So, its proportion on the Web logically approaches its proportion in the real world, although keeping a historical advantage. However, the African continent continues to fall behind and the digital divide is still huge[21].
2020-2030	We are entering a new phase of globalisation where the demographic weight is starting to be the dominant factor, at least at the heart of the Arab and Asian worlds. In this period, the Internet's linguistic centre of gravity will move towards Asian languages and Arabic, and if the African continent manages to overcome its digital divide, its demography could entail some surprises...

In conclusion, we believe that the idea that English is the lingua franca of the Internet is an illusion: the Internet features more and more **multilingualism[22]** and the digital economy is clearly becoming more influenced by multilingualism.

The fight against misinformation has become the main theme in this healthcare crisis period, where misinformation can lead to death. In keeping with our vision (Pimienta, Rodríguez, 2020), the need to develop broad and thorough **information literacy** programmes is an emergency as acute as global warming. Clearly, these programmes must include citizen education to critically think about the data offered on the Web, and firmly demand methodological and algorith-

19.- And again, focused on L1+L2.

20.- Even in that initial period, computing options were adopted, such as the 7-bit ASCII code which prevented the use of accents and other signs like the diacritical mark, which took several years to overcome.

21.- Of the 56 countries with an Internet connection rate below 30%, 34 are African, 7 Asian, and 6 from the Pacific. Out of these 34 African countries, 14 have less than 10% of people connected.

22.- Indeed, it is the human space where multilingualism is better and more widely expressed, given its boundless characteristics; the Internet's degree of multilingualism could well surpass that of humans, whether in terms of content, traffic, uses or interfaces...

mic transparency, including an honest presentation of biases inherent to all constructed data approaches, whether using statistical or other methods. It is clear that the progress of artificial intelligence, based on intensive data use, makes this need even more critical.

REFERENCES

Pimienta D. (2021) *Enhanced and second version of an alternative approach to produce indicators of languages in the Internet.* Observatory of Linguistic and Cultural Diversity in the Internet
http://funredes.org/lc2021/ALI%20V2-EN.pdf

Pimienta D., Rodríguez L.G. (2020) "¡Va de retro Internet! Una visión crítica de la evolución de la Internet desde la sociedad civil", *Revista Ibero-Americana de Ciência da Informação, V13 N3, Pp. 979-1000* - https://periodicos.unb.br/index.php/RICI/article/view/33041/27497

English versión: http://funredes.org/RockInternetBlues/Va%20de%20retro%20Internet.en.pdf

Pimienta D. (2019) *An alternative approach to produce indicators of languages in the Internet*. Observatory of Linguistic and Cultural Diversity in the Internet
http://funredes.org/lc2017/Alernative%20Languages%20Internet.docx

Pimienta D., Prado D., Blanco A. (2009) Twelve years of measuring linguistic diversity on the Internet: balance and perspectives, *UNESCO, Publications for World Summit on the Information Society*, *CI-2009/WS/1*- http://unesdoc.unesco.org/ulis/cgi-bin/ulis.pl?catno=187016

Pimienta D. (2005) Linguistic Diversity in cyberspace: models for development and measurement", in *Measuring Linguistic Diversity on the Internet, UNESCO, Publications for World Summit on the Information Society, 2005*- http://unesdoc.unesco.org/images/0014/001421/142186e.pdf

INTERNET I DIVERSITAT LINGÜÍSTICA: CIBERGEOGRAFIA DE LES LLENGÜES AMB EL NOMBRE MÉS GRAN DE LOCUTORS

Daniel Pimienta

L'*Observatori de la diversitat lingüística i cultural a l'Internet*[1] acaba de publicar els resultats del seu últim estudi[2], actualitzant i millorant el seu treball previ del 2017 destinat a produir indicadors de la presència de les llengües amb més de 5 milions de locutors L1[3]. A aquest article proposem analitzar les dades obtingudes i interpretar el seu significat en termes de **cibergeografia de llengües**.

Primer, una explicació respecte a les llengües seleccionades: no és una decisió política assumida la d'enfocar les llengües amb el nombre més gran de locutors i deixar de banda les altres, i en particular les llengües qualificades d'indígenes, a aquest període que la UNESCO ha marcat, l'any 2019, com l'*any de les llengües indígenes* (https://es.iyil2019.org/), seguit de la *dècada de les llengües indígenes 2022-2032*[4]. La selecció de llengües amb el nombre més gran de locutors és senzillament una restricció resultant de la metodologia adoptada, per la qual l'anàlisi dels biaixos porta a concloure que aquests serien massa alts per a llengües amb un nombre de parlants L1 per sota del milió.

Això ens porta, abans de discutir els resultats, a presentar breument la metodologia, les fonts principals i els biaixos que poden afectar les dades produïdes per a les llengües considerades.

La font demolingüística per a aquesta nova edició és el "Global Dataset #24" d'Ethnologue, del març del 2021, sens dubte la font més completa en quant a llengües, així com la més actualitzada i fiable, tot i que ha de quedar clar que la perfecció no existeix en aquest terreny i que professionals del camp poden objectar algunes de les xifres. També hem adoptat el reagrupament que fa Ethnologue d'algunes llengües en macrollengües[5]. Les llengües amb L1>5M són 130 i es pot consultar la llista a Pimienta (2021). Convé anotar que hi ha un nou estudi en curs que busca eliminar o disminuir els biaixos restants i estendre la cobertura cap a L1>1M, el que representa 330 llengües, una xifra que s'apropa més al nombre estimat de 500 llengües presents a l'Internet. Utilitzarem els resultats intermedis d'aquest últim estudi que s'han difós[6].

La metodologia detallada, les fonts i els biaixos associats, així com els resultats es troben totalment documentats a Pimienta (2019, 2021). Es tracta d'una aproximació indirecta

---

1.- Preferim utilitzar l'expressió *a l'Internet* enlloc de la recomanada *a Internet* perquè ens sembla que confondre el protocol de comunicació (Internet) amb la xarxa mundial de persones i informació (l'Internet) projecta una història de l'Internet massa simplificada i que amaga moltes contribucions valuoses procedents de xarxes que van existir abans de la convergència protocol·lària (com Bitnet o Usenet, per exemple).

2.- Consulti http://funredes.org/lc2021. Aquest estudi ha enfocat especialment el portuguès i ha estat possible gràcies al recolzament del *Departament de Cultura i Educació del Ministeri de Relacions Exteriors de Brasil* dins el marc de l'Institut Internacional de Llengua Portuguesa i sota la coordinació de la Càtedra UNESCO de Polítiques Lingüístiques per al Multilingüisme.

3.- Utilitzem el terme L1 per referir-nos a la llengua materna i L2 per la o les segones llengües.

4.- https://en.unesco.org/news/unesco-launches-global-task-force-making-decade-action-indigenous-languages

5.- Les macro-llengües estan senyalades en cursiva.

6.- http://funredes.org/lc2021/Results1M.xlsx

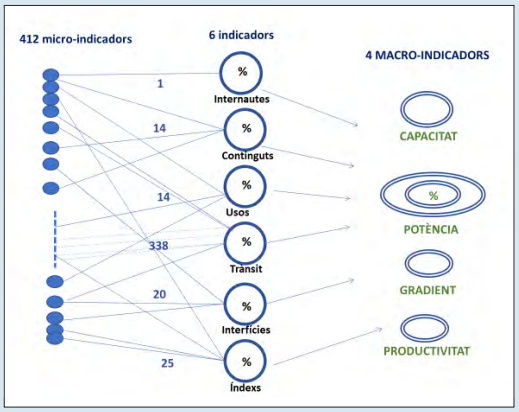
a la mesura de la presència de les llengües a l'Internet, a partir d'un gran nombre de fonts de dades sobre llengües o països a l'Internet. Les dades per país s'han transformat en dades per llengües mitjançant una tècnica de ponderació amb les dades demolingüístiques, essent aquesta una de les innovacions més grans del mètode[7].

La presència es mesura amb càlculs estadístics realitzats a partir de les fonts primàries, respecte el percentatge global sobre la població L1+L2, segons 6 indicadors[8]; a partir d'aquests 6 indicadors es calculen 4 macroindicadors:

- **Potència**: la mitjana dels 6 indicadors (pes absolut de la llengua a Internet, el qual òbviament afavoreix les llengües amb més locutors)
- **Capacitat**: la potència dividida pel nombre de locutors (un pes relatiu que permet mesurar la força de les llengües independentment del seu nombre de locutors)
- **Gradient**: la potència dividida pel nombre de locutors connectats (un valor que mesura el dinamisme dels locutors connectats a l'hora de produir continguts, generar trànsit, subscriure's a plataformes...)
- **Productivitat de continguts**: indicador de continguts dividit pel nombre de locutors

Totes les dades es processen en forma de percentatges sobre el **nombre mundial de locutors L1+L2**, valor òbviament superior a la població mundial[9]. Segons la última versió d'Ethnologue, les dades globals són les següents:

- Població mundial (total mundial de parlants L1): 7.231.699.136
- Total mundial de parlants L1 + L2: 10.361.716.756
- La "taxa mundial de multilingüisme" és, per tant, 10.361.716.756 / 7.231.699.136 = 1,4328
  (en altres paraules, el 43% de la població mundial és, almenys, bilingüe).



Cap mètode estadístic és lliure de biaixos i és molt important identificar i analitzar-los i els seus efectes sobre els resultats. La indicació de la repartició de locutors L2 per país que ofereix ara Ethnologue ha permès eliminar el biaix més gran del mètode que consistia en extrapolar els resultats en termes de L1 per a L2 (biaix que afavoria llengües com l'anglès i francès amb una alta població de L2 a països amb baixa taxa de connectivitat). Queda un biaix notable al mètode, el de considerar que dins d'un país el percentatge de persones connectades a l'Internet és idèntic per a totes les llengües existents. Així, el percentatge de locutors de català connectats a

7.- Crèdits a Daniel Prado, qui va originar aquest concepte l'any 2012.

8.- Els 6 indicadors són:
Internautes: percentatge global de locutors connectats
Trànsit: percentatge de trànsit en la llengua considerada
Usos: percentatge de participació a plataformes o a recursos de connectivitat
Continguts: percentatge de continguts en la llengua considerada
Interfícies: presència de la llengua a interfícies d'aplicacions o com a llengua de traducció en línia
Índexs: transformació pel que fa a llengües del rang dels països a vàries classificacions de paràmetres relacionats amb la societat de la informació

9.- A aquest recompte, la mateixa persona compta un cop per la seva llengua materna i tants cops més per cada segona llengua.

INDICADOR	VALOR	BIAIXOS
INTERNAUTES	19>16	La font principal és la UIT[10]. L'any 2017, va ser la font millor qualificada amb un 19/20, però a aquesta versió la puntuació cau a 16 perquè la UIT ha deixat de proporcionar la seva pròpia estimació quan el país no produeix dades oficials.
ÍNDEX	15>18	Aquest indicador es deriva d'una combinació de 25 micro-indicadors (l'any 2017 hi havia una sola font amb 5 paràmetres). Les fonts són organitzacions internacionals, ONGs o universitats.
CONTINGUTS	5>8	Només hi ha 13 micro-indicadors per construir aquest indicador, 11 dels quals provenen de les excel·lents estadístiques de Wikimedia. Tanmateix, Wikimedia no reflecteix la diversitat real de la Web, estant esbiaixada cap a una visió occidental. S'ha implementat un sistema de ponderació per reduir una mica aquesta dependència. És un biaix bastant sensible.
TRÀNSIT	13>11	Aquest indicador es deriva de la mesura del trànsit per país utilitzant Alexa.com en una selecció de 338 llocs Web. L'any 2017, l'anàlisi va mostrar que Alexa estava esbiaixat negativament cap als països asiàtics i el Brasil. L'any 2021, es detecten nous biaixos que ara perjudiquen als països europeus.
INTERFÍCIES	19	Aquestes són dades objectives. No obstant, segueix essent un "indicador radical" que omet la gran majoria de les llengües del món i se centra en un subconjunt molt limitat.
USOS	12	Aquest indicador es basa principalment en les dades de subscripció per país a les xarxes socials més conegudes (Facebook, Twitter, LinkedIn, etc.), el que implica un biaix contra els països no occidentals on existeixen aplicacions alternatives.

l'Internet a Espanya està calculat de manera idèntica al percentatge de locutors de castellà o d'àrab, siguin L1 o L2, encara que la realitat probablement sigui diferent, amb algunes llengües amb taxes més altes que la mitjana nacional i d'altres amb taxes menors. Aquest biaix es considera acceptable si no es pretén comparar llengües dins d'un país i si no es processen llengües amb un baix nombre de locutors; en qualsevol cas, és important conèixer-lo a l'hora d'interpretar els resultats.

Els altres biaixos resulten de la selecció de fonts pels càlculs dels indicadors i es resumeixen al quadre presentat més amunt, puntuant de 0 (biaixos inacceptables) a 20 (totalment lliure de biaixos) cada indicador i mostrant els canvis entre 2017 i 2021.

L'estudi en curs està enfocat al francès, amb el suport de la OIF[11], i pretén reduir els biaixos esmentats, utilitzant maneres alternatives de reflectir les aplicacions dels països asiàtics que ofereixen serveis similars als de Wikimedia o les xarxes socials més conegudes. Mentrestant, els documents als quals hem fet referència ofereixen uns resultats amb correcció "a mà" de biaixos (però aplicat només a una petita part dels resultats).

El lector pot considerar una mica avorrida aquesta lectura introductòria presentant, abans dels resultats, metodologia, fonts i biaixos. No obstant això, considerem un deure ètic donar els elements d'apreciació dels biaixos abans d'oferir resultats per evitar la pràctica tan comuna i tan nefasta d'utilitzar dades trobades a l'Internet com una realitat, sense la precaució que hauria d'implicar l'anàlisi de la seva metodologia de producció i biaixos.

Què ens diuen els resultats de l'estudi de l'any 2021 en comparació amb els de l'any 2017 i anteriors[12]?

10.- La Unió Internacional de les Telecomunicacions (http://itu.int), l'organisme de les Nacions Unides que proporciona estadístiques sobre telecomunicacions, inclòs el percentatge de persones connectades a l'Internet per país.

11.- http://francophonie.org

12.- L'Observatori ha dut a terme campanyes de mesura, amb una altra metodologia, entre els anys 1998 i 2007, i es va veure obligat a interrompre-les degut a que l'evolució dels motors de cerca (perdent la fiabilitat com a eina científica) ha fet que el mètode esdevingui obsolet. Els resultats encara es poden consultar a http://funredes.org/lc.

Veiem primer les llengües més *potents*.

	POTÈNCIA	CONTINGUTS
Anglès	25%	30,0%
Xinès	15%	10%
Castellà	8%	6%
Francès	3,8%	4,5%
Hindi	3,8%	3,0%
Portuguès	3,5%	2,8%

Després del portuguès, es troben molt a prop el **rus** i l'**àrab**, seguits, en aquest ordre, per l'**alemany**, el **japonès**, el **malai**, el **turc**, el **coreà** i el **bengalí**. Els 3 factors que determinaran l'evolució futura són, per ordre d'importància: la *demografia*, la superació de la *bretxa digital* i la capacitat de crear *continguts*. La demografia afavoreix l'hindi que probablement passarà ràpidament per davant del francès i podria, a mitjà termini, arribar a superar el castellà. La demografia acabarà afavorint l'àrab per sobre d'altres idiomes propers, inclòs el portuguès, que podria afermar la seva posició per davant del rus.

Podem observar clarament que el centre de gravetat de l'Internet s'està movent ràpidament des del món occidental, on va néixer i prosperar a la seva fase inicial, cap a les llengües asiàtiques i l'àrab. La demografia hauria d'afavorir a llarg termini les llengües del continent africà, i també les llengües europees de la colonització africana, però la bretxa digital africana és encara molt marcada i més lenta en reduir-se comparada amb el creixement global de l'Internet. Aquest quadre, realitzat amb els últims resultats de 330 llengües, presenta clarament aquesta situació:

El quadre es llegeix així: hi ha 138 llengües d'Àfrica tractades a l'estudi; de mitjana, 28,6% dels seus locutors estan connectats a l'Internet, el conjunt representa el 9,15% del total mundial de parlants L1 + L2, no obstant, juntes només representen el 2,55% del pes total de l'Internet i el 5,18% de la població L1+L2 connectada.

Les llengües d'origen europeu segueixen liderant a l'Internet, per sobre de la mitjana, però la recent empenta de les llengües asiàtiques les col·loca en millor posició en quant a persones connectades, i el seu pes a l'Internet va creixent ràpidament. Recordant que l'instrument de mesura queda per ara notablement esbiaixat contra les llengües asiàtiques, és probable que la diferència en quant a potència sigui bastant inferior a l'indicat. En qualsevol cas, a mesura que la taxa de connectivitat dels països asiàtics vagi creixent, i apropant-se a l'excepcional taxa mitjana de més del 80% de les llengües europees, accediran també al primer lloc en quant a potència.

El macroindicador *potència*, per definició, afavoreix les llengües amb un nombre més gran de parlants. Observem doncs les llengües que lideren els macroindicadors *capacitat* i *gradient*, així com les llengües més connectades per tenir una indicació que no tingui en compte el nombre de locutors.

	Llengües d'Àfrica	**Llengües de les Amèriques**	Àrab com a macrollengua	**Llengües d'Àsia**	**Llengües d'Europa**	**Resta de llengües**
Internautes %	28,6%	59,7%	60,2%	46,6%	81,1%	54,2%
Potència	2,55%	0,19%	3,11%	35,71%	54,93%	3,47%
Capacitat	0,24	0,63	0,88	0,57	1,61	0,45
Gradient	0,45	0,59	0,80	0,65	1,07	0,55
Població L1+L2	9,15%	0,31%	3,53%	48,21%	30,91%	7,81%
Població connectada	5,18%	0,32%	3,89%	44,60%	39,51%	6,45%
Nombre de llengües amb L1>1M	**138**	**8**	**1**	**135**	**47**	**0**

	CAPACITAT
Noruec	4,65
Hebreu	4,40
Estonià	*3,93*
Finlandès	3,49
Serbocroat	3,18
Suec	2,64
Holandès	2,28
Danès	2,21
Català	2,14
Italià	2,11
Alemany	2,11
Macedoni	2,08
Japonès	2,08

El biaix resultant de l'ús de les estadístiques de Wikimedia afavoreix sensiblement les llengües que han invertit en aquest espai (el cas de l'hebreu o el suec, per exemple). És important considerar la llista globalment, sense confiar massa en l'ordre, i descobrir que clarament apunta a les llengües nacionals de països i regions reconeguts pel seu lideratge al camp de la societat de la informació.

Les llengües més connectades són les següents:

	% LOCUTORS CONNECTATS
Noruec	97,87%
Danès	97,82%
Suec	93,49%
Japonès	92,62%
Holandès	92,03%
Limburguès	91,90%
Suís Alemany	91,56%
Català	90,47%
Flamenc occidental	90,43%
Finlandès	89,67%
Estonià	89,10%
Letó	88,95%
Gallec	88,61%
Saxó superior	88,13%
Alemany	87,68%
Bavarès	87,68%

I finalment, les llengües amb un gradient més gran:

	GRADIENT
Hebreu	2,82
Noruec	2,60
Estonià	2,41
Serbocroat	2,24
Finlandès	2,13
Malgaix	2,13
Anglès	1,73
Suec	1,54
Italià	1,53
Macedoni	1,37
Holandès	1,36
Alemany	1,31
Eslovè	1,30
Català	1,30
Polonès	1,27
Castellà	1,25

La presència a aquest quadre del **malgaix**, una llengua amb un nombre de persones connectades summament baix (menys del 10%), planteja interrogants sobre el mètode. Passa que el malgaix té una presència, en alguns dels quadres de Wikimedia[13], altament desproporcionada en relació a la seva presència a l'Internet i, pel joc de les mitjanes, la desproporció és tan enorme que aconsegueix impactar els resultats. És un dels símptomes dels biaixos de l'indicador de continguts sobre el qual seguim treballant.

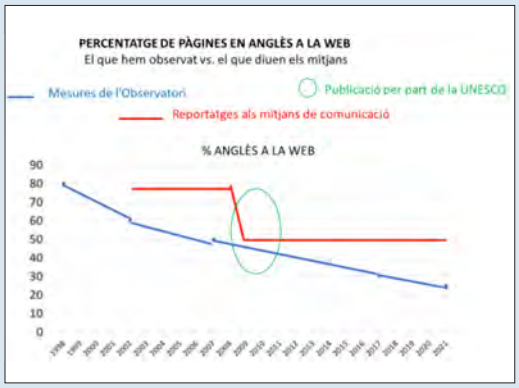
Les noves llengües que apareixen altes en *capacitat* i *gradient* a l'edició amb 330 llengües confirmen el diagnòstic del 2017: el **noruec**, l'**eslovè**, l'**estonià** i el **català** estan associats a països o regions cotitzades als paràmetres de la societat de la informació.

Ens centrem ara en les primeres llengües de l'Internet, començant per l'**anglès**. La font de dades sobre la presència de llengües a la

13.- Especialment el "Wiktionary" (https://es.wiktionary.org/wiki/Wikcionario:Portada) on arriba a tenir el 18% del total d'entrades.

Web més utilitzada, i durant molt de temps la única, és W3Techs[14].

Com és possible que aquesta font indiqui un percentatge de continguts en anglès del 62,5% l'any 2021 quan l'Observatori proposa la xifra del 30%? Mesurant directament les pàgines web amb un algoritme de reconeixement de llengües, no haurien d'equivocar-se. *Doncs sí, és el multilingüisme, estúpid*! Si se'ns permet reutilitzar aquesta famosa i provocativa expressió... i tractar un problema històric de desinformació sobre l'espai de l'anglès a l'Internet, il·lustrat amb aquestes dues corbes:



Fins l'any 2009, els mitjans publicàven informes sobre la presència de l'anglès a la Web posicionant-la, sense canvis durant una dècada, al 80%, mentre que les nostres mesures indicàven un declivi progressiu cap al 50%. Els mitjans es recolzaven en 3 publicacions que proposaven, amb la mateixa metodologia, els mateixos resultats, els anys 1997, 1999 i 2002. La metodologia no era realment esbiaixada, sinó científicament invàlida[15], veure Pimienta (2009) per a més detalls. Després de les publicacions de la UNESCO sobre el tema (Pimienta 2006, 2009), els mitjans de comunicació[16] van adoptar progressivament aquest nou valor del 50%. Després va aparèixer W3Techs com a única font, els resultats de la qual van mantenir un valor entre 50% i 60% des de l'any 2011[17].

Com procedeix W3Techs, i quin és el problema que ocorre en relació amb el multilingüisme?

W3Techs selecciona els 10 milions de llocs més visitats de la Web, segons indica l'aplicació de màrqueting digital Alexa (http://alexa.com). Deixem de banda els biaixos a favor de l'anglès dels algoritmes de reconeixement de llengua i el fet de que seleccionar els 10 milions de llocs més visitats (sobre els prop de 1,2 bilions de llocs Web existents[18], és a dir, menys de l'1%) afavoreix els llocs en anglès. Avui dia, analitzar la Web sencera sembla una tasca inassolible i la nostra intenció no és menysprear la feina encomiable de W3Techs, que proporciona moltes dades molt útils. Ens centrem sobretot en la seva gestió del multilingüisme. W3Techs aplica el seu algoritme, a diari, sobre la **pàgina d'entrada** d'aquests 10 milions de llocs. La decisió de limitar-se a la pàgina d'entrada, sense compensar-la d'alguna manera, forma part del problema. Per descomptat, la comptabilitat de les llengües a la Web hauria de fer-se a nivell de pàgines i no a nivell de llocs, ja que no es pot comptabilitzar igual un lloc web limitat a una pàgina i un altre

14.- https://w3techs.com/technologies/overview/content_language
15.- S'aplicava un algoritme de reconeixement de llengües sobre la pàgina d'entrada de 3000 llocs triats a l'atzar un únic cop a partir de números d'IP, i es calculaven els percentatges. El mètode estadístic per a validar aquesta aproximació seria repetir un gran nombre de vegades el mètode i llavors analitzar la variable aleatòria amb eines estadístiques (mitjana, variància, etc.). Un sol tir amb arc sobre una diana no acostuma a informar sobre la capacitat del tirador!
16.- I també malauradament Wikipedia (https://en.wikipedia.org/wiki/Languages_used_on_the_Internet) de qui s'espera més cautela.
17.- Veure https://w3techs.com/technologies/history_overview/content_language/ms/y).
18.- Font: https://news.netcraft.com/archives/category/web-server-survey/

amb milers de pàgines. A això s'afegeix la possibilitat de que aquell lloc web amb milers de pàgines tingui pàgines en diferents llengües, tot i que la pàgina d'entrada estigui principalment en anglès, incrementant així la dimensió de l'error. A dia d'avui, bona part dels llocs més visitats (com Facebook, per exemple) ofereixen, des de la pàgina d'entrada, desenes de versions lingüístiques; comptabilitzar la pàgina d'entrada en anglès és deixar de banda totes aquestes versions. Finalment, és molt comú que la pàgina d'entrada d'un lloc en una llengua diferent de l'anglès tingui algunes paraules en anglès (per exemple, de navegació o de copyright); comptabilitzar-la com a pàgina en anglès, com probablement passa amb l'algoritme de W3Techs, és obviar desenes de pàgines en altres idiomes.

No cal ser un expert en estadística per entendre que el mètode, per no considerar la realitat del multilingüisme, es pot equivocar en proporcions gegants... Què podria fer l'algoritme de W3Techs per a millorar els seus productes, sense abandonar un enfocament pragmàtic, és a dir, sense entrar al repte d'analitzar totes les pàgines de tots els llocs?

- Analitzar les opcions de llengües ofertes a la pàgina d'entrada i comptabilitzar cada opció igual que la versió anglesa.
- Trobar un mètode per obtenir una estimació, encara que sigui aproximativa, del nombre de pàgines del lloc web i multiplicar cada versió lingüística per aquest nombre per tenir una comptabilitat de pàgines i no de llocs.
- Quan l'algoritme reporta més d'una llengua a la pàgina d'entrada, per principi, no comptabilitzar-la com anglès.

Altres factors haurien d'alertar-nos sobre la inversemblança de les dades de W3Techs i cridar l'atenció sobre alguna anomalia estadística simptomàtica d'un error gros:

- no té sentit que la proporció de continguts en anglès s'hagi mantingut estable durant els últims 10 anys mentre que durant el mateix període els països asiàtics i àrabs han envaït la Web i que un conjunt de llengües no europees[19] ara n'ocupen prop d'un terç;
- la presència d'internautes anglòfons (L1+L2) ha passat del 32% l'any 2017 al 13% avui dia;
- mostrar el xinès amb només l'1,3% dels continguts i l'hindi amb el 0,1% quan aquestes dues llengües representen respectivament el 17,5% i el 4,2% de les persones connectades.

Per tancar aquest capítol, que la proporció de pàgines Web en anglès disminueixi no significa de cap manera que la presència en termes absoluts de l'anglès disminueixi, ni tampoc que hagi acabat de créixer; només significa que noves llengües estan ocupant cada cop més espai, cosa que redueix la proporció de l'anglès. Per descomptat l'anglès segueix essent una llengua líder a l'Internet, de la qual la proporció estimada de continguts (30%) supera en un factor 2 la proporció d'internautes (15%).

Hem dissertat a Pimienta (2017) dels biaixos a diferents projectes i com la falta de consideració del multilingüisme pot dur a errors flagrants. El cas freqüent més típic és el càlcul d'elements basats sobre L1+L2, dividint per la població mundial, cosa que provoca errors de magnitud, amagats entre els valors de la resta de llengües. El nombre de parlants L1+L2 és molt superior a la població mundial, havíem estimat en 25% la proporció de persones multilingües l'any 2017, en aquesta nova versió, Ethnologue ens ofereix una xifra més encertada del 43%.

19.- El xinès, l'hindi, l'àrab, el turc, el bengalí, el vietnamita, l'urdú, el persa i el marathi.

Observem ara les llengües que segueixen a l'anglès. El **xinès** està en segona posició en quant a *potència* i *continguts* però ja ocupa el primer lloc en quant a *persones connectades* al món i, a diferència dels països occidentals, on molts es troben per sobre del 90%, queda espai per progressar. La taula següent mostra dades amb biaixos mínims que s'obtenen a partir de les dades de la UIT de persones connectades a l'Internet i de les dades demolingüístiques d'Ethnologue, ponderant els primers amb els segons[20]. La presència de les llengües asiàtiques i de l'àrab és notable.

	% MUNDIAL INTERNAUTES	% MUNDIAL LOCUTORS	% LOCUTORS CONNECTATS
Xinès	17,5%	14,6%	65,59%
Anglès	15,2%	12,9%	64,35%
Castellà	7,0%	5,2%	73,04%
Hindi	4,2%	5,8%	40,18%
Àrab	3,9%	3,5%	60,25%
Rus	3,5%	2,5%	77,21%
Portuguès	3,0%	2,5%	66,96%
Francès	3,0%	2,6%	63,33%
Alemany	2,2%	1,4%	87,68%
Malai	2,2%	2,3%	51,01%
Japonès	2,0%	1,2%	92,62%
Turc	1,3%	0,9%	77,95%
Bengalí	1,1%	2,6%	24,16%
Urdú	1,0%	2,2%	24,13%
Persa	0,9%	0,8%	63,99%
Vietnamita	0,9%	0,7%	69,00%
Coreà	0,9%	0,8%	64,73%
Italià	0,9%	0,7%	75,66%

Proposem aquest anàlisi històric de les llengües a l'Internet.

PERÍODE	CARACTERÍSTIQUES
1970-1990	L'Internet va néixer al món occidental, molt marcada per la llengua anglesa a la seva fase històrica inicial, tant per raons tecnològiques (la llengua dels professionals de les xarxes[21]) com per la naturalesa dels seus primers usuaris (el món de la investigació), on una alta proporció utilitza l'anglès com a L2 encara que no sigui la seva llengua materna. L'anglès va dominar la xarxa durant aquest període.
1990-2010	Aquest període correspon al naixement de la Web (1992): les llengües europees van invertir a l'Internet, la qual es va transformar en un espai privilegiat per aquestes llengües, amb un domini de l'anglès que va disminuir del 80% al 50% com a resultat de l'empenta de les altres llengües europees.
2010-2020	L'Internet va ser alhora motor i subjecte de la globalització i la proporció d'internautes que tenien l'anglès com a L1 o L2 va disminuir ràpidament per apropar-se al percentatge mundial real, inferior al 20%. Així, la seva proporció a la Web s'apropa lògicament a la seva proporció al món real, tot i guardar un avantatge històric. No obstant, el continent africà segueix endarrerit i la bretxa digital és encara enorme[22].
2020-2030	Entrem a una nova fase de globalització on el pes demogràfic comença a ser el factor dominant, almenys al si del món àrab i asiàtic. En aquest període, el centre de gravetat lingüístic de l'Internet passarà cap a les llengües asiàtiques i l'àrab, i si el continent africà aconsegueix superar la seva bretxa digital, la seva demografia podria reservar sorpreses...

Com a conclusió, pensem que la creença segons la qual la llengua franca de l'Internet és l'anglès és una il·lusió: el que caracteritza l'Internet és cada cop més el **multilingüisme[23]** i l'economia digital clarament està cada dia més caracteritzada pel factor del multilingüisme.

La lluita contra la desinformació s'ha transformat en un tema principal en aquest

20.- I de nou centrat en L1+L2.

21.- En aquell període inicial, fins i tot es van adoptar opcions informàtiques, com el codi ASCII de 7 bits que impedia l'ús dels accents i altres signes com la titlla, que van trigar diversos anys a superar-se.

22.- Dels 56 països amb taxa de connexió a l'Internet inferior al 30%, 34 són africans, 7 asiàtics i 6 del Pacífic. D'aquests 34 països africans, 14 tenen menys del 10% de persones connectades.

23.- Per descomptat, és l'espai humà on el multilingüisme s'expressa de la millor manera i en més mesura, donades les seves característiques sense fronteres; el grau de multilingüisme de l'Internet bé podria ser superior al dels humans, ja sigui en termes de continguts, de trànsit, d'usos o d'interfícies…

període de crisi sanitària on la desinformació pot conduir a la mort. D'acord amb la nostra visió (Pimienta, Rodríguez, 2020), la necessitat de desenvolupar programes amplis i complets d'alfabetització informacional és una emergència tan aguda com la de l'escalfament global. Clarament, aquests programes han d'incloure l'educació de la ciutadania per bregar amb les dades que s'ofereixen a la Web, amb una ment crítica, i una exigència ferma sobre la transparència metodològica i algorítmica, incloent una presentació honesta dels biaixos inherents a tota aproximació de dades construïdes, sigui amb mètodes estadístics o d'altres. És evident que els progressos de la intel·ligència artificial, basada en un ús intensiu de les dades, fan que aquesta necessitat sigui encara més crítica.

REFERÈNCIES

Pimienta D. (2021) *Versión nueva y mejorada de un enfoque alternativo para la producción de indicadores lingüísticos en la Internet.* Observatorio de la diversidad lingüística y cultural en la Internet http://funredes.org/lc2021/ALI%20V2-ES.pdf

Pimienta D., Rodríguez L.G. (2020) "¡Va de retro Internet! Una visión crítica de la evolución de la Internet desde la sociedad civil", *Revista Ibero-Americana de Ciência da Informação, V13 N3, Pp. 979-1000* - https://periodicos.unb.br/index.php/RICI/article/view/33041/27497

Pimienta D. (2019) *Un enfoque alternativo para producir indicadores de la presencia de las lenguas en la Internet.* Observatorio de la diversidad lingüística y cultural en la Internet http://funredes.org/lc2019/Alternativa%20Lengua%20Internet.docx

Pimienta D., Prado D., Blanco A. (2009) Twelve years of measuring linguistic diversity on the Internet: balance and perspectives, *UNESCO, Publications for World Summit on the Information Society*, *CI-2009/WS/1*- http://unesdoc.unesco.org/ulis/cgi-bin/ulis.pl?catno=187016

Pimienta D. (2005) Linguistic Diversity in cyberspace: models for development and measurement", in *Measuring Linguistic Diversity on the Internet, UNESCO, Publications for World Summit on the Information Society, 2005*- http://unesdoc.unesco.org/images/0014/001421/142186e.pdf

INTERNET Y DIVERSIDAD LINGÜÍSTICA: CIBER-GEOGRAFÍA DE LAS LENGUAS CON MAYOR NÚMERO DE LOCUTORES.

Daniel Pimienta

El *Observatorio de la diversidad lingüística y cultural en la Internet*[1] acaba de publicar los resultados de su último estudio[2], actualizando y mejorando su trabajo previo del 2017 destinado a producir indicadores de la presencia de las lenguas de más de 5 millones de locutores L1[3]. Proponemos en este artículo analizar los datos obtenidos e interpretar su significado en términos de **ciber-geografía de lenguas.**

Primero, una explicación en cuanto a las lenguas seleccionadas: no es una decisión política asumida la de enfocar las lenguas con mayor número de locutores y dejar de lado las otras, y en particular las lenguas calificadas de indígenas, en ese periodo que la UNESCO ha marcado, en 2019, como *año de las lenguas indígenas* (https://es.iyil2019.org/), seguido de, la *década de las lenguas indígenas 2022-2032*[4]. La selección de las lenguas con mayor número de locutores es sencillamente una restricción resultado de la metodología adoptada por la cual el análisis de los sesgos conduce a concluir que éstos serían demasiado altos para lenguas con número de hablantes inferior al millón.

Eso nos conduce, antes de discutir los resultados, a presentar brevemente la metodología, las fuentes principales y los sesgos que pueden afectar los datos producidos para las lenguas consideradas.

La fuente demo-lingüística para esta nueva edición es el "Global Dataset #24" de Ethnologue, de marzo 2021, sin duda la fuente más completa en términos de lenguas, así como la más actualizada y confiable, aunque debe quedar claro que la perfección no existe en este terreno y que profesionales del campo podrán objetar algunas cifras. Hemos también adoptado de Ethnologue el reagrupamiento de algunas lenguas en macro-lenguas[5]. Las lenguas con L1>5 M son 130 y se puede consultar la lista en Pimienta (2021). Es de notar que está en curso un nuevo estudio que busca eliminar o disminuir los sesgos restantes y extender la cobertura hacia L1 >1M, lo que representa 330 lenguas, una cifra que se acerca más al estimado número de 500 lenguas presentes en la Internet. Utilizaremos los resultados intermedios de ese último estudio que se han difundido[6].

La metodología detallada, las fuentes y los sesgos asociados, así como los resultados están totalmente documentados en Pimienta (2019, 2021). Se trata de una aproximación

---

1.- Preferimos usar la forma de expresión *en la Internet* en vez de la recomendada *en Internet* porque nos parece que confundir el protocolo de comunicación (Internet) con la red mundial de personas y de información (la Internet) proyecta una historia de la Internet demasiado simplificada y que esconde muchas contribuciones valiosas procedentes de redes que existieron antes de la convergencia protocolar (como Bitnet o Usenet, por ejemplo).

2.- Consulte http://funredes.org/lc2021. Ese estudio ha enfocado especialmente el portugués y ha sido posible gracias al apoyo del *Departamento de Cultura y Educación del Ministerio de Relaciones Exteriores de Brasil* en el marco del Instituto Internacional de Lengua Portuguesa y bajo la coordinación de la Cátedra UNESCO de Políticas Lingüísticas para el Multilingüismo.

3.- Usamos la terminología L1 para referirnos a la lengua materna y L2 para la o las lenguas segundas.

4.- https://en.unesco.org/news/unesco-launches-global-task-force-making-decade-action-indigenous-languages

5.- Las macro-lenguas están señaladas en cursiva.

6.- http://funredes.org/lc2021/Results1M.xlsx

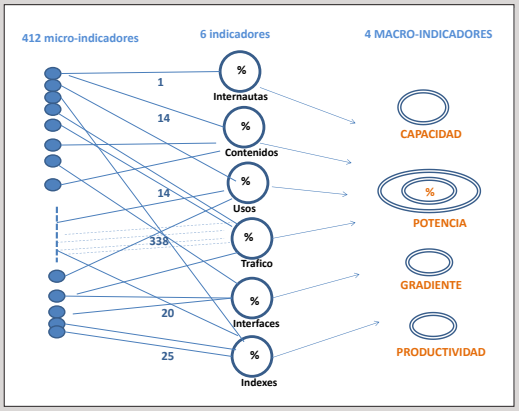
indirecta a la medición de la presencia de las lenguas en la Internet, a partir de un gran número de fuentes de datos sobre lenguas o países en la Internet. Los datos por país han sido transformados en datos por lenguas por una técnica de ponderación con los datos demo-lingüísticos, siendo esta una de las originalidades mayores del método[7].

La presencia se mide con cálculos estadísticos realizados a partir de las fuentes primarias, en términos de porcentaje global sobre la población L1+L2, según 6 indicadores[8]; a partir de esos 6 indicadores se calculan 4 macro-indicadores:

- **Potencia**: el promedio de los 6 indicadores (peso absoluto de la lengua en Internet, el cual obviamente favorece a las lenguas con más locutores)
- **Capacidad**: la potencia dividida por el número de locutores (un peso relativo que permite medir la fuerza de las lenguas independientemente de su número de locutores)
- **Gradiente**: la potencia dividido por el número de locutores conectados (un valor que mide el dinamismo de los locutores conectados a la hora de producir contenidos, generar tráfico, suscribirse a plataformas…)
- **Productividad de contenidos**: Indicador de contenidos dividido por número de locutores.

Todos los datos están procesados en forma de porcentajes sobre el **número mundial de locutores L1+L2**, valor obviamente superior a la población mundial[9]. Según la última versión de Ethnologue, los datos globales son los siguientes:

- Población mundial (total mundial de hablantes L1): 7.231.699.136
- Total mundial de hablantes L1 + L2: 10.361.716.756
- La "tasa mundial de multilingüismo" es, por tanto, 10.361.716.756 / 7.231.699.136 = 1,4328
  (en otras palabras, el 43% de la población mundial es, al menos, bilingüe).



Ningún método estadístico está libre de sesgos y es muy importante identificar y analizar esos sesgos y sus efectos en los resultados. La indicación de la repartición de locutores L2 por país que ofrece ahora Ethnologue ha permitido eliminar el mayor sesgo del método que consistía en extrapolar los resultados en términos de L1 para L2 (sesgo que favorecía lenguas como inglés y francés con alta población de L2 en países con baja tasa de conectividad). Queda un sesgo notable en el

7.- Créditos a Daniel Prado quien originó este concepto en 2012.

8.- Los 6 indicadores son:
Internautas: porcentaje global de locutores conectados
Tráfico: porcentaje de tráfico en la lengua considerada
Usos: porcentaje de participación en plataformas o en recursos de conectividad
Contenidos: porcentaje de contenidos en la lengua considerada
Interfaces: presencia de la lengua en interfaces de aplicaciones o como lengua de traducción en línea
Índices: transformación en términos de lenguas del rango de los países en varias clasificaciones de parámetros relacionados con la sociedad de la información.

9.- En este recuento, la misma persona cuenta una vez por su lengua materna y tantas veces más por cada lengua segunda.

INDICADOR	VALOR	SESGOS
INTERNAUTAS	19>16	La fuente principal es la UIT[10]. En 2017, fue la fuente mejor calificada con un 19/20, pero en esta versión la puntuación cae a 16 porque la UIT ha dejado de proporcionar su propia estimación cuando el país no produce datos oficiales.
ÍNDICE	15>18	Este indicador se deriva de una combinación de 25 microindicadores (en 2017 había una sola fuente con 5 parámetros). Las fuentes son organizaciones internacionales, ONGs o universidades.
CONTENIDOS	5>8	Solo hay 13 microindicadores para construir este indicador y 11 de ellos provienen de las excelentes estadísticas de Wikimedia. Sin embargo, Wikimedia no refleja la diversidad real de la Web, siendo sesgado hacia una visión occidental. Se ha implementado un sistema de ponderación para reducir un poco esta dependencia. Es un sesgo bastante sensible.
TRÁFICO	13>11	Este indicador se deriva de la medición del tráfico por país utilizando Alexa.com en una selección de 338 sitios en la Web. En 2017, el análisis mostró que Alexa estaba negativamente sesgado hacia los países asiáticos y Brasil. En 2021, se detectan nuevos sesgos que ahora perjudican a los países europeos.
INTERFACES	19	Estos son datos objetivos. Sin embargo, sigue siendo un "indicador radical" que omite la gran mayoría de las lenguas del mundo y se centra en un subconjunto muy limitado.
USOS	12	Este indicador se basa principalmente en los datos de suscripción por país a las redes sociales más conocidas (Facebook, Twitter, Linkedin, etc.) lo que implica un sesgo contra los países no occidentales donde existen aplicaciones alternativas.

método, el de considerar que dentro de un país el porcentaje de personas conectadas a la Internet es idéntico para todas las lenguas existentes. Así el porcentaje de locutores de catalán conectados a la Internet en España está calculado de manera idéntica al porcentaje de locutores del castellano o del árabe, sean L1 o L2, aunque la realidad es probablemente distinta, con algunas lenguas con tasas mayores que el promedio nacional y otras con tasas menores. Este sesgo ha sido considerado aceptable si no se pretende comparar lenguas dentro de un país y si no se procesan lenguas con números bajos de locutores; en cualquier caso es importante conocerlo a la hora de interpretar los resultados.

Los demás sesgos resultan de la selección de fuentes para los cálculos de los indicadores y están resumidos en el cuadro superior, puntuando de 0 (sesgos inaceptables) a 20 (totalmente libre de sesgos) cada indicador y mostrando los cambios entre 2017 y 2021.

El estudio en curso está enfocado al francés, con el apoyo de la OIF[11], y pretende reducir los sesgos mencionados, tratando maneras alternativas de reflejar las aplicaciones de los países asiáticos que ofrecen servicios similares a los de Wikimedia o las redes sociales más conocidas. Mientras tanto, los documentos referenciados ofrecen unos resultados con corrección "a mano" de sesgos (pero aplicado sólo a una parte pequeña de los resultados).

El lector puede considerar algo aburrida esta lectura introductoria presentando, antes de los resultados, metodología, fuentes y sesgos. Sin embargo, consideramos un deber ético dar los elementos de apreciación de los sesgos antes de ofrecer resultados para evitar la práctica tan común y tan nefasta de usar datos encontrados en la Internet como una realidad, sin la precaución que debería implicar el análisis de su metodología de producción y sesgos.

10.- La Unión Internacional de las Telecomunicaciones (http://itu.int), el organismo de las Naciones Unidas que proporciona estadísticas sobre telecomunicaciones, incluido el porcentaje de personas conectadas a la Internet por país.

11.- http://francophonie.org

¿Que nos dicen los resultados del estudio del 2021 en comparación con los del 2017 y anteriores[12]?

Veamos primero las lenguas más *potentes*.

	POTENCIA	CONTENIDOS
Inglés	25%	30,0%
Chino	15%	10%
Español	8%	6%
Francés	3,8%	4,5%
Hindi	3,8%	3,0%
Portugués	3,5%	2,8%

Después del portugués, muy cerca están el **ruso** y el **árabe**, seguidos, en este orden, por **alemán**, **japonés**, **malayo**, **turco**, **coreano** y **bengalí**. Los 3 factores que determinarán la evolución futura son, por orden de importancia: la *demografía*, la superación de *la brecha digital* y la capacidad de crear *contenidos*. La demografía favorece al hindi que probablemente pasará rápidamente por delante del francés y podría, a medio plazo, llegar a superar al español. La demografía terminará favoreciendo al árabe sobre otros idiomas cercanos, incluido el portugués, que podría afianzar su posición frente al ruso.

Podemos observar claramente que el centro de gravedad de Internet se está moviendo rápidamente desde el mundo occidental, donde nació y prosperó en su fase inicial, hacia las lenguas asiáticas y el árabe. La demografía debería favorecer a largo plazo a las lenguas del continente africano, y también a las lenguas europeas de la colonización africana, pero la brecha digital africana es aún muy marcada y más lenta en reducirse comparada con el crecimiento global de la Internet. Este cuadro, realizado con los últimos resultados de 330 lenguas, presenta claramente esta situación:

El cuadro se lee así: hay 138 lenguas de África tratadas en el estudio; en promedio, 28.6 % de sus locutores están conectados a la Internet, el conjunto representa el 9.15% del total mundial de hablantes L1+L2, sin embargo, juntas solo representan el 2.55% del peso total de la Internet y el 5.18% de la población L1+L2 conectada.

Las lenguas de origen europeo siguen liderando en la Internet, por encima del promedio, pero el reciente empuje de las lenguas asiáticas las coloca, desde ya, en mejor posición en términos de personas conectadas, y su peso en la Internet va creciendo rápidamente. Recordando que el instrumento de medición queda por el momento notablemente sesgado contra las lenguas asiáticas, es probable que la diferencia en términos de

	Lenguas de África	**Lenguas de las Américas**	Árabe como macro-lengua	**Lenguas de Asia**	**Lenguas de Europa**	**Resto de lenguas**
Internautas %	28,6%	59,7%	60,2%	46,6%	81,1%	54,2%
Potencia	2,55%	0,19%	3,11%	35,71%	54,93%	3,47%
Capacidad	0,24	0,63	0,88	0,57	1,61	0,45
Gradiente	0,45	0,59	0,80	0,65	1,07	0,55
Población L1+L2	9,15%	0,31%	3,53%	48,21%	30,91%	7,81%
Población conectada	5,18%	0,32%	3,89%	44,60%	39,51%	6,45%
Numero de lenguas con L1>1M	**138**	**8**	**1**	**135**	**47**	**0**

12.- El Observatorio ha conducido campañas de medición, con otra metodología, entre 1998 y 2007, estando obligado a interrumpirlas debido a que la evolución de los motores de búsqueda (perdiendo confiabilidad como herramienta científica) ha hecho obsoleto el método. Los resultados siguen siendo consultables en http://funredes.org/lc.

potencia sea bastante inferior a lo indicado. En cualquier caso, en la medida que la tasa de conectividad de los países asiáticos vaya elevándose, y acercándose a la excepcional tasa promedio de más del 80% de las lenguas europeas, accederán también al primer lugar en términos de potencia.

El macroindicador *potencia,* por definición, favorece a las lenguas con mayor número de hablantes. Observemos entonces las lenguas que lideran los macroindicadores *capacidad* y *gradiente,* así como las lenguas más conectadas para tener una indicación que no tenga en cuenta el número de locutores.

	CAPACIDAD
Noruego	4,65
Hebreo	4,40
Estonio	*3,93*
Finlandés	3,49
Serbocroata	3,18
Sueco	2,64
Holandés	2,28
Danés	2,21
Catalán	2,14
Italiano	2,11
Alemán	2,11
Macedonio	2,08
Japonés	2,08

El sesgo resultado del uso de las estadísticas de Wikimedia favorece sensiblemente a las lenguas que han invertido en este espacio (caso del hebreo o el sueco, por ejemplo). Es importante considerar la lista globalmente, sin confiar demasiado en el orden, y descubrir que claramente apunta a las lenguas nacionales de países y regiones reconocidos por **su liderazgo en el campo de la sociedad de la información.**

Las lenguas más conectadas son las siguientes:

	% LOCUTORES CONECTADOS
Noruego	97,87%
Danés	97,82%
Sueco	93,49%
Japonés	92,62%
Holandés	92,03%
Limburgués	91,90%
Suizo Alemán	91,56%
Catalán	90,47%
Flamenco occidental	90,43%
Finlandés	89,67%
Estonio	89,10%
Letón	88,95%
Gallego	88,61%
Sajón superior	88,13%
Alemán	87,68%
Bávaro	87,68%

Y finalmente las lenguas con mayor gradiente:

	GRADIENTE
Hebreo	2,82
Noruego	2,60
Estonio	2,41
Serbocroata	2,24
Finlandés	2,13
Malgache	2,13
Inglés	1,73
Sueco	1,54
Italiano	1,53
Macedonio	1,37
Holandés	1,36
Alemán	1,31
Esloveno	1,30
Catalán	1,30
Polaco	1,27
Español	1,25

La presencia en este cuadro del **malgache**, una lengua con un número de personas conectadas sumamente bajo (menos de 10%), plantea interrogantes sobre el método. Ocurre que el malgache tiene una presencia, en algunos de los cuadros de Wikimedia[13], altamente desproporcionada con relación a su presencia en la Internet y, por el juego de los promedios, la desproporción es tan enorme que logra impactar los resultados. Es uno de los síntomas

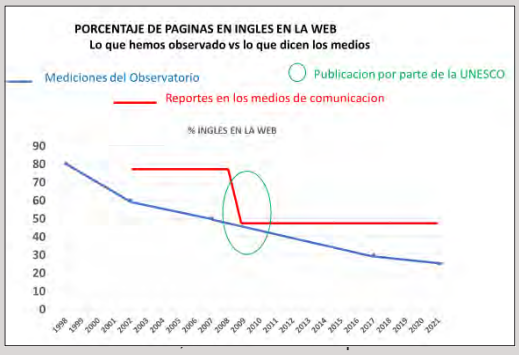
13.- Especialmente el "wiktionnary" (https://es.wiktionary.org/wiki/Wikcionario:Portada) donde llega tener 18% del total de entradas.

de los sesgos del indicador de contenidos sobre el cual seguimos trabajando.

Las nuevas lenguas que aparecen altas en *capacidad* y *gradiente* en la edición con 330 lenguas, confirman el diagnóstico del 2017: **noruego**, **esloveno**, **estonio** y **catalán** están asociadas a países o regiones cotizados en los parámetros de la sociedad de la información.

Nos centramos ahora en las primeras lenguas de la Internet, comenzando con el **inglés**. La fuente de datos sobre la presencia de lenguas en la Web más utilizada, y durante mucho tiempo la única, es W3Techs[14].

¿Como es posible que esta fuente indique un porcentaje de contenidos en inglés del 62.5% en 2021 cuando el Observatorio propone la cifra del 30%? Midiendo directamente las páginas web con un algoritmo de reconocimiento de lenguas no deberían equivocarse. ¡*Pues sí, es el multilinguismo, estúpido*! Si se nos permite reutilizar esa famosa y provocativa expresión… y tratar un problema histórico de desinformación acerca del espacio del inglés en la Internet, ilustrado con estas dos curvas:



formes sobre la presencia del inglés en la Web posicionándola, sin cambios durante una década, en el 80%, mientras nuestras mediciones indicaban un declive progresivo hacia el 50%. Los medios se apoyaban en 3 publicaciones que proponían, con la misma metodología, los mismos resultados, en 1997, 1999 y 2002. La metodología no era realmente sesgada sino científicamente no válida[15], ver Pimienta (2009) para más detalles. Después de las publicaciones de la UNESCO sobre el tema (Pimienta 2006, 2009), los medios de comunicación[16] adoptaron progresivamente ese nuevo valor de 50%. Luego apareció W3Techs como única fuente, cuyos resultados mantuvieron un valor entre 50 y 60% desde el 2011[17].

¿Cómo procede W3Techs y cuál es el problema que ocurre en relación con el multilingüismo?

W3Techs selecciona los 10 millones de sitios más visitados de la Web, según indica la aplicación de márketing digital Alexa (http://alexa.com). Dejemos de lado los sesgos a favor del inglés de los algoritmos de reconocimiento de lengua y el hecho de que seleccionar los 10 millones de sitios más visitados (sobre los cerca de 1.2 billones de sitios Web existentes[18], es decir menos de 1% de ellos) favorece a los sitios en inglés. Hoy en día, analizar la Web entera parece una tarea inalcanzable y nuestra intención no es despreciar el trabajo encomiable de W3Techs, que provee de muchos datos muy útiles. Nos centramos sobre todo en su gestión del multilingüismo. W3Techs

14.- https://w3techs.com/technologies/overview/content_language
15.- Se aplicaba un algoritmo de reconocimiento de lenguas sobre la página de entrada de 3000 sitios escogidos al azar una sola vez a partir de números de IP, y se calculaban los porcentajes. El método estadístico para validar esta aproximación sería repetir un gran número de veces el método y luego analizar la variable aleatoria con herramientas estadísticas (promedio, varianza, etc.). ¡Un solo tiro al arco sobre una diana no suele informar sobre la capacidad del tirador!
16.- Y también desafortunadamente Wikipedia (https://en.wikipedia.org/wiki/Languages_used_on_the_Internet) de quien uno espera más cautela.
17.- Ver https://w3techs.com/technologies/history_overview/content_language/ms/y).
18.- Fuente: https://news.netcraft.com/archives/category/web-server-survey/

aplica su algoritmo, a diario, sobre **la página de entrada** de esos 10 millones de sitios. La decisión de limitarse a la página de entrada, sin compensarla de alguna manera, es parte del problema. Desde luego la contabilidad de las lenguas en la Web debería hacerse a nivel de páginas y no a nivel de sitios, pues no se puede contabilizar igual un sitio web limitado a una página y otro con miles de páginas. A eso se suma la posibilidad de que ese sitio web con miles de páginas tenga páginas en distintas lenguas, a pesar de que la página de entrada esté principalmente en inglés, incrementando así el tamaño del error. Hoy en día, una buena parte de los sitios más visitados (como Facebook por ejemplo) ofrecen, desde la página de entrada, decenas de versiones lingüísticas; contabilizar la página de entrada en inglés es dejar de lado todas esas versiones. Finalmente, es muy común que la página de entrada de un sitio en una lengua diferente del inglés tenga algunas palabras en inglés (por ejemplo, de navegación o de copyright); contabilizarla como página en inglés, como probablemente ocurre con el algoritmo de W3Techs, es obviar decenas de páginas en otros idiomas.

No hay que ser un experto en estadística para entender que el método, por no considerar la realidad del multilingüismo, se puede equivocar en proporciones gigantescas… ¿Qué podría hacer el algoritmo de W3Techs para mejorar sus productos, sin abandonar un enfoque pragmático, es decir sin entrar en el reto de analizar todas las páginas de todos los sitios?

- Analizar las opciones de lenguas ofertadas en la página de entrada y contabilizar cada opción al igual que la versión inglesa.
- Encontrar un método para obtener una estimación, aunque sea aproximativa, del número de páginas del sitio web y multiplicar cada versión lingüística por ese número para tener una contabilidad en páginas y no en sitios;
- Cuando el algoritmo reporta más de una lengua en la página de entrada, por principio, no contabilizarla como inglés.

Otros factores deberían alertar sobre lo inverosímil de los datos de W3Techs y llamar la atención sobre alguna anomalía estadística sintomática de un error grueso:

- no tiene sentido que la proporción de contenidos en inglés se haya quedada estable durante los últimos 10 años mientras en el mismo periodo los países asiáticos y árabes han invadido la Web y que un conjunto de lenguas no europeas[19] ocupa ahora cerca de su tercera parte;
- la presencia de internautas anglófonos (L1+L2) ha pasado de 32% en 2017 a 13% hoy en día;
- mostrar el chino con solo el 1.3% de los contenidos y el hindi con el 0.1% cuando esas dos lenguas representan respectivamente 17.5% y 4.2% de las personas conectadas.

Para cerrar este capítulo, que la proporción de páginas Web en inglés disminuya de ninguna manera significa que la presencia en términos absolutos del inglés disminuya, ni tampoco que haya terminado de crecer; solo significa que nuevas lenguas están ocupando más y más espacio, lo que reduce la proporción del inglés. Desde luego el inglés sigue siendo una lengua líder en la Internet, cuya proporción estimada de contenidos (30%) supera en un factor 2 la proporción de internautas (15%).

Hemos disertado en Pimienta (2017) de los sesgos en distintos proyectos y cómo la falta de consideración del multilingüismo puede llevar a errores flagrantes. El caso frecuente

19.- Chino, hindi, árabe, turco, bengalí, vietnamita, urdu, persa y maratí.

más típico es el cálculo de elementos basados sobre L1+L2, dividiendo por la población mundial, cosa que provoca errores de magnitud, escondidos en los valores del resto de lenguas. El número de hablantes L1+L2 es muy superior a la población mundial, habíamos estimado en 25% la proporción de personas multilingües en el 2017, en esa nueva versión, Ethnologue nos ofrece una cifra más acertada de 43%.

Observemos ahora las lenguas que siguen al inglés. El **chino** está en segunda posición en términos de *potencia* y *contenidos* pero ya ocupa la primera plaza en términos de *personas conectadas* en el mundo y, a diferencia de los países occidentales, donde muchos están por encima de los 90%, queda espacio para progresar. La tabla siguiente muestra datos cuyos sesgos son mínimos y se obtienen a partir de los datos de la UIT de personas conectadas a la Internet y de los datos demo-lingüísticos de Ethnologue, ponderando los primeros con los segundos[20]. La presencia de las lenguas asiáticas y del árabe es notable.

	% MUNDIAL INTERNAUTAS	% MUNDIAL LOCUTORES	% LOCUTORES CONECTADOS
Chino	17,5%	14,6%	65,59%
Inglés	15,2%	12,9%	64,35%
Español	7,0%	5,2%	73,04%
Hindi	4,2%	5,8%	40,18%
Árabe	3,9%	3,5%	60,25%
Ruso	3,5%	2,5%	77,21%
Portugués	3,0%	2,5%	66,96%
Francés	3,0%	2,6%	63,33%
Alemán	2,2%	1,4%	87,68%
Malayo	2,2%	2,3%	51,01%
Japonés	2,0%	1,2%	92,62%
Turco	1,3%	0,9%	77,95%
Bengalí	1,1%	2,6%	24,16%
Urdu	1,0%	2,2%	24,13%
Persa	0,9%	0,8%	63,99%
Vietnamita	0,9%	0,7%	69,00%
Coreano	0,9%	0,8%	64,73%
Italiano	0,9%	0,7%	75,66%

Proponemos ese análisis histórico de las lenguas en la Internet.

PERIODO	CARACTERISTICAS
1970-1990	La Internet nació en el mundo occidental, muy marcada por la lengua inglesa en su fase histórica inicial, tanto por razones tecnológicas (la lengua de los profesionales de las redes[21]) como por la naturaleza de sus primeros usuarios (el mundo de la investigación), donde una alta proporción utiliza el inglés como L2 aunque no sea su lengua materna. El inglés dominó la red durante ese periodo.
1990-2010	Este periodo corresponde al nacimiento de la Web (1992): las lenguas europeas invirtieron en la Internet, la cual se transformó en un espacio privilegiado para esas lenguas, con un dominio del inglés que disminuyó de 80% a 50% como resultado del empuje de las otras lenguas europeas.
2010-2020	La Internet fue a la vez motor y sujeto de la globalización y la proporción de internautas que tenían el inglés como L1 o L2 disminuyó rápidamente para acercarse al porcentaje mundial real, inferior al 20%. Así, su proporción en la Web se acerca lógicamente a su proporción en el mundo real, aunque guardando una ventaja histórica. El continente africano sigue sin embargo rezagado y la brecha digital es aún enorme[22].
2020-2030	Entramos en una nueva fase de globalización donde el peso demográfico comienza a ser el factor dominante, al menos en el seno del mundo árabe y asiático. En este periodo, el centro de gravedad lingüístico de la Internet va ir pasando hacia las lenguas asiáticas y el árabe y si el continente africano logra superar su brecha digital, su demografía puede reservar sorpresas…

Como conclusión, pensamos que la creencia según la cual la *lengua franca* de la Internet

20.- Y de nuevo centrado en L1+L2.

21.- Incluso en ese periodo inicial se adoptaron opciones informáticas, como el código ASCII de 7 bits que impedía el uso de los acentos y otros signos como la tilde, que tardaron varios años en ser superadas.

22.- De los 56 países con tasa de conexión a la Internet inferior a 30%, 34 son africanos, 7 asiáticos y 6 del Pacifico. De esos 34 países africanos, 14 tienen menos de 10% de personas conectadas.

es el inglés es una ilusión: lo que caracteriza la Internet más y más es el **multilingüismo**[23] y la economía digital claramente está cada día más caracterizada por el factor del multilingüismo.

La lucha contra la desinformación se ha transformado en un tema principal en este periodo de crisis sanitaria donde la desinformación puede conducir a la muerte. Acorde con nuestra visión (Pimienta, Rodríguez, 2020), la necesidad de desarrollar programas amplios y completos de **alfabetización informacional** es una emergencia tan aguda como la del calentamiento global. Claramente, esos programas deben incluir la educación de la ciudadanía para lidiar con los datos que se ofrecen en la Web, con una mente crítica, y una exigencia firme sobre la transparencia metodológica y algorítmica, incluyendo una presentación honesta de los sesgos inherentes a toda aproximación de datos construidos, sean con métodos estadísticos u otros. Es evidente que los progresos de la inteligencia artificial, basados en un uso intensivo de los datos, hacen esa necesidad aún más crítica.

REFERENCIAS

Pimienta D. (2021) *Versión nueva y mejorada de un enfoque alternativo para la producción de indicadores lingüísticos en la Internet.* Observatorio de la diversidad lingüística y cultural en la Internet http://funredes.org/lc2021/ALI%20V2-ES.pdf

Pimienta D., Rodríguez L.G. (2020) "¡Va de retro Internet! Una visión crítica de la evolución de la Internet desde la sociedad civil", *Revista Ibero-Americana de Ciência da Informação, V13 N3, Pp. 979-1000* - https://periodicos.unb.br/index.php/RICI/article/view/33041/27497

Pimienta D. (2019) *Un enfoque alternativo para producir indicadores de la presencia de las lenguas en la Internet.* Observatorio de la diversidad lingüística y cultural en la Internet http://funredes.org/lc2019/Alternativa%20Lengua%20Internet.docx

Pimienta D., Prado D., Blanco A. (2009) Twelve years of measuring linguistic diversity on the Internet: balance and perspectives, *UNESCO, Publications for World Summit on the Information Society*, *CI-2009/WS/1-* http://unesdoc.unesco.org/ulis/cgi-bin/ulis.pl?catno=187016

Pimienta D. (2005) Linguistic Diversity in cyberspace: models for development and measurement", in *Measuring Linguistic Diversity on the Internet, UNESCO, Publications for World Summit on the Information Society, 2005*- http://unesdoc.unesco.org/images/0014/001421/142186e.pdf

23.- Desde luego, es el espacio humano donde el multilingüismo se expresa de la mejor y mayor manera, dado sus características sin fronteras; el grado de multilingüismo de la Internet bien podría ser superior al de los humanos, sea en términos de contenidos, de tráfico, de usos o de interfaces…

XENOGRAPHY: THE IMPACT OF LINGUISTIC DIVERSITY ON THE EVOLUTION OF WRITING

Juan Carlos Moreno Cabrera
Retired professor of General linguistics

Introduction

Among the many positive properties of linguistic diversity, we find one that is directly related to one of the technologies that has notably determined a considerable part of current human societies: writing. Certainly, the invention of writing, as far as we currently know, has occurred on rare occasions in some societies of Antiquity in the Near East and Egypt, Central America, and China. Nevertheless, this technology quickly spread across the planet. This expansion caused writing designed and structured to transcribe a specific language to come into general contact with languages with vastly different and diverse genealogy and typology. We could think that, due to adapting writing designed for a specific language to another quite different language, this script could become deteriorated or degenerated; therefore, linguistic diversity would constitute a threat or an obstacle to its development. Nonetheless, in this article I will demonstrate that the opposite is in fact true. Adopted scripts are adapted to new languages they transcribe and, because of this, make writing itself evolve. Linguistic diversity has not only not hindered the evolution of writing, rather it has been, without a doubt, one of the factors that has strengthened and driven it.

Xenography

When a writing system is designed for a specific language, it is with that language in mind, and not another. That said, linguistic diversity is present from the beginning because it will need to transcribe not only the words of that language, but also those of other languages from communities with whom it maintains or has maintained friendly or hostile contact. This involves transcribing, at least, place names or personal names and certain typical words or expressions. I call this phenomenon *internal xenography* (Moreno Cabrera 2016: 117-128). When we write a French proper noun in English, we use the French, not the English, spelling; therefore, we write *Rousseau*, so, only by knowing how to pronounce that language, can we read this proper noun properly. In this case, internal xenography is not used. But other languages indeed do this; for example, in Russian, foreign proper nouns are transcribed making internal xenographic use of the Cyrillic alphabet: Руссо [Russó] transcribes Rousseau. Japanese also does this:ルソー [Rusō]. These are cases of internal xenography: script designed for one language is used to transcribe words from a different language. If English resorted to this method, in this case we would write *Roosoh* to name this French philosopher. The Royal Spanish Academy (RAE) used this same procedure in its proposal to write the English word of Irish origin *whisky* as *güisqui* in Spanish. Russian uses internal xenography to transcribe it as виски [viski] and Japanese as ウイスキー [wisukī

External xenography (Moreno Cabrera 2016: 117-128), also called *alloglottography* (Kornicki 2018: 36), occurs when an illiterate community decides to take on the script of another community, whose writing may have been adopted and adapted from another community or invented by the same.

In many of these cases, writing specifically designed to transcribe one language is used to visualise another language with a different phonological, morphological, and lexical system. This brings up linguistic issues that are not always easy, whose resolution requires applying practically the same degree of inventiveness and ingenuity as to invent a new writing system. So, a community that adapts an existing script due to not creating its own writing for its own language is under no circumstances considered less intelligent or ingenious than the community that invented the borrowed script, as we will see in the specific cases set forth in the following sections.

In communities where xenography is used, written words that are meaningless to this community come into play. The Spanish surname *Zapatero* has a particular meaning for Spanish speakers, which is lost on English speakers. For them, the sound associated to that written name is what matters. The same can be said of the English surname *Smith* for Spanish speakers. If we transcribe this English surname using internal xenography into Spanish, *Esmiz*, for example, it becomes obvious that internal xenography brings the surname's phonetics to the forefront. Therefore, we have the following phoneticisation law of writing systems:

> Xenographic processes cause the progressive phoneticisation of writing systems with lexical foundations.

This trend determines the evolution of lexically based writing systems towards phonetically based scripts.

Linguistic units in the origin of writing

All known writing systems, at their source, are arranged around the two basic language units corresponding to the first and second articulations of human language: the natural word and the syllable, respectively. So, we could say that the original writing systems are logosyllabic: this is the case of Egyptian, Sumerian, and Chinese scripts.

All languages have two kinds of natural words: lexical and grammatical words. In a simple expression such as *Mary's car*, the words *Mary* and *car* are lexical and the possessive form *'s* is grammatical. Egyptian, Sumerian, and Chinese scripts, three of the most anciently attested, are clearly logosyllabic: graphemes are interpreted as syllables and/or lexical or grammatical words. Some sinology specialists describe Chinese writing as *morphosyllabic* (Rovira 2010: 63-68). It is true that in Chinese there are lexical words such as *rén* 'human being, person' or *shān* 'mountain, hill' and grammatical words such as the perfect aspect marker *le* or the connecting morpheme *de*. Therefore, linguistics includes lexical words and grammatical words, those with no lexical context whose function is exclusively grammatical; morphemes are always grammatical, not lexical, and only appear affixed to a lexical word. This is precisely what happens with the aforementioned Chinese morphemes *le* and *de*; but it would not be linguistically accurate to describe *rén* and *shān* as morphemes: they are clearly lexemes or lexical elements. So, it would be more accurate to describe Chinese writing as *logomorphosyllabic*. Even so, since linguistics distinguishes between lexical and grammatical words, we can resort to the term *logophonographic* (Moreno Cabrera 2005: 82-86) or the more traditional *logosyllabic* (Moreno Cabrera 2021: 227-232) to describe Chinese writing, given that it is much less shocking in terms of terminology to say that in Chinese *le* and *de* are grammatical words than to declare that *rén* and *shān* are morphemes, at least from the perspective of linguistic science.

Chinese writing and sinography

Chinese writing is, without a doubt, one of the most influential in the world. Z. Handel (2019: 10) suggests the term *sinographosphere* to denote areas in which the influence of Chinese writing has been decisive. Adapting Chinese script to write Korean, Japanese, or Vietnamese brought into play many phonetic, lexical, and grammatical mechanisms that modified that script in several ways, and Japanese is the most interesting and radical case (Taylor & Taylor, 2014: 255-361, Handel 2019: 166-211, Li 2020: 87-100).

In the previous section, I said that Chinese writing is logosyllabic. Indeed, each Chinese character is associated to a syllable and, with very few exceptions, to a meaning, whether lexical or grammatical. Here are the characters of the aforementioned Chinese words:

Lexical words
人 *rén* 'human being, person, people' / 山 *shān* 'hill, mountain'
Grammatical words
了 *le* / 的 *de*

An issue that immediately arises is how to transcribe foreign terms in Chinese. This is what I call *internal xenography*. Well, only the syllabic value of the characters is used. To do so, the characters whose syllables are most similar to the syllables of the foreign word to be transcribed are chosen. This creates an expression that is poorly formed or absurd semantically, and its purely phonetic value comes to the forefront. Let us see a specific example to illustrate this point. To transcribe *Catalunya* in Chinese script, characters associated to Chinese syllables that are phonetically similar to the Catalan word's four syllables are chosen. Since the syllabic structure of Catalan is quite different to the syllabic structure of the Chinese language, Chinese resorts to five characters instead of the four that would be expected. The written Catalan syllable *nya* contains a voiced palatal nasal consonant that does not exist in Chinese. To imitate it, this language resorts to the sequence of two syllables, *ní* and *yǎ*, resulting in five characters corresponding to five syllables instead of the four syllables in the Catalan word. Here is the Chinese transcription of *Catalunya* with the original meaning of the characters used:

加泰啰尼哑
Jiā tài luō ní yǎ

The word *jiā* means 'add', *tài* means 'safe, peaceful', *luō* is a purely phonetic component without any meaning, *ní* means 'Buddhist nun' and *yǎ* means 'mute, dumb'.

As you can see, this Chinese expression does not mean anything, so it can only be interpreted phonetically.

Internal xenography, an obvious consequence of linguistic diversity, allows a purely syllabic interpretation of logosyllabographic script. That is, the occasional transformation of logosyllabographic script into syllabographic script. This potentiality of Chinese writing is systematically exploited in the case of adapting Chinese script to the Japanese language, a truly remarkable case of external xenography or alloglottography. We will see this below.

A remarkable example of external xenography: Japanese writing

Adopting and adapting Chinese script to write Japanese is, without a doubt, a complex and challenging task, since Chinese and Japanese have vastly different phonological, lexical, morphological, and syntactic structures. However, that process has been the source of one of the most interesting and fruitful writing systems that have been conceived in the history of humanity.

By using Chinese characters to write Japanese words, a Chinese lexeme that consists of a single syllable often corresponds to a Japanese lexeme with more than one syllable. Let us see a specific example. The Chinese lexeme that means 'mountain' that we saw in the previous section:

山 *shān* 'hill, mountain'

By using the Chinese logogram to write the corresponding Japanese word, which is *yama*, this results in:

山 *yama* 'hill, mountain'

This disobeys the principle that governs Chinese writing, according to which each character is associated to one syllable. In this case, the *kanji* (Chinese logogram) for *mountain* is associated with the two-syllable word *yama*. Thus, the Chinese logosyllabographic system is converted into a new pure logographic system, where each character is associated with one word, regardless of the number of syllables it contains. Chinese writing has been stripped of its phonetic base, leaving only its logographic aspect. Hence, this results in a pure logographic script.

Additionally, in Japanese, *kanji* not only have a Japanese reading, called *kun*, but also keep their original Chinese reading, called *on* (Taylor & Taylor 2014: 276-278). This second reading often appears in compound words. In this way, the logogram for 'mountain' can be read as *yama* (*kun* reading) or *san* (*on* reading). For example, Mount Fuji is written in *kanji* as follows, where the last character is read according to its Chinese *on* reading:

富士山
Fuji san

But adapting Chinese script to write in Japanese does not end here, given that some Chinese characters were simplified to create two syllabaries, exploiting the phonetic aspect of Chinese writing.

These syllabaries based on simplified Chinese logosyllabograms are exclusively interpreted phonetically. The two syllabaries are called *hiragana*, which comes from simplified cursive versions of Chinese characters, and *katakana*, the result of simplified versions of Chinese characters. *Hiragana* is used to write grammatical morphemes, which are abundant in Japanese; *katakana* is used to write foreign words, among other things. That is, it is used to perform Japanese internal xenography. Let us see how we can write *Catalunya* in Japanese:

カタロニア
Ka ta ro ni a

This uses the *katakana* syllabary. As with the case of Chinese writing seen in the previous section, there are five instead of four syllabograms, given the differences between the syllabic structures of Catalan and Japanese.

So, Japanese writing combines logosyllabography (*on* reading of the *kanji*, using compound words of Chinese origin), pure logography (*kun* reading of the *kanji*, used for simple Japanese words) and two moraic syllabaries, given that they distinguish between short (one mora) and long (two morae) syllables: both long vowels and consonants are written with special graphemes. This explains the fact that there is a syllabic coda grapheme for the [-*n*] sound.

To illustrate this, let us see the following example:

日本語
に ほん ご
Ni hon go
'Japanese language'

The first line is the expression 'Japanese (*Nihon*) language (*go*)' in logosyllabography based on the Chinese or *on* reading of the *kanji*. We can see that each character corresponds to one syllable: *ni*, *hon* and *go*. But two of those characters, in their *kun* reading, are not monosyllabic.

The second character reads as *moto* 'source, root'; and the third, *kataru*, 'conversation, account, theme'.

The second line shows the same word in the *hiragana* syllabary. We can see that there are four rather than three graphemes as would be expected given that *nihongo* consists of three syllables. But the syllable *hon* consists of two morae (*ho-* and *-n*) and, therefore, is made up of two graphemes.

In his detailed study on the adaptation of the Chinese script to write Japanese, Korean, and Vietnamese, Handel (2019) reveals how typological diversity between the languages involved is one of the factors that determines the most radical transformations of that script. Japanese is a vastly different language to Chinese in the typological sense, both in terms of phonetic, morphological, and syntactic aspects. This is one of the factors that made Chinese script transform into three scripts: one pure logographic script (*kanji*) and two syllabographic scripts (*katakana* and *hiragana*). As we have seen, pure logography has not eliminated Chinese logosyllabography in Japanese, so, both coexist in Japanese writing, one of the most rich and fascinating scripts in the world.

Linguistic diversity and the origin of the alphabet

Logosyllabic cuneiform script was invented by the Sumerians towards 3200 BC and was used in Antiquity to transcribe many languages: Akkadian, Hittite, Eblaite, Elamite, Hurrian, Luwian and Urartian. One of the most interesting by-products of cuneiform script is the Ugaritic alphabet from 14th century BC, which uses grapheme tracing techniques from cuneiform writing.

It is this xenographic dynamic, where one script was adopted and adapted by different languages than that for which it was invented, which gave rise to the most widely spread writing system at this time.

Most alphabets in use today come from the Greek alphabet. How did this alphabet arise? It was adapted from the Phoenician abjad. Phoenician is a Semitic language, and Greek is an Indo-European language. The phonological systems of both languages are quite different and, the alphabet emerged precisely from the confrontation of this difference. That is, linguistic diversity is once more at the heart of the evolution of writing systems.

The Phoenician abjad was invented around the 12th century BC and consists of only twenty-two letters or phonemograms, which transcribe consonants.

Why are there no letters for vowels? Many fantastic and naive explanations have been given to answer this question. But the answer is extremely simple and is not due to the lack or imperfection of the Phoenician script, quite the opposite. Based on the acrophonic principle, each letter represents the initial consonant of the syllables. We know that all or practically all the phonological oppositions of a language occur in this syllabic position, given that many less phonological oppositions occur in the final or syllabic coda position. For example, in Spanish, the opposition between phonemes /n/, /m/ and /ñ/ is verified at the onset, not at the coda.

Oddly enough, in the case of Semitic languages, including Phoenician, no syllables start with a vowel, because they are always preceded by a glottal stop in this position. Subsequently, in line with the acrophonic principle, there are no letters for vowels in Semitic languages. The Ugaritic cuneiform alphabet, also a Semitic language, precisely includes graphemes for vowels, including the glottal stop. That is, it does not transcribe vowels as /a/, /i/, /u/, rather as /'a/, /'i/, /'u/, where the apostrophe indicates the glottal stop.

In the Phoenician abjad, the letter <A>, the first in the list, precisely transcribes the glottal stop and not vowel /a/.

According to Powell (1991), the Phoenician abjad was adapted to write Greek by someone from the island of Euboea around 800 BC, who consulted a Phoenician informer to write down Homer's poems. Regardless of the reliability or authenticity of this hypothesis, the truth is that the person who adapted the Phoenician abjad to Greek used the letters for Phoenician consonant sounds that did not exist in Greek to transcribe Greek vowels. The cases of the letter A and the letter O are particularly enlightening in this regard. The letter A in the Phoenician abjad transcribes the glottal stop. This consonant is not audible in itself but affects the vowel it precedes by giving it a sharp or abrupt onset, as is the case, for example, with the German *ein* 'one', which is preceded by an oral glottal stop. It is more than likely that the person or people who adapted Phoenician script to Greek did not identify the glottal stop associated to the letter A and used it to transcribe the Greek vowel /a/. The same can be said of the vowel O from the Phoenician abjad. It transcribed a voiced pharyngeal fricative consonant. This consonant, typical of Semitic languages, did not exist in ancient Greek and its rear position may have led those who adapted the alphabet to transcribe the vowel /o/.

The opinion that vowels were invented in Greece to build the first alphabet and, thus, that it was the Greeks who invented the alphabet, is not plausible. We must keep in mind that the Greeks never invented their own script, even though their civilisation constituted the basis of the philosophy, science, and literature of all modern western culture. The most ancient Greek testimonies date back to the 16th century BC and are written in a syllabary known as linear B, a script that was borrowed from the Minoan script known as linear A. This script fell into disuse after the 12th century BC and it was not until several centuries later that the Greeks adapted another script to transcribe Greek, as I have explained. Therefore, under no circumstances can the invention of a script be considered a mandatory sign of the cultural excellence of a civilisation.

A new perspective of the history of writing

The dominant idea today about how writing arose and evolved throughout the history of humanity was suggested by I. Gelb in the 1950s in his monograph on the history of writing (Gelb 1963). According to this author (Gelb 1963: 191), writing evolved from the initial logosyllabographic stage of the Egyptian, Sumerian, and Chinese scripts, through the syllabographic stage of the Phoenician and Japanese scripts to the alphabetic stage of Greek script. One of the qualms of this proposal is that the Phoenician script is not syllabographic, rather phonemographic, and the Egyptians already included a series or phonemographic elements in their logographic script. That is, phonemography, which is the basis of the alphabet, already existed in the logosyllabographic scripts

of Antiquity. One question that sometimes comes up is that if ancient writing systems such as Egyptian were able to transcribe isolated sounds, phonemes, or syllables, why did it not become widespread at the expense of the logosyllabic system, which was much more complex and difficult to learn? The answer is related to the religious and sociocultural function of writing in ancient societies, where writing was restricted to a more or less privileged caste and was in no way designed to be used generally by ordinary folk.
Within the classic Chinese education system, the exam system for civil servants demanded a level of dedication that was within reach of very few people:

> "In their long existence, the Confucian classics —the Five Classics and the Four Books— have received much commentary by scholars, and have swollen to a massive body of literature, containing some 500,000 characters. Generations of scholars studied and memorized it in preparation for a series of civil service examinations. Its long and arduous study promoted scholarly literacy among a tiny upper crust of males while preventing the spread of functional literacy among the masses." (Taylor & Taylor 2014: 89)

The Chinese literary tradition, largely decisive in the entire cultural area under Chinese influence, featured precisely this elitist nature:

> "Sinitic remained, after all, a sacred language in the context of Buddhist scriptures throughout east Asia, and the Chinese classics were not only the language of the civil service examinations and the bureaucracy, but they were also constantly being re-examined, re-interpreted, and re-oriented to deal with present problems." (Kornicki 2018: 185)

Inventing a script is one of the most remarkable and important human intellectual feats. The societies of Antiquity that created writing systems were as intellectually advanced as modern societies. Writing's evolution from a logographic phase towards a phonemographic stage has nothing to do with an allegedly more advanced mental or intellectual development of modern societies compared to ancient civilisations. Although the ancient Greek society never invented its own script, the people who developed the basis of modern western culture and philosophy, Socrates, Plato, or Aristotle among them, were as smart as modern industrialised societies.

Xenography or alloglottography has been precisely one of the mechanisms that has contributed to writing's evolution, as I have tried to prove in this brief article.

Conclusions

Linguistic diversity is one of the most conspicuous manifestations of the impressive adaptability and creativity of this curious animal we call *human being*. Like human language, linguistic diversity is an unmistakable sign of human nature itself. Writing is further proof of our creativity and adaptability. As I have tried to demonstrate in this article, the convergence between writing and linguistic diversity has been precisely one of the factors that has made writing itself evolve towards becoming more phonetic and accessible to people in general. This convergence has in no way been a negative or counter-productive factor, rather the opposite, as we have seen.

From here we can draw an important lesson about the nature of linguistic diversity. Since the myth of Tower of Babel, linguistic diversity has been portrayed as a cause of lack of communication, isolation, lack of understanding and hostility. This attitude is typical of linguistic imperialism and hegem-

onic ideologies in general. Overlooking the importance of linguistic diversity to understand the cultural development of humanity can be a side effect of this attitude. In this article, I have aimed to show that we cannot brush aside linguistic diversity to understand the invention and development of a key technology in the history of humanity: writing.

References

Gelb, Ignace. 1963. *A Study of Writing.* Chicago: The University of Chicago Press.

Handel, Zev. 2019. *Sinography. The borrowing and adaptation of the Chinese script*. Leiden: Brill.

Kornicki, Peter Francis. 2018. *Languages, Scripts, and Chinese Texts in East Asia.* Oxford: Oxford University Press.

Li, Yu. 2020. *The Chinese Writing System in Asia. An interdisciplinary Perspective*. London: Routledge.

Moreno Cabrera, Juan Carlos. 2005. *Las lenguas y sus escrituras. Tipología, Evolución e Ideología*. Madrid: Síntesis.

Moreno Cabrera, Juan Carlos. 2016. *Multilingüismo y lenguas en contacto*. Madrid: Síntesis.

Moreno Cabrera, Juan Carlos. 2021. *La clasificación de las lenguas. Introducción a la taxonomía lingüística*. Madrid: Síntesis.

Rovira i Esteva, Sara. 2010. *Lengua y escritura chinas: mito y realidades*. Barcelona: Bellaterra.

Powell, Barry. 1991. *Homer and the origin of the Greek alphabet*. Cambridge: Cambridge University Press.

Taylor, Insup & Taylor, M. Martín. 2014. *Writing and Literacy in Chinese, Korean and Japanese.* Amsterdam: John Benjamins.

XENOGRAFIA: L'IMPACTE DE LA DIVERSITAT LINGÜÍSTICA EN L'EVOLUCIÓ DE L'ESCRIPTURA

Juan Carlos Moreno Cabrera
Catedràtic jubilat de Lingüística general

Introducció

Entre les moltes propietats positives de la diversitat lingüística trobem una que està directament relacionada amb una de les tecnologies que ha determinat notablement una part considerable de les societats humanes actuals. Es tracta de l'escriptura. Certament, la invenció de l'escriptura, pel que sabem actualment, ha tingut lloc en comptades ocasions a algunes societats de l'Antiguitat situades a l'Orient Pròxim i Egipte, a l'Amèrica central i a la Xina. No obstant això, aquesta tecnologia es va expandir ràpidament per tot el planeta. Aquesta expansió va ocasionar el contacte generalitzat d'una escriptura pensada i estructurada per transcriure una llengua concreta amb llengües d'una genealogia i tipologia molt diferent i diversa. Es podria dir que, a l'adaptar una escriptura pensada per una llengua concreta a una altra llengua molt diferent, es podria produir un deteriorament o una degeneració d'aquella escriptura i, per tant, la diversitat lingüística constituiria una amenaça o un obstacle pel seu desenvolupament. Tanmateix, en aquest article demostraré que exactament el contrari és cert. Les escriptures adoptades s'adapten a les noves llengües que transcriuen i, gràcies a això, fan evolucionar la pròpia escriptura. La diversitat lingüística no només no ha impedit l'evolució de l'escriptura, sinó que ha estat, sens dubte, un dels factors que l'ha potenciat i impulsat.

La xenografia

Quan es dissenya un sistema d'escriptura per una llengua concreta, es pensa lògicament en aquella llengua i no en una altra. Malgrat això, la diversitat lingüística és ja present des de l'inici perquè caldrà transcriure no només les paraules d'aquella llengua, sinó també les de les altres llengües de comunitats amb les quals manté o ha mantingut algun contacte amistós o hostil. Això implica transcriure, almenys, topònims o antropònims i determinades paraules o expressions característiques. Anomeno aquest fenomen *xenografia interna* (Moreno Cabrera 2016: 117-128). Quan escrivim un nom propi francès en català, utilitzem l'ortografia francesa, no la catalana; d'aquesta manera escrivim *Rousseau*, per tant, només sabent com es pronuncia en aquella llengua, podem llegir l'antropònim adequadament. En aquest cas, no s'ha utilitzat la xenografia interna. Però altres llengües sí que ho fan; per exemple, en rus els noms propis estrangers es transcriuen fent ús xenogràfic intern de l'alfabet ciríl·lic: Русco [Russó] transcriu Rousseau. També ho fa el japonès: ルソー [Rusō]. Aquests són casos de xenografia interna: l'escriptura pensada per una llengua s'utilitza per transcriure paraules d'una llengua diferent. Si el català fes servir aquest mètode, en aquest cas escriuria *Russó* per anomenar aquest filòsof francès. La RAE[1] va fer servir aquest mateix procediment en la seva proposta per escriure en castellà la paraula anglesa d'origen irlandès *whisky* com a *güisqui*. Al rus es transcriu amb xenografia interna com a виски [viski] i en japonès com a ウイスキー [wisukī].

1.- Real Academia Española

La *xenografia externa* (Moreno Cabrera 2016: 117-128), també anomenada *al·loglotografia* (Kornicki 2018: 36), es produeix quan una comunitat illetrada decideix prendre l'escriptura d'una altra comunitat, l'escriptura de la qual pot haver estat adoptada i adaptada d'una altra comunitat o inventada per aquesta mateixa. En molts d'aquests casos es tracta d'utilitzar una escriptura dissenyada específicament per transcriure una llengua amb l'objectiu de visualitzar una altra llengua amb un sistema fonològic, morfològic i lèxic diferent. Això planteja problemes lingüístics no sempre fàcils, per solucionar els quals cal aplicar gairebé el mateix grau d'enginy i inventiva que per inventar un nou sistema d'escriptura. Per aquest motiu, una comunitat que adapta una escriptura ja existent per no haver-ne creat una de pròpia per la seva llengua, en cap cas pot considerar-se menys intel·ligent o enginyosa que la comunitat que ha inventat l'escriptura prestada, tal com veurem als casos concrets que exposo a les següents seccions.

A la xenografia entren en joc paraules escrites sense sentit per la comunitat a la qual es produeix aquest fenomen. El cognom castellà *Zapatero* té un sentit per a persones castellanoparlants, però no significa res per les angloparlants. Per aquestes, el so associat a aquest nom escrit és el que compta. Podem dir el mateix del cognom anglès *Smith* per les persones castellanoparlants. Si transcrivim amb xenografia interna aquest cognom anglès en *Esmiz*, posem per cas, és evident que la fonètica del cognom és el que es mostra en primer pla a la xenografia interna. Tenim, doncs, la següent llei de fonetització dels sistemes d'escriptura:

> Els processos xenogràfics provoquen la progressiva fonetització dels sistemes d'escriptura que tenen un fonament lèxic.

Aquesta tendència és la que determina l'evolució dels sistemes d'escriptura de base lèxica cap a escriptures de base fonètica.

Les unitats lingüístiques a l'origen de l'escriptura

Tots els sistemes d'escriptura coneguts, en el seu origen, s'organitzen al voltant de les dues unitats bàsiques de les llengües corresponents a la primera i segona articulació del llenguatge humà: la paraula natural i la síl·laba, respectivament. Per això, podem dir que els sistemes d'escriptura originals són logosil·làbics: és el cas de les escriptures egípcia, sumèria i xinesa.

A totes les llengües trobem dos tipus de paraules naturals: les lèxiques i les gramaticals. A una expressió senzilla com *el cotxe de la Maria*, les paraules *cotxe* i *Maria* són lèxiques i les paraules *el*, *de* i *la* són gramaticals. Les escriptures egípcia, sumèria i xinesa, tres de les més antigament evidenciades, són clarament logosil·làbiques: els grafemes s'interpreten com a síl·labes i/o com a paraules lèxiques o gramaticals. Algunes persones especialistes en sinologia qualificarien l'escriptura xinesa de *morfosil·làbica* (Rovira 2010: 63-68). És cert que en xinès hi ha paraules lèxiques com *rén* 'ésser humà, persona' o *shān* 'muntanya, turó' i paraules gramaticals com el marcador d'aspecte perfecte tipus *le* o el morfema relacionant *de*. Per això, la lingüística parla de paraules lèxiques i paraules gramaticals, aquelles sense context lèxic la funció de la qual és exclusivament gramatical; els morfemes són sempre gramaticals, no lèxics i només apareixen afixats a una paraula lèxica. Això és precisament el que ocorre amb els morfemes xinesos *le* i *de* que acabo d'esmentar; però no seria lingüísticament adequat qualificar *rén* i *shān* de morfemes: són clarament lexemes o elements lèxics. Per aquest motiu, seria més adequat qualifi-

car l'escriptura xinesa de *logomorfosil·làbica*. Amb tot això, com la lingüística sol distingir entre paraules lèxiques i gramaticals, podem recórrer perfectament al terme *logofonogràfic* (Moreno Cabrera 2005: 82-86) o al més tradicional *logosil·làbic* (Moreno Cabrera 2021: 227-232) per caracteritzar l'escriptura xinesa, ja que resulta molt menys terminològicament xocant dir que en xinès *le* i *de* són paraules gramaticals que afirmar que *rén* i *shān* són morfemes, almenys des de la ciència lingüística.

L'escriptura xinesa i la sinografia

L'escriptura xinesa és, sens dubte, una de les més influents del món. Z. Handel (2019: 10) proposa el terme *sinografoesfera* per indicar aquelles àrees a les quals la influència de l'escriptura xinesa ha estat determinant. L'adaptació de l'escriptura xinesa per escriure coreà, japonès o vietnamita va posar en joc nombrosos mecanismes fonètics, lèxics i gramaticals que van modificar aquesta escriptura de diverses maneres, essent el cas del japonès el més interessant i radical (Taylor & Taylor, 2014: 255-361, Handel 2019: 166-211, Li 2020: 87-100).

A la secció anterior he dit que l'escriptura xinesa és logosil·làbica. Efectivament, cada caràcter xinès s'associa a una síl·laba i, amb molt poques excepcions, a un significat, o bé lèxic o bé gramatical. Vet aquí els caràcters de les paraules xineses esmentades a la secció anterior:

Paraules lèxiques
人 *rén* 'ésser humà, persona, gent' / 山 *shān* 'turó, muntanya'
Paraules gramaticals
了 *le* / 的 *de*

Una qüestió que sorgeix immediatament és com es poden transcriure en l'escriptura xinesa els termes estrangers. Es tracta del que anomeno *xenografia interna*. Doncs bé, el que es fa és utilitzar els caràcters tan sols amb el seu valor sil·làbic. Per fer-ho, es trien caràcters les síl·labes de les quals s'assemblen més a les síl·labes de la paraula estrangera que es vol transcriure. D'aquesta manera, s'obté una expressió semànticament mal formada o absurda, i el valor purament fonètic passa a primer pla. Anem a veure un exemple concret per il·lustrar aquest punt. Per transcriure *Catalunya* amb l'escriptura xinesa, se seleccionen caràcters associats a síl·labes xineses que s'assemblen fonèticament a les quatre síl·labes de la paraula catalana. Com que l'estructura sil·làbica del català és molt diferent de l'estructura sil·làbica de la llengua xinesa, el xinès empra cinc caràcters en comptes dels quatre previsibles. La síl·laba catalana escrita *nya* conté una consonant nasal palatal inexistent en xinès; per imitar-la es recorre a la seqüència de les dues síl·labes *ní* i *yǎ*, per tant, obtenim cinc caràcters corresponents a cinc síl·labes enlloc de les quatre síl·labes de la paraula catalana. Vet aquí la transcripció xinesa de *Catalunya* amb la indicació dels significats originals dels caràcters empleats:

加泰啰尼哑
Jiā tài luō ní yǎ

La paraula *jiā* significa 'sumar, afegir, *tài* significa 'segur, pacífic', *luō* és un component purament fonètic sense significat, *ní* significa 'monja budista' i *yǎ* significa 'mut, ronc'.

Com es pot comprovar, aquesta expressió xinesa no té cap sentit coherent, així que només es pot interpretar fonèticament.

La xenografia interna, una clara conseqüència de la diversitat lingüística, fa que es produeixi una interpretació purament sil·làbica d'una escriptura logosil·labogràfica. És a dir, la transformació ocasional d'una escriptura

logosil·labogràfica en sil·labogràfica. Aquesta potencialitat de l'escriptura xinesa s'explota de forma sistemàtica en el cas de l'adaptació de l'escriptura xinesa a la llengua japonesa, un cas realment notable de xenografia externa. Veiem-ho a continuació.

Un exemple notable de xenografia externa: l'escriptura japonesa

L'adopció i adaptació de l'escriptura xinesa per escriure japonès és sens dubte una tasca complexa i difícil, donat que la llengua xinesa i la japonesa tenen estructures fonològiques, lèxiques, morfològiques i sintàctiques molt diferents. Tanmateix, aquest procés ha estat l'origen d'un dels sistemes d'escriptura més interessants i fructífers que s'hagin ideat a la història de la humanitat.

A l'utilitzar els caràcters xinesos per escriure paraules japoneses, amb freqüència es produeix la circumstància on un lexema xinès que consta d'una sola síl·laba correspon a un lexema japonès de més d'una síl·laba. Veiem un exemple concret. El lexema xinès que significa 'muntanya' que hem vist a la secció anterior:

山 *shān* 'turó, muntanya'

A l'emprar el logograma xinès per escriure la paraula japonesa corresponent, que és *yama*, obtenim:

山 *yama* 'turó, muntanya'

Això contravé el principi que regeix l'escriptura xinesa, segons el qual cada caràcter s'associa a una síl·laba. En aquest cas, al *kanji* (logograma xinès) per a *muntanya* se li associa la paraula bisíl·laba *yama*. D'aquesta manera, el sistema logosil·labogràfic xinès es converteix en un nou sistema logogràfic pur, on cada caràcter s'associa a una paraula, independentment del nombre de síl·labes que tingui. S'ha despossseït l'escriptura xinesa de la seva base fonètica, deixant únicament el seu aspecte logogràfic. Així, s'obté una escriptura logogràfica pura.

A més a més, en japonès els *kanji* no només tenen la lectura japonesa, anomenada *kun*, sinó que també retenen la seva lectura original xinesa, anomenada *on* (Taylor & Taylor, 2014: 276-278). Aquesta última apareix habitualment a paraules compostes. D'aquesta manera, el logograma per 'muntanya' es pot llegir com *yama* (lectura *kun*) o com *san* (lectura *on*). Per exemple, el mont Fuji s'escriu en *kanji* de la següent manera, on l'últim caràcter es llegeix amb la seva lectura xinesa *on*:

富士山
Fuji san

Però l'adaptació de l'escriptura xinesa per escriure japonès no s'acaba aquí, ja que es van simplificar alguns caràcters xinesos per crear dos sil·labaris i, per tant, s'explota la vessant fonètica de l'escriptura xinesa.

Aquest sil·labaris procedents de logosil·labogrames xinesos simplificats tenen una interpretació exclusivament fonètica. Els dos sil·labaris s'anomenen *hiragana*, que procedeix de versions cursives simplificades de caràcters xinesos i *katakana*, resultat de versions simplificades dels caràcters xinesos. El *hiragana* s'utilitza per escriure els morfemes gramaticals, molt abundants a la llengua japonesa; el *katakana* es fa servir per escriure paraules estrangeres, entre d'altres coses. És a dir, s'utilitza per dur a terme la xenografia interna japonesa. Veiem com es pot escriure *Catalunya* en japonès:

カタロニア
Ka ta ro ni a

S'utilitza el sil·labari *katakana*. Com en el cas de l'escriptura xinesa vista a la secció ante-

rior, hi ha cinc i no quatre síl·labes, donades les diferències entre les estructures sil·làbiques del català i el japonès.

L'escriptura japonesa és, doncs, una combinació d'una logosil·labografia (lectura *on* dels *kanji*, emprada en paraules compostes d'origen xinès), una logografia pura (lectura *kun* dels *kanji*, emprada en paraules simples pròpiament japoneses) i dos sil·labaris de naturalesa moraica, ja que es distingeix entre les síl·labes breus (una mora) i les llargues (dues mores): tant les vocals com les consonants llargues s'escriuen mitjançant grafemes especials. Això explica que hi hagi un grafema pel so [-*n*] de coda sil·làbica.

Per il·lustrar-ho, veiem el següent exemple:

日本語
に ほん ご
Ni hon go
'Llengua japonesa'

A la primera línia tenim l'expressió 'llengua (*go*) japonesa (*Nihon*)' amb la logosil·labografia basada en la lectura *on* o xinesa dels *kanji*. S'observa que cada a caràcter li correspon una síl·laba: *ni*, *hon* i *go*. Però dos d'aquests mateixos caràcters, amb lectura *kun*, no són monosil·làbics.

El segon caràcter es llegeix *moto* 'origen, arrel'; i el tercer, *kataru* 'conversa, narració, tema'.

A la segona línia podem veure la mateixa paraula emprant el sil·labari *hiragana*. S'observa que hi ha quatre grafemes i no tres com seria d'esperar donat que *nihongo* consta de tres síl·labes. Però la síl·laba *hon* consta de dues mores (*ho-* i *-n*) i, per tant, es compon de dos grafemes.

Al seu detallat estudi sobre l'adaptació de l'escriptura xinesa per escriure japonès, coreà i vietnamita, Handel (2019) mostra com la diversitat tipològica entre les llengües involucrades és un dels factors que determinen les transformacions més radicals d'aquesta escriptura. El japonès és una llengua molt diferent del xinès a nivell tipològic, tant en aspectes fonètics com morfològics i sintàctics. Aquest és un dels factors que va provocar la transformació de l'escriptura xinesa en tres escriptures: una logogràfica pura (*kanji*) i dues sil·labogràfiques (*katakana* i *hiragana*). Com hem vist, la logografia pura no ha eliminat la logosil·labografia xinesa en japonès, per la qual cosa ambdues conviuen a l'escriptura japonesa, una de les escriptures més riques i fascinants del món.

La diversitat lingüística i l'origen de l'alfabet

L'escriptura cuneïforme logosil·làbica va ser inventada pels sumeris cap a l'any 3200 abans de Crist i es va utilitzar a l'Antiguitat per transcriure nombroses llengües: l'accadi, la hitita, l'eblaïta, l'elamita, la hurrita, el luvi o l'urartià. Una de les derivacions més interessants de l'escriptura cuneïforme és l'alfabet ugarític del segle XIV abans de Crist, on s'utilitzaven les tècniques del traçat de grafemes de l'escriptura cuneïforme.

És aquesta dinàmica xenogràfica, on una escriptura va ser adoptada i adaptada per llengües diferents a la llengua per la qual es va inventar, el que va donar lloc al sistema d'escriptura més difós actualment.

La majoria dels alfabets en ús avui en dia procedeixen de l'alfabet grec. Com va sorgir aquest alfabet? Va ser una adaptació del consonantari fenici. El fenici és una llengua semita i el grec, una llengua indoeuropea. Els sistemes fonològics d'ambdues llengües són molt diferents i l'alfabet va sorgir precisament de la confrontació d'aquesta diferència. És a dir, la

diversitat lingüística és un cop més l'origen de l'evolució dels sistemes d'escriptura.

El consonantari fenici es va inventar cap al segle XII abans de Crist i consta de només 22 lletres o fonemogrames, que transcriuen consonants.

Per què no hi ha lletres per les vocals? S'han donat explicacions molt fantasioses i innocents per respondre aquesta pregunta. Però la resposta és molt senzilla i no es deu a un dèficit o imperfecció de l'escriptura fenícia, sinó ben al contrari. Basant-se en el principi acrofònic, cada lletra representa la consonant inicial de les síl·labes. Sabem que en aquesta posició sil·làbica es dona la totalitat o quasi totalitat de les oposicions fonològiques d'una llengua, ja que a la posició final o coda sil·làbica es donen moltes menys oposicions fonològiques. Per exemple, en castellà l'oposició entre els fonemes /n/, /m/ i /ñ/ es verifica a l'atac sil·làbic, no a la coda sil·làbica.

Curiosament, a les llengües semítiques, entre les quals s'hi troba el fenici, cap síl·laba comença per vocal, perquè sempre estan precedides en aquesta posició per una oclusió glotal. Per tant, en coherència amb el principi acrofònic, no hi ha lletres per a les vocals a les llengües semítiques. Precisament, a l'alfabet cuneïforme ugarític, llengua també semítica, hi ha grafemes per a les vocals, que inclouen l'oclusió glotal. És a dir, no es transcriuen les vocals /a/, /i/, /u/, sinó /ʾa/, /ʾi/, /ʾu/, on la cometa volada indica l'oclusió glotal.

Al consonantari fenici, la lletra <A>, la primera de la llista, transcriu precisament l'oclusió glotal i no la vocal /a/.

Segons Powell (1991), l'adaptació del consonantari fenici per escriure el grec la va dur a terme una persona de l'illa d'Eubea cap al 800 abans de Crist, qui va consultar un informant fenici per posar per escrit els poemes homèrics. Independentment de la veracitat o la versemblança d'aquesta hipòtesi, el cert és que qui va adaptar el consonantari fenici al grec va utilitzar les lletres dels sons consonàntics fenicis inexistents en grec per transcriure les vocals gregues. Els casos de la lletra A i la lletra O són molt il·lustratius al respecte. La lletra A al consonantari fenici transcriu l'oclusió glotal. Aquesta consonant no és audible en sí mateixa, però afecta la vocal a la qual precedeix donant-li un inici tallant o abrupte, com passa, per exemple, al *ein* 'un' de l'alemany, que va precedit per una oclusió glotal no escrita. És més que probable que la o les persones que van adaptar l'escriptura fenícia al grec no van identificar l'oclusió glotal associada a la lletra A i la van utilitzar per transcriure la vocal /a/ del grec. El mateix es pot dir de la vocal O del consonantari fenici. Transcrivia una consonant fricativa faríngia sonora. Aquesta consonant, típica de les llengües semítiques, no existia en grec antic i la seva localització posterior potser va dur les persones que en van fer l'adaptació a utilitzar-la per transcriure la vocal /o/.

No és plausible l'opinió de que les vocals es van inventar a Grècia per construir el primer alfabet i, per tant, que va ser el poble grec l'inventor de l'alfabet. Cal tenir en compte que els grecs mai van inventar una escriptura pròpia, tot i que la seva civilització va constituir la base de la filosofia, la ciència i la literatura de tota la cultura occidental actual. Els testimonis més antics del grec daten del segle XVI abans de Crist i estan escrits en un sil·labari conegut com a lineal B, una escriptura que va manllevar de l'escriptura minoica coneguda com a lineal A. Aquesta escriptura va deixar d'utilitzar-se a partir del segle XII a. de C. i van passar diversos segles fins que els grecs van tornar a adaptar una escriptura per transcriure el grec, com ja he explicat. La invenció d'una escriptura, per tant, de cap manera pot considerar-se un signe necessari de l'excel·lència cultural d'una civilització.

Una nova perspectiva de la història de l'escriptura

La idea dominant actualment de com va sorgir i evolucionar l'escriptura a la història de la humanitat la va proposar I. Gelb als anys cinquanta del segle passat a la seva monografia sobre la història de l'escriptura (Gelb 1963). Segons aquest autor (Gelb 1963: 191), l'escriptura va evolucionar des de la fase inicial logosil·labogràfica de les escriptures egípcia, sumèria i xinesa passant per la fase sil·labogràfica de l'escriptura fenícia i japonesa fins a arribar a la fase alfabètica de l'escriptura grega. Un dels inconvenients d'aquesta proposta és que l'escriptura fenícia no és sil·labogràfica, sinó fonemogràfica i, a més, els egipcis ja incloïen a la seva escriptura logogràfica una sèrie d'elements fonemogràfics. És a dir, la fonemografia, que és la base de l'alfabet, ja existia a les escriptures logosil·làbiques de l'Antiguitat. Una qüestió que sorgeix a vegades és que si als sistemes d'escriptura antics com l'egipci hi havia la possibilitat de transcriure sons aïllats, fonemes o síl·labes, per què no es va generalitzar, en detriment del sistema logosil·làbic, molt més complex i difícil d'aprendre? La resposta té a veure amb la funció religiosa i sòciocultural de l'escriptura a les societats antigues, on l'escriptura estava restringida a una casta més o menys privilegiada i no estava pensada de cap manera pel seu ús generalitzat per part del poble planer.

Dins del sistema educatiu xinès clàssic, el sistema d'exàmens pels funcionaris públics exigia una dedicació que estava a l'abast de molt poques persones:

> "In their long existence, the Confucian classics —the Five Classics and the Four Books— have received much commentary by scholars, and have swollen to a massive body of literature, containing some 500,000 characters. Generations of scholars studied and memorized it in preparation for a series of civil service examinations. Its long and arduous study promoted scholarly literacy among a tiny upper crust of males while preventing the spread of functional literacy among the masses." (Taylor & Taylor 2014: 89)[2]

La tradició literària xinesa, determinant a gran part de tota l'àrea cultural d'influència xinesa, va tenir precisament aquest caràcter elitista:

> "Sinitic remained, after all, a sacred language in the context of Buddhist scriptures throughout east Asia, and the Chinese classics were not only the language of the civil service examinations and the bureaucracy, but they were also constantly being re-examined, re-interpreted, and re-oriented to deal with present problems." (Kornicki 2018: 185)[3]

La invenció de l'escriptura és una de les proeses intel·lectuals humanes més notables i importants. Les societats de l'Antiguitat que van crear els sistemes d'escriptura eren tan avançades intel·lectualment com les societats actuals. L'evolució de l'escriptura des d'una etapa logogràfica cap a una fase fonemogràfica no té res a veure amb un suposat desenvolupament mental o intel·lectual més

2.- Nota del traductor (N. del T.): "A la seva llarga existència, els clàssics confucians —els Cinc Clàssics i els Quatre Llibres— han rebut molts comentaris de part d'erudits, i han crescut fins a una literatura enorme, constant d'uns 500.000 caràcters. Generacions d'erudits els han estudiat i après de memòria per preparar-se per una sèrie d'oposicions públiques. El seu llarg i feixuc estudi va promoure una alfabetització erudita entre una diminuta elit masculina mentre que impedia la propagació d'una alfabetització funcional entre les masses."

3.- N. del T.: "Al cap i a la fi, la llengua sinítica seguia essent una llengua sacra en el context de les escriptures budistes per tota l'Àsia Oriental, i els clàssics xinesos no només eren la llengua de les oposicions públiques i la burocràcia, sinó que també eren constantment re-examinats, re-interpretats, i re-orientats per resoldre problemes presents.

avançat de les societats modernes respecte les antigues. Tot i que la societat grega de l'Antiguitat mai va inventar una escriptura pròpia, les persones que van desenvolupar les bases de la filosofia i la cultura occidental actual, entre elles Sòcrates, Plató o Aristòtil, eren tan intel·ligents com les societats industrialitzades actuals.

La xenografia o al·loglotografia ha estat, precisament, un dels mecanismes que ha propiciat l'evolució de l'escriptura, tal com he intentat mostrar a aquest breu article.

Conclusió

La diversitat lingüística és una de les manifestacions més conspícues de l'impressionant adaptabilitat i creativitat d'aquest curiós animal que anomenem ésser humà. Com el llenguatge humà, la diversitat de les llengües és una marca inconfusible de la pròpia naturalesa humana. L'escriptura és una altra mostra de la nostra creativitat i adaptabilitat. La confluència entre l'escriptura i la diversitat lingüística ha estat precisament, tal com he volgut demostrar en aquest article, un dels factors que ha fet evolucionar la pròpia escriptura per fer-la cada cop més fonètica i més accessible per la gent en general. Aquesta confluència no ha estat de cap manera un factor negatiu o contraproduent sinó tot al contrari, tal com hem vist.

A partir d'aquí, podem treure una lliçó important sobre la naturalesa de la diversitat lingüística. Des del mite de la torre de Babel s'ha vist la diversitat lingüística com un factor d'incomunicació, d'aïllament, de incomprensió o d'hostilitat. Aquesta actitud és típica de l'imperialisme lingüístic i de les ideologies hegemòniques en general. Ignorar la importància de la diversitat lingüística per comprendre el desenvolupament cultural de la humanitat pot ser un efecte col·lateral d'aquesta actitud. A aquest article he volgut mostrar que no podem deixar de banda la diversitat lingüística per entendre la invenció i desenvolupament d'una tecnologia clau a la història de la humanitat: l'escriptura.

Referències

Gelb, Ignace. 1963. *A Study of Writing.* Chicago: The University of Chicago Press.

Handel, Zev. 2019. *Sinography. The borrowing and adaptation of the Chinese script*. Leiden: Brill.

Kornicki, Peter Francis. 2018. *Languages, Scripts, and Chinese Texts in East Asia.* Oxford: Oxford University Press.

Li, Yu. 2020. *The Chinese Writing System in Asia. An interdisciplinary Perspective*. London: Routledge.

Moreno Cabrera, Juan Carlos. 2005. *Las lenguas y sus escrituras. Tipología, Evolución e Ideología*. Madrid: Síntesis.

Moreno Cabrera, Juan Carlos. 2016. *Multilingüismo y lenguas en contacto*. Madrid: Síntesis.

Moreno Cabrera, Juan Carlos. 2021. *La clasificación de las lenguas. Introducción a la taxonomía lingüística*. Madrid: Síntesis.

Rovira i Esteva, Sara. 2010. *Lengua y escritura chinas: mito y realidades*. Barcelona: Bellaterra.

Powell, Barry. 1991. *Homer and the origin of the Greek alphabet*. Cambridge: Cambridge University Press.

Taylor, Insup & Taylor, M. Martín. 2014. *Writing and Literacy in Chinese, Korean and Japanese.* Amsterdam: John Benjamins.

XENOGRAFÍA: EL IMPACTO DE LA DIVERSIDAD LINGÜÍSTICA EN LA EVOLUCIÓN DE LA ESCRITURA

Juan Carlos Moreno Cabrera
Catedrático jubilado de Lingüística general

Introducción

Entre las muchas propiedades positivas de la diversidad lingüística está la que se relaciona directamente con una de las tecnologías que ha determinado de forma muy notable una parte apreciable de las sociedades humanas actuales. Se trata de la escritura. Ciertamente, la invención de la escritura, por lo que sabemos en la actualidad, ha tenido lugar en contadas ocasiones en algunas sociedades de la Antigüedad situadas en el próximo Oriente y Egipto, en América central y en China. Sin embargo, esta tecnología se expandió de forma rápida por todo el planeta. Esta expansión ocasionó el contacto generalizado de una escritura pensada y estructurada para transcribir una lengua concreta con lenguas de muy distinta y diversa genealogía y tipología. Se podría pensar que, al adaptar una escritura pensada para una lengua concreta a otra lengua muy diferente, se podría producir un deterioro o una degeneración de esa escritura, con lo que la diversidad lingüística constituiría una amenaza o un obstáculo para su desarrollo. Sin embargo, voy a mostrar en este artículo que lo cierto es exactamente lo contrario. Las escrituras adoptadas se adaptan a las nuevas lenguas que transcriben y, gracias a ello, hacen evolucionar la propia escritura. La diversidad lingüística no sólo no ha impedido la evolución de la escritura, sino que ha sido, sin lugar a ninguna duda, uno de los factores que la ha potenciado e impulsado.

La xenografía

Cuando se diseña un sistema de escritura para una lengua concreta, se piensa lógicamente en esa lengua y no en otra. Sin embargo, la diversidad lingüística está ya presente desde el principio porque será necesario transcribir no solo las palabras de esa lengua, sino también las de otras lenguas de comunidades con las que se mantiene o ha mantenido algún contacto amistoso u hostil. Eso implica transcribir, como mínimo, topónimos o antropónimos y determinadas palabras o expresiones características. Este fenómeno lo denomino *xenografía interna* (Moreno Cabrera 2016: 117-128). Cuando en castellano escribimos un nombre propio francés utilizamos la ortografía francesa, no la española; de este modo escribimos *Rousseau*, con lo cual, solo sabiendo cómo se pronuncia en esa lengua, podemos leer ese antropónimo adecuadamente. En este caso, no se echa mano de la xenografía interna. Pero en otras lenguas sí se hace esto; por ejemplo, en ruso los nombres propios extranjeros se transcriben haciendo uso xenográfico interno del alfabeto cirílico: Руссо [Russó] transcribe Rousseau. Lo propio se hace en japonés: ルソー [Rusō]. Estos son casos de xenografía interna: se utiliza la escritura pensada para una lengua con el fin de transcribir palabras de una lengua diferente. Si el castellano recurriera a este método en este caso escribiría *Rusó* para nombrar a este filósofo francés. Este mismo procedimiento utilizó la RAE en su propuesta para escribir la palabra inglesa de origen irlandés *whisky* como *güisqui*. En ruso se transcribe por xenografía interna como виски [viski] y en japonés como ウイスキー [wisukī].

La *xenografía externa* (Moreno Cabrera 2016: 117-128), que se ha denominado también *aloglotografía* (Kornicki 2018: 36), se produce cuando una comunidad iletrada

decide tomar la escritura de otra comunidad, cuya escritura puede haber sido adoptada y adaptada de la de otra comunidad o puede haber sido inventada por ella. En muchos de estos casos se trata de utilizar una escritura específicamente diseñada para transcribir una lengua con el fin de visualizar otra lengua con un sistema fonológico, morfológico y léxico diferente. Ello plantea problemas lingüísticos no siempre fáciles para solucionar los cuales es necesario aplicar casi el mismo grado de ingenio e inventiva que el necesario para inventar un sistema de escritura nuevo. Por ello, una comunidad que adapta una escritura ya existente por no haber creado una propia para su lengua, en ningún caso puede concebirse como menos inteligente o ingeniosa que la comunidad que ha inventado la escritura que se toma prestada, tal como vamos a ver en los casos concretos que expongo en las secciones siguientes.

En la xenografía entran en juego palabras escritas sin sentido para la comunidad en la que se produce este fenómeno. El apellido castellano *Zapatero* tiene un sentido para las personas castellanohablantes pero no significa nada para las anglohablantes. Para ellas, el sonido asociado a ese nombre escrito es lo que cuenta. Lo mismo puede decirse del apellido inglés *Smith* para las personas castellanohablantes. Si transcribimos por xenografía interna ese apellido inglés mediante *Esmiz*, pongamos por caso, es evidente que la fonética del apellido es lo que está en el primer plano en la xenografía interna. Tenemos, pues la siguiente ley de la fonetización de los sistemas de escritura:

> Los procesos xenográficos provocan la progresiva fonetización de los sistemas de escritura que tienen un fundamento léxico.

Esta tendencia es la que determina la evolución de los sistemas de escritura de base léxica hacia escrituras de base fonética.

Las unidades lingüísticas en el origen de la escritura

Todos los sistemas de escritura conocidos, en su origen, se organizan en torno a las dos unidades básicas de las lenguas correspondientes a la primera y segunda articulaciones del lenguaje humano: la palabra natural y la sílaba respectivamente. Por ello, se puede decir que los sistemas de escritura originales son logosilábicos: es el caso de las escrituras egipcia, sumeria y china.

En todas las lenguas encontramos dos tipos de palabras naturales: las léxicas y las gramaticales. En una expresión sencilla como *el coche de María,* las palabras *coche* y *María* son léxicas y las palabras *el* y *de* son gramaticales. Las escrituras egipcia, sumeria y china, tres de las más antiguamente atestiguadas, son claramente logosilábicas: los grafemas se interpretan como sílabas y/o como palabras léxicas o gramaticales. Algunas personas especialistas en sinología califican la escritura china de *morfosilábica* (Rovira 2010: 63-68). Es cierto que en chino hay palabras léxicas como *rén* ‘ser humano, persona’ o *shān* ‘montaña, colina’ y palabras gramaticales como el marcador de aspecto perfecto tipo *le* o el morfema relacionante *de*. Por ello, en lingüística se habla de palabras léxicas y palabras gramaticales, aquellas sin contenido léxico cuya función es exclusivamente gramatical; los morfemas son siempre gramaticales, no léxicos y solo aparecen afijados a una palabra léxica. Eso es precisamente lo que ocurre con los morfemas chinos *le* y *de* que acabo de mencionar; pero no sería adecuado lingüísticamente calificar *rén* y *shān* como morfemas: son claramente lexemas o elementos léxicos. Por ello, sería más adecuado calificar la escritura china como *logomorfosilábica*. Con todo, como en lingüística se suele distinguir entre palabras léxicas y gramaticales, podemos perfectamente recurrir al término *logofonográfico* (Moreno Cabrera 2005: 82-86) o al más tradicional *logosilábico* (Moreno Cabrera

2021: 227-232) para caracterizar la escritura china, ya que resulta mucho menos chocante terminológicamente decir que en chino *le* y *de* son palabras gramaticales que afirmar que *rén* y *shān* son morfemas, por lo menos desde la ciencia lingüística.

La escritura china y la sinografía

La escritura china, sin duda, es una de las más influyentes del mundo. Z. Handel (2019: 10) propone el término *sinografoesfera* para denotar aquellas áreas en las que la influencia de la escritura china ha sido determinante. La adaptación de la escritura china para escribir coreano, japonés o vietnamita puso en juego numerosos mecanismos fonéticos, léxicos y gramaticales que modificaron esa escritura de diversos modos, siendo el caso del japonés el más interesante y radical (Taylor & Taylor, 2014: 255-361, Handel 2019: 166-211, Li 2020: 87-100).

En la sección anterior dije que la escritura china es logosilábica. En efecto, cada carácter chino se asocia a una sílaba y, con muy pocas excepciones, a un significado, ya sea éste léxico o gramatical. He aquí los caracteres de las palabras chinas mencionadas en la sección anterior:

Palabras léxicas
人 *rén* ‘ser humano, persona, gente’ / 山 *shān* ‘colina, montaña’
Palabras gramaticales
了 *le* / 的 *de*

Una cuestión que surge inmediatamente es cómo se pueden transcribir en la escritura china los términos extranjeros. Se trata de lo que denomino *xenografía interna*. Pues bien, lo que se hace es usar los caracteres en su valor silábico solamente. Para ello, se eligen aquellos caracteres cuyas sílabas se asemejan más a las sílabas de la palabra extranjera que se desea transcribir. Con ello, se obtiene una expresión semánticamente mal formada o absurda, con lo que pasa a primer plano el valor puramente fonético. Veamos un ejemplo concreto para ilustrar este punto. Para transcribir *Catalunya* en la escritura china, se seleccionan caracteres asociados a sílabas chinas que se parecen fonéticamente a las cuatro sílabas de la palabra catalana. Como la estructura silábica del catalán es muy diferente de la estructura silábica de la lengua china, el chino recurre a cinco caracteres en vez de los cuatro esperables. La sílaba catalana escrita *nya* contiene una consonante nasal palatal no existente en chino, para imitar la cual se recurre a la secuencia de las dos sílabas *ní* y *yǎ*, por lo que obtenemos cinco caracteres correspondientes a cinco sílabas en vez de las cuatro sílabas de la palabra catalana. He aquí la transcripción china de *Catalunya* con la indicación de los significados originales de los caracteres empleados:

加泰啰尼哑
Jiā tài luō ní yǎ

La palabra *jiā* significa ‘sumar, añadir, *tài* significa ‘seguro, pacífico’, *luō* es un componente puramente fonético sin significado, *ní* significa ‘monja budista’ y *yǎ* significa ‘mudo, ronco’.

Como puede comprobarse, esta expresión china no tiene sentido coherente alguno, por lo que sólo se puede interpretar fonéticamente.

La xenografía interna, una clara consecuencia de la diversidad lingüística, hace que se produzca una interpretación puramente silábica de una escritura logosilabográfica. Es decir, la transformación ocasional de una escritura logosilabográfica en silabográfica. Esta potencialidad de la escritura china es explotada de modo sistemático en el caso de la adaptación de la escritura china a la lengua japonesa, un caso realmente notable de xenografia externa. Esto lo vamos a ver a continuación.

Un ejemplo notable de xenografía externa: la escritura japonesa

La adopción y adaptación de la escritura china para escribir japonés es sin duda una tarea compleja y difícil, dado que la lengua china y la japonesa tienen estructuras fonológicas, léxicas, morfológicas y sintácticas muy diferentes. Sin embargo, ese proceso ha sido el origen de uno de los sistemas de escritura más interesantes y fructíferos que se hayan ideado en la historia de la humanidad.

Al usar los caracteres chinos para escribir palabras japonesas se produce frecuentemente la circunstancia de que un lexema chino que consta de una sola sílaba se corresponde con un lexema japonés de más de una sílaba. Veamos un ejemplo concreto. Sea el lexema chino que significa 'montaña' que vimos en la sección anterior:

山 *shān* 'colina, montaña'

Al utilizar el logograma chino para escribir la palabra japonesa correspondiente, que es *yama*, obtenemos:

山 *yama* 'colina, montaña'

Con esto se contraviene el principio que rige la escritura china, según el cual cada carácter se asocia con una sílaba. En este caso, al *kanji* (logograma chino) para *montaña* se le asocia la palabra bisílaba *yama*. Con ello, se produce la conversión del sistema logosilabográfico chino en un nuevo sistema logográfico puro, en donde cada carácter se asocia con una palabra, independientemente del número de sílabas que tenga. Se ha despojado la escritura china a de su base fonética, dejando únicamente su aspecto logográfico. Con ello, se obtiene una escritura logográfica pura.

Además, en japonés los *kanji* no sólo tienen la lectura japonesa, denominada *kun,* sino que también retienen su lectura original china, denominada *on* (Taylor & Taylor 2014: 276-278). Esta última aparece habitualmente en palabras compuestas. De este modo, el logograma para 'montaña' se puede leer como *yama* (lectura *kun*) o como *san* (lectura *on*). Por ejemplo, el monte Fuji se escribe en *kanji* de la siguiente manera, en donde el último carácter se lee en su lectura china *on*:

富士山
Fuji san

Pero con esto no acaba la adaptación de la escritura china para escribir japonés, ya que se simplificaron algunos caracteres chinos para crear dos silabarios, con lo cual se explota la vertiente fonética de la escritura china.

Estos silabarios procedentes de logosilabogramas chinos simplificados tienen una interpretación exclusivamente fonética. Los dos silabarios se denominan *hiragana*, que procede de versiones cursivas simplificadas de caracteres chinos y *katakana*, resultado de versiones simplificadas de los caracteres chinos. El *hiragana* se utiliza para escribir los morfemas gramaticales, muy abundantes en la lengua japonesa; el *katakana* se utiliza, entre otras cosas, para escribir palabras extranjeras. Es decir, se utiliza para llevar a cabo la xenografía interna japonesa. Veamos cómo se puede escribir en japonés *Catalunya*:

カタロニア
Ka ta ro ni a

Se utiliza el silabario *katakana*. Como en el caso de la escritura china vista en la sección anterior, hay cinco y no cuatro silabogramas, dadas las diferencias entre las estructuras silábicas del catalán y del japonés.

La escritura japonesa es entonces una combinación de una logosilabografía (lectura *on*

de los *kanji*, usada en palabras compuestas de origen chino), una logografía pura (lectura *kun* de los *kanji*, usada en palabras simples propiamente japonesas) y dos silabarios de naturaleza moraica, ya que se diferencian las sílabas breves (una mora) de las largas (dos moras): tanto las vocales como las consonantes largas se anotan mediante grafemas especiales. Esto explica que haya un grafema para el sonido [-*n*] de coda silábica.

Para ilustrar lo anterior veamos el siguiente ejemplo:

日本語
に ほん ご
Ni hon go
'Lengua japonesa'

En la primera línea tenemos la expresión 'lengua (*go*) japonesa (*Nihon*)' en la logosilabografía basada en la lectura *on* o china de los *kanji*. Se observa que a cada carácter le corresponde una sílaba: *ni*, *hon* y *go*. Pero dos de esos mismos caracteres, en su lectura *kun*, no son monosilábicos.

El segundo carácter se lee *moto* 'origen, raíz'; y el tercero *kataru* 'conversación, narración, tema'.

En la segunda línea podemos ver la misma palabra en silabario *hiragana*. Se observa que hay cuatro grafemas y no tres como sería esperable dado que *nihongo* consta de tres sílabas. Pero la sílaba *hon* consta de dos moras (*ho-* y *-n*) y, por tanto, se compone de dos grafemas.

En su detallado estudio sobre la adaptación de la escritura china para escribir japonés, coreano y vietnamita, Handel (2019) muestra cómo la diversidad tipológica entre las lenguas implicadas es uno de los factores que determinan las transformaciones más radicales de esa escritura. El japonés es una lengua muy distinta del chino tipológicamente hablando, tanto en los aspectos fonéticos, como morfológicos y sintácticos. Este es uno de los factores que provocó la transformación de la escritura china en tres escrituras: una logográfica pura (*kanji*) y dos silabográficas (katakana y hiragana). Como hemos visto, la logografía pura no ha eliminado la logosilabografía china en japonés por lo que ambas conviven en la escritura japonesa, una de las escrituras más ricas y fascinantes del mundo.

La diversidad lingüística y el origen del alfabeto

La escritura cuneiforme logosilábica fue inventada por los sumerios hacia el 3200 antes de Cristo y se utilizó en la Antigüedad para transcribir numerosas lenguas: el acadio, el hitita, el eblaíta, el elamita, el hurrita, el luvita o el urartiano. Una de las derivaciones más interesantes de la escritura cuneiforme es el alfabeto ugarítico del siglo XIV antes de Cristo, en el que se utilizan las técnicas del trazado de grafemas de la escritura cuneiforme.

Es esta dinámica xenográfica, en la que una escritura fue adoptada y adaptada por lenguas diferentes de aquella para la que se inventó, lo que dio origen al sistema de escritura más difundido en la actualidad.

La mayor parte de los alfabetos hoy en uso proceden del alfabeto griego. ¿Cómo surgió este alfabeto? Fue una adaptación del consonantario fenicio. El fenicio es una lengua semita y el griego una lengua indoeuropea. Los sistemas fonológicos de ambas lenguas son muy diferentes y, precisamente de la confrontación de esa diferencia surgió el alfabeto. Es decir, la diversidad lingüística está una vez más en el origen de la evolución de los sistemas de escritura.

El consonantario fenicio se inventó hacia el siglo XII antes de Cristo y consta tan sólo de

22 letras o fonemogramas, que transcriben consonantes.

¿Por qué no hay letras para las vocales? Se han dado explicaciones muy fantasiosas e inocentes para contestar esta pregunta. Pero la respuesta es muy sencilla y no se debe a un déficit o imperfección de la escritura fenicia, sino a todo lo contrario. Basándose en el principio acrofónico, cada letra representa la consonante inicial de las sílabas. Sabemos que en esta posición silábica se da la totalidad o casi totalidad de las oposiciones fonológicas de una lengua, ya que en la posición de final o coda silábica se dan muchas menos oposiciones fonológicas. Por ejemplo, en castellano la oposición entre los fonemas /n/, /m/ y /ñ/ se verifica en cabeza silábica, no en coda silábica.

Curiosamente, en las lenguas semíticas, entre las cuales está el fenicio, ninguna sílaba empieza por vocal, porque estas siempre están precedidas en esta posición por una oclusión glotal. Por consiguiente, en coherencia con el principio acrofónico, no hay letras para las vocales en las lenguas semíticas. Precisamente, en el alfabeto cuneiforme ugarítico, lengua también semítica, hay grafemas para las vocales, que incluyen la oclusión glotal. Es decir, no se transcriben las vocales /a/, /i/, /u/, sino /ʾa/, /ʾi/, /ʾu/, en donde la comilla superscrita indica la oclusión glotal.

En el consonantario fenicio, la letra <A>, la primera de la lista, transcribe precisamente la oclusión glotal y no la vocal /a/.

Según Powell (1991) la adaptación del consonantario fenicio para escribir griego fue llevada a cabo por una persona de la isla de Eubea hacia el 800 antes de Cristo, que consultó a un informante fenicio para poner por escrito los poemas homéricos. Independientemente de la veracidad o verosimilitud de esta hipótesis, lo cierto es que quien adaptó el consonantario fenicio al griego utilizó las letras de los sonidos consonánticos fenicios inexistentes en griego para transcribir las vocales griegas. Los casos de la letra A y de la letra O son muy ilustrativos al respecto. La letra A en el consonantario fenicio transcribe la oclusión glotal. Esta consonante no es audible en sí misma, pero afecta a la vocal a la que precede dándole un inicio cortante o abrupto, como ocurre, por ejemplo, en el alemán *ein* ‘uno’, que va precedido de una oclusión glotal no escrita. Es más que probable que la o las personas que adaptaron la escritura fenicia al griego no identificaran la oclusión glotal asociada a la letra A y la utilizaran para transcribir la vocal /a/ del griego. Lo mismo se puede decir de la vocal O del consonantario fenicio. Transcribía una consonante fricativa faringal sonora. Esta consonante, típica de las lenguas semíticas, no existía en griego antiguo y su localización posterior pudo llevar a las personas que hicieron la adaptación a utilizarla para transcribir la vocal /o/.

No es plausible la opinión de que en Grecia se inventaron las vocales para construir el primer alfabeto y por tanto que fue el pueblo griego quien inventó el alfabeto. Hay que tener en cuenta que los griegos nunca inventaron una escritura propia, a pesar de constituir su civilización la base de la filosofía, de la ciencia y de la literatura de toda la cultura occidental actual. Los testimonios más antiguos del griego datan del siglo XVI antes de Cristo y están escritos en un silabario conocido como lineal B, una escritura que se tomó prestada de la escritura minoica conocida como lineal A. Esta escritura dejó de usarse a partir del siglo XII a. de C. y pasaron varios siglos hasta que los griegos volvieron a adaptar una escritura para transcribir el griego, como ya he explicado. La invención de una escritura, por consiguiente, en modo alguno puede considerarse como un signo necesario de excelencia cultural de una civilización.

Un nuevo enfoque de la historia de la escritura

La idea dominante hoy en día de cómo surgió y evolucionó la escritura en la historia de la humanidad fue propuesta por I. Gelb en los años cincuenta del siglo pasado en su monografía sobre la historia de la escritura (Gelb 1963). Según este autor (Gelb 1963: 191), la escritura evolucionó desde el estadio inicial logosilabográfico de las escrituras egipcia, sumeria y china pasando por el estadio silabográfico de la escritura fenicia y japonesa hasta llegar al estadio alfabético de la escritura griega. Uno de los inconvenientes de esta propuesta es que la escritura fenicia no es silabográfica, sino fonemográfica y, además los egipcios ya incluían en su escritura logográfica una serie de elementos fonemográficos. Es decir, la fonemografía, que es la base del alfabeto, ya existía en las escrituras logosilabográficas de la Antigüedad. Una cuestión que a veces surge es que si en los sistemas de escritura antiguos como el egipcio había la posibilidad de transcribir sonidos aislados, fonemas o sílabas, ¿por qué no se generalizó a expensas del sistema logosilábico, mucho más complejo y difícil de aprender? La respuesta tiene que ver con la función religiosa y socio-cultural de la escritura en las sociedades antiguas, en donde la escritura estaba restringida a una casta más o menos privilegiada y no estaba pensada en modo alguno para su uso generalizado por parte del pueblo llano.

Dentro del sistema educativo chino clásico, el sistema de exámenes para los funcionarios públicos exigía una dedicación que estaba al alcance de muy pocas personas

> "In their long existence, the Confucian classics – the Five Classics and the Four Books – have received much commentary by scholars, and have swollen to a massive body of literature, containing some 500,000 characters. Generations of scholars studied and memorized it in preparation for a series of civil service examinations. Its long and arduous study promoted scholarly literacy among a tiny upper crust of males while preventing the spread of functional literacy among the masses." (Taylor & Taylor 2014: 89)[1]

La tradición literaria china, determinante en buena parte de toda el área cultural de influencia china, tuvo precisamente este carácter elitista:

> "Sinitic remained, after all, a sacred language in the context of Buddhist scriptures throughout east Asia, and the Chinese classics were not only the language of the civil service examinations and the bureaucracy, but they were also constantly being re-examined, re-interpreted, and re-oriented to deal with present problems." (Kornicki 2018: 185)[2]

La invención de una escritura es una de las hazañas intelectuales humanas más notables e importantes. Las sociedades de la Antigüedad que crearon los sistemas de escritura estaban tan avanzadas intelectualmente como las sociedades actuales. La evolución de la escritura desde una etapa logográfica hacia un estadio fonemográ-

---

1.- Nota del traductor (N. del T.): "En su larga existencia, los clásicos confucianos —los Cinco Clásicos y los Cuatro Libros— han recibido muchos comentarios de parte de eruditos, y han crecido hasta una literatura enorme, constando de unos 500.000 caracteres. Generaciones de eruditos los han estudiado y aprendido de memoria para prepararse para una serie de oposiciones públicas. Su largo y arduo estudio promovió una alfabetización erudita entre una diminuta élite masculina mientras que impedía la propagación de una alfabetización funcional entre las masas."

2.- N. del T.: "Al fin y al cabo, la lengua sinítica seguía siendo una lengua sacra en el contexto de las escrituras budistas por todo Asia Oriental, y los clásicos chinos no solo eran la lengua de las oposiciones públicas y la burocracia, sino que también eran constantemente reexaminados, reinterpretados, y reorientados para resolver problemas presentes."

fico no tiene nada que ver con un supuesto desarrollo mental o intelectual más avanzado de las sociedades modernas frente a las antiguas. Aunque la sociedad griega de la Antigüedad nunca inventó una escritura propia, las personas que desarrollaron las bases de la filosofía y de la cultura occidental actual, entre ellas Sócrates, Platón o Aristóteles, era tan inteligentes como las de las sociedades industrializadas actuales.

La xenografía o aloglotografía ha sido, precisamente, uno de los mecanismos que ha propiciado la evolución de la escritura, tal como he intentado mostrar en este breve artículo.

Conclusión

La diversidad lingüística es una de las manifestaciones más conspicuas de la impresionante adaptabilidad y creatividad de ese curioso animal que denominamos *ser humano*. Como el lenguaje humano, la diversidad de las lenguas es una marca inconfundible de la propia naturaleza humana. La escritura es otra muestra de nuestra creatividad y adaptabilidad. La confluencia entre la escritura y la diversidad lingüística ha sido precisamente, tal como he pretendido probar en este artículo, uno de los factores que ha hecho evolucionar la propia escritura para hacerla cada vez más fonética y más accesible para la gente en general. Esta confluencia no ha sido en modo alguno un factor negativo o contraproducente sino todo lo contrario, tal como hemos visto.

A partir de aquí podemos extraer una importante lección sobre la naturaleza de la diversidad lingüística. Desde el mito de la torre de Babel se ha visto la diversidad lingüística como un factor de incomunicación, de aislamiento, de incomprensión o de hostilidad. Esta actitud es típica del imperialismo lingüístico y de las ideologías hegemónicas en general. Pasar por alto la importancia de la diversidad lingüística para comprender el desarrollo cultural de la humanidad puede ser un efecto colateral de esta actitud. He pretendido mostrar en este artículo que no podemos dejar de lado la diversidad lingüística para entender la invención y desarrollo de una tecnología clave en la historia de la humanidad: la escritura.

Referencias

Gelb, Ignace. 1963. *A Study of Writing.* Chicago: The University of Chicago Press.

Handel, Zev. 2019. *Sinography. The borrowing and adaptation of the Chinese script*. Leiden: Brill.

Kornicki, Peter Francis. 2018. *Languages, Scripts, and Chinese Texts in East Asia.* Oxford: Oxford University Press.

Li, Yu. 2020. *The Chinese Writing System in Asia. An interdisciplinary Perspective*. London: Routledge.

Moreno Cabrera, Juan Carlos. 2005. *Las lenguas y sus escrituras. Tipología, Evolución e Ideología*. Madrid: Síntesis.

Moreno Cabrera, Juan Carlos. 2016. *Multilingüismo y lenguas en contacto*. Madrid: Síntesis.

Moreno Cabrera, Juan Carlos. 2021. *La clasificación de las lenguas. Introducción a la taxonomía lingüística*. Madrid: Síntesis.

Rovira i Esteva, Sara. 2010. *Lengua y escritura chinas: mito y realidades*. Barcelona: Bellaterra.

Powell, Barry. 1991. *Homer and the origin of the Greek alphabet*. Cambridge: Cambridge University Press.

Taylor, Insup & Taylor, M. Martín. 2014. *Writing and Literacy in Chinese, Korean and Japanese.* Amsterdam: John Benjamins.

LINGUISTIC DIVERSITY IN THE AGE OF BIG DATA

Tunde Adegbola
African Languages Technology Initiative (Alt-i)

Introduction

Language is one of the natural endowments of the human race, and languages come as many as there are bases for distinction between cultures. As a cultural heritage, language stands as a fundamental basis for expressing differences in cultures. Even though language is used primarily for the purpose of communication it also serves the equally important purpose of identity. Hence, cultures and communities may be defined by geography and other spatial realities, but more often than not, language stands as an important intangible heritage based on which a people is defined and known.

Diversity is one of the most important forces of nature. It manifests at vital levels that ensure the sustenance of the natural world by its role in bringing about the necessary variety required for balance in nature. It permeates the whole of nature, and it can be observed at various levels in the natural world. Probably the most obvious level of diversity in nature is biodiversity but abiotic diversity is also as important as biodiversity. Biodiversity describes the multitudes of living organisms, consisting of an estimated 8 million fungi, bacteria, plants, and animals. According to National Geographic, only about 1.2 million of these have been identified. The power of diversity derives from the fact that each element in a diverse ecosystem contributes something unique to the integrity of the ecosystem.

Linguistic diversity

Human beings are social animals. Hence, they live in and as communities. Humans communicate as members of communities, building their communities naturally around the issues they communicate, and how they communicate these issues, guided by the intrinsic linguistic structure of the human mind. For this reason, human languages evolve naturally in response to the need of humans to express themselves within their environments and the structure of their minds. All humans face the same primary problems of survival and sustenance, and therefore discuss essentially similar issues using the same set of speech apparatus. It is expected therefore that various human languages will manifest high levels of similarity. The similarities in the issues they discuss and the apparatus with which they express them are important factors responsible for the similarities in languages. However, random perturbations in the way humans express themselves, as well as the peculiarities of their various environments tend to manifest structurally in their language, taking what would have remained single languages in diverse directions due to separation in time and space between two or more speaker groups of a single language. Skutnabb-Kangas observed that countries with mega-diversities have had more varied micro-environments to observe, analyse, describe, and discuss than countries with less diversity, and all of these knowledges have been encoded in their many languages (Skutnabb-Kangas 2002). Hence, both the similarities and dissimilarities that manifest in human languages are products of features that characterise the human mind as well as the human environment. According to Lera Boroditsky (2018), linguistic diversity reveals the ingenuity of the human mind in its invention of 7000 cognitive universes, each one corresponding to one of the languages spoken somewhere or the other in the world.

Linguistic diversity serves an important purpose of characterising the structure of the human mind as well as the methods by which humans adapt to their environment in critically diverse ways. In addition, the sheer diversity of the ways by which humans express themselves in language might provide insights into other aspects of diversity of the human environment. Hence, the possibility of a contemplative, non-destructive manipulation of language as an ideational rather than material reality may provide facilities for studying these other aspects of natural diversity as analogues of human language.

The intrinsic value of diversity is felt at all levels of diversity encountered in the natural world. Language is a vital component of the human cognitive system and every element of knowledge that humans encounter is coded in some language or the other. Linguistic diversity therefore remains one of the forces of nature that give support to an understanding of the diverse variabilities in, as well as the unity of, the natural world.

The importance of biodiversity is relatively easily felt and well understood, and awareness about it is now quite widespread, but the importance of linguistic diversity is much less appreciated. The equal importance of linguistic diversity may however be realised if we note that efforts to maintain and promote biodiversity need to be supplemented with efforts to also maintain and promote linguistic diversity if the efforts towards the maintenance and promotion of biodiversity is to be maximally useful. Biodiversity produces a wide variety of material reality that call upon language to label and describe them. The labels and descriptions bestowed upon these elements of material reality by language are usually drawn from relevant characteristic features of such materials. These characteristic features may be latent, but the labels and descriptions given to them serve as vital pointers to their potentialities. Very much like what software represents to hardware, biodiversity without the corresponding linguistic diversity is bound to deny the global knowledge community of the potentials of the myriads of material realities localised in specific areas within the ecology of our biosphere.

Factors that work against linguistic diversity

The quest for survival of living organisms usually puts them in paths of conflict both within the same group and also between different groups. Such conflicts sometimes lead to domination causing certain groups to be exterminated, either materially or ideationally. In human societies such domination might manifest in genocide, with the end result of the obliteration of a dominated people. More often however, it manifests as a gradual recession in the culture of the dominated due to an imposition of the culture of the dominator. The recession of a culture is usually accompanied by the recession of the language that is supposed to give vitality to the culture. The colonial experiences in various parts of the world speak volumes to the recession of languages due to domination by a foreign culture. Apart from colonisation, various levels of interactions between neighbouring communities, in which one culture becomes dominated by another and the language that expresses the dominated culture falls victim to the domination, are common.

Beyond these traditional causes of language endangerment, which tend to work against linguistic diversity, the development, deployment, and exploitation of modern Information Communication Technologies (ICTs) have of recent become yet another potent force that may work against linguistic diversity. From the point of view of basic intuition, by virtue of the dependence of communication

on information and information on language, an information communication technology is expected to work to promote the cause of information communication through language and thereby contribute to language development. However, modern ICTs on the whole have the potential to either promote or work against linguistic diversity depending on the way they are employed. The purpose of this article is to look critically at contemporary trends in ICT development, particularly the development and utilisation of big data in Natural Language Processing (NLP), analyse its possible negative effects on linguistic diversity and consider ways of mitigating the possible negative effects.

ICT and linguistic diversity

Developments in digital technology have served the purpose of producing and promoting information communication technologies as important tools that extend the reach of human communication capacities. Both information and its communication rest squarely on language, and the proliferation of information and communication technology devices and their widespread use in human communication are bound to have profound effects on languages and their development. To start with, languages for which information communication technology devices have been designed to work will be at an advantage over languages which these devices do not respond to yet. Hence, the state of ICT development as it concerns the languages to which ICT devices have not been adapted will have implications for language development and the survival of languages.

Intergenerational communication is probably the most important factor of language survival. As long as parents speak to their children in the language of their culture, such a language stands a chance for survival. However, if parents are constrained to speak to their children in a foreign language, the survival of their own language falls into jeopardy. The circumstances that may constrain parents to speak to their children in a foreign language are many but the one that is of importance to this article is the effect of modern ICT.

We now live in a world in which humans communicate with one another more and more through machines. In addition, more and more, humans now also have the need to communicate with machines. Young people, being cybernatives embrace ICT devices with enthusiasm and as it were, now live most of their lives through these devices in cyberspace. Rather than living fully in the real world of mass and volume, they now live an appreciable portion of their lives in the virtual world of social media, interacting more with their peers in the virtual world of global social media and less with their parents in the real world of mass and volume, the only world that their parents know well. This implies that young people's language may be determined more by the interaction with their peers on global social media than by interactions with their parents within a local geographical space. This is the first level at which ICT may affect language development. To make matters worse, if the devices they use to interact with their peers are not well adapted to the language of their parents from whom they are supposed to inherit their language, not only will they be disadvantaged in their interactions with their peers, but their interactions with their parents too will be negatively affected. Invariably, these young people will expect their parents to be able to communicate with them through ICT devices. However, their parents, being cybermigrants, continue to struggle with these devices, hopefully with some form of guidance and encouragement from the children. If, however, these ICT devices are ill-adapted to the language that

parents seek to bequeath to their children, the intergenerational facility of language survival becomes compromised. Parents in such a culture are constrained to communicate with their children in a foreign language, putting their indigenous language in jeopardy and thereby making it difficult for parents to communicate important nuances of their cultural environment to their children. It stands to reason therefore, that beyond the traditional factors that jeopardise the survival of languages, any language that modern ICT devices have not been designed to work in is in danger of extinction.

There are some ICT devices that are language agnostic, and the adoption rate of such technologies are a testimony to the importance of language to the adoption of information and communication technologies. A good example of a language agnostic information communication technology is the telephone as a speech-based communication device. The high adoption rate of the telephone around the world bears eloquent testimony to the language agnosticism of the telephone as a speech-based device. However, with the progressive integration of communication technologies with information technologies, the information communication scope of the "plain old telephone service" (POTS) has widened to accommodate many "pretty amazing new services" (PANS) to the extent that even intergenerational communication through these devices now demands the use of more than mere speech. Communicating on such devices with more than mere speech and taking full advantage of the facilities they offer, at the most basic level, demands the development of locales, keyboards, spell checkers and grammar correction facilities. At a more advanced level, there will be a need for speech recognisers, speech synthesisers, machine translators and other amenities that facilitate communication through and with machines.

The development of ICT devices that accommodate the myriads of languages they may need to be adapted to, presents a chicken and egg situation for most of the 7000 languages spoken in the world today. Apart from many languages of Europe and some of the languages of Asia, most of the other languages of the world would at best be seen as resource-scarce languages. A resource-scarce language in this context is a language lacking in electronic resources needed to be used to develop human language technology relevant to the language in question. Some of the most basic resources required for the development of human language technology for any language include written language corpora, spoken language corpora, monolingual dictionaries, and bilingual dictionaries, all in computer readable form. Also of relevance are terminology collections, grammars, taggers, morphological analysers, and parsers as well as speech recognisers, text-to-speech synthesisers, annotation standards, annotation tools, corpus exploration and exploitation tools as well as bilingual corpora for training machine translators. These and more, usually referred to collectively as the Basic LAnguage Resource Kit (BLARK), are the basic computer readable electronic resources required to build ICT devices that can accommodate a given language (Krauwer, 2003). Unfortunately, however, creating the BLARK for a language from scratch requires the availability of some of the elements of BLARK itself, bringing about the already mentioned chicken and egg situation, a confounding situation in which a chicken is needed to lay the egg and an egg is needed to hatch the chicken. The solution to such a problem involves a strategy that turns the proverbial vicious cycle into a virtuous spiral. An iterative cyclic process in which circular motion on a two-dimensional plane is nudged along a third dimension thereby turning what would have remained a circle into a spiral that features some incre-

mental improvement even if only marginally during every cycle of movement.

Fundamental to BLARK is the locale. As a term in computing, the locale specifies the language environment of a software through which user interface parameters such as the formatting of dates and times, display format for numbers, currency symbol and other such parameters specific to the region of the world in which a language is used are defined. To develop the locale relevant to the culture of a given language, it may be enough to undertake a careful examination of the culture in question and benchmark its locale needs against the elements of some existing locales. With a bit of insightful consideration, it will be possible to develop a credible locale for a culture and language this way. To develop an efficient keyboard layout for a language, it may be necessary to consider the statistical distribution of the characters defined by the orthography of the language and thereby place the more frequently used characters in positions where they can be easily reached by the more convenient fingers. For various historical reasons, this might have not been done in the design of the QWERTY keyboard layout, but every opportunity to take advantage of ergonomics should be positively exploited. To obtain this statistical distribution, we need a sizeable corpus of a computer readable text and to build a sizeable corpus of computer readable text we need an efficient keyboard layout and therein lies the chicken and egg circumstance of the situation. The strategy to turn this seeming vicious cycle into the earlier described virtuous spiral is to start with a tentative modest corpus developed on an existing though inefficient keyboard layout and the tentative modest corpus is used to develop a new layout that improves the efficiency of the existing layout iteratively.

Such a strategy may work well for the development of modest corpora required to develop basic ICT tools such as keyboards and spell checkers. However, to take full advantage of some of the more modern techniques offered by deep machine learning in the development of speech recognisers, speech synthesisers, machine translators and other advanced NLP technologies, there is a necessity to resort to the use of Big Data. These technologies require computers to “learn” by immersion in large volumes of historical data.

Big data and linguistic diversity

Big data describes huge volumes of ever-growing data, accumulated as a result of the storage of data generated in ordinary everyday processes of formal and informal transactions. It may be considered a by-product of these processes because the production of data is not the primary purpose of such transactions. It is characterised by three “Vs”; volume, velocity and variety. As the term itself suggests, volume is the very essence of big data. Of necessity, it involves huge volumes of data that may not yield to efficient storage and processing with traditional data management tools. It therefore demands novel methods for efficient storage and processing. The velocity character of big data addresses the fact that even though it is characteristically of huge volumes, yet it still continues to grow at a rather fast pace thereby making it even more voluminous. Furthermore, even though some of the content of big data may be structured, they are not always necessarily structured because big data features a wide variety of structures as well as unpredictable formats. A typical example of big data is the accumulation of data generated in the stock exchanges around the world. Of course, the primary objective of a stock exchange is not the generation of data but data does get generated anyway, in large volumes, at a fast pace and in a wide variety of structures as well as a lack of struc-

ture and in all sorts of unpredictable formats. For example, the New York stock exchange is said to produce close to a terabyte of data in a single day of trading (Groenfeldt, 2013).

The term Big Data refers not only to the wide variety of data accumulated at high velocity into huge volumes. It also describes the novel frameworks within which the data is stored as well as the tools and techniques employed in the processing and analyses of the voluminous data. Its value derives from the fact that knowledge is embedded in patterns and patterns become apparent only in significant volumes of data. Hence, big data by virtue of its volume, velocity and variety has the capability to reveal hitherto unrecognised patterns that harbour hitherto unavailable knowledge and, in many instances, knowledge unavailable to computers otherwise. This is the reason why big data has played and continues to play a very important role in the development of artificial intelligence.

Recently, big data has emerged as a very important source of linguistic data. With a lot of human activities now taking place on, or at least passing through, the internet, the data accumulated as a by-product of these activities constitute useful data that grows in volume into big data, owing mainly to the velocity and variety at which they are generated. Newsletters, newspapers and magazines are now normally found on the internet. Documentation of governance activities, as well as judicial proceedings and legislative debates, reports of meetings, academic publications as well as public and private opinions now find their ways to the internet with relative ease, turning it into a vital and convenient repository. At less formal levels, both professional and social group chats, as well as many other social activities, have elevated the internet to an arena where many, particularly young people, now live a significant portion of their lives. Hence, all the texts generated as a result of these activities become natural sources of linguistic data, providing an unprecedented depth of insight into the human mind as the source of this data and human language as the vehicle of human thought.

Unfortunately, most of the languages spoken in the world today are not used in domains that accumulate the linguistic data generated in their use into big data. Linguistic data is inadvertently generated in various everyday processes that involve the use of languages. For a few languages, most of the texts generated in these simple everyday interactions leave documentary footprints on the internet, thereby offering an opportunity for the accumulation of large amounts of data according to the volume, velocity and variety conditions that promote accumulation into big data. However, most languages are not used in domains that promote such accumulation, leaving little or no chances for such languages to produce texts that may be accumulated into big data. Languages that fall within this category are bound to be disadvantaged going by the imperatives of the information age.

Before the information age, it was enough for a language to be used in everyday transactions to qualify as a living language. In the information age in which we now find ourselves, however, given the need for humans to communicate through and with machines and the value these have given to big data, for a language to be regarded as an actively living language, the use of such a language must produce documental footprints that can be accumulated into big data to be used to further vitalise the use of the language. This is a major challenge for the overwhelming majority of the 7000 odd languages spoken in the world today. Specifically, the challenge is how to push these languages into use in domains where their use will not be tantamount to linguistic “gas flaring”. Gas flaring

is a process in the petro-chemical industry in which otherwise useful combustible gases such as methane, butane, and propane, produced as by-products, are deliberately flared as a disposal strategy. In the context of language use, we liken the unharvested linguistic data generated in the processes of everyday human interactions to flared valuable gas. These elements of linguistic data that are allowed to vaporise into thin air, whenever undocumented text is generated in the processes of everyday human interaction, constitute valuable linguistic resources but are rendered as waste products because they are not harvested. Considering the importance of big data in the information age, strategies need to be put in place to ensure the capture and accumulation of these valuable resources into useful linguistic data.

A local problem with global implications

Every organism in the biosphere plays important roles in the ecology of the biosphere and such roles are known to have global significance. We may extrapolate this fact to the language space. In this vein, it is not unreasonable to assume that a language may code some important elements of knowledge which may not be available in other languages. This may be due to the fact that certain natural phenomena may be more easily observable in the locality in which such a language is spoken. Based on the significant diversity of languages, all of them seeking to achieve the same purposes of communication and identity, we can make the bold conjecture that every language has the potential of some unique highly efficient information coding strategy. Such unique and highly efficient information coding strategies may offer some ways of coding general or specialised information in very productive but yet unknown ways. The sheer length of time over which human languages have evolved and the success with which modern computing has employed genetic algorithms suggest the viability of such a conjecture. For example, the wide variety of the morphology of various languages suggests that humans have been widely creative in evolving their languages. They have been able to find ways to use a relatively limited finite set of linguistic symbols to represent the infinite number of thoughts that pass through their minds. The wide variety of creative arrangements of this relatively small set of linguistic symbols as combinations of phones into phonemes, phonemes into morphemes, morphemes into words, words into sentences and sentences into discourses to achieve the coding of an infinite set of thoughts remains a feat even if commonplace. There must be a lot of lessons for the global cognoscenti to learn from this naturally evolved feat of humankind.

The analogical correspondences between various widely differing natural phenomena present opportunities for the development of creative information coding strategies. Apparent and thereby simply comprehensible phenomena can be used to understand more complex but analogical ones. Recent developments in Bioinformatics have demonstrated the values of an information processing approach to molecular biology. The use of artificial intelligence in addressing the problem of protein folding, particularly the works of DeepMind in Alphafold suggests a strong linguistic basis of genomics. For all we care, natural language may provide simply comprehensible analogues of some of the complex genomic patterns, structures and processes. Natural Language Processing might offer well-tested algorithms to address the more complex problems that emerge in genomic sequencing. If the languages that most apparently demonstrate the application of some of these algorithms become extinct before such algorithms are discovered, humankind would have lost a very important

element of knowledge. These sorts of analogical explanations cannot be limited only to bioinformatics.

In cognitive science for example, language plays a major role in theorizing the structure of the human mind because language is one intangible and abundant output of the human mind. For this reason, therefore, the study of the phonology, morphology, syntax, and pragmatics of the linguistic output of the human mind gives insight into the structure and the workings of the human mind. Unfortunately, however, all that we know in cognitive science today might have come from about 30 or definitely less than 50 largely similar languages of the 7,000 languages of the world. From a statistical point of view, this constitutes convenience sampling which, based on statistical theory, is not generalisable. There is a need therefore to actively seek to radically widen the base of languages available to studies in cognitive science and other related fields. We might have been probing into the human mind through a few dark and small windows while many other lighter and probably much wider windows are available but yet unexplored.

Hence, there is much greater value beyond the merely romantic though valid sentiments of linguistic diversity and the economic value it offers in the sustenance of local livelihoods. The value of the lessons that the information coding strategy of some obscure natural languages may teach humankind remains yet largely unexplored. The global community is definitely worse off with the loss of any single language, how much more the loss of linguistic diversity. For all we care, the data structure or algorithmic solutions to some still pending scientific problems might have been developed genetically and used in some unknown human language.

The primary responsibility for the development of a language lies on the speakers of the language. First and foremost, it is the responsibility of the speakers of a language to identify domains relevant and useful to their lives and livelihoods and use their language in such domains. However, every natural language is a gift of nature, not only to the speakers of the language but also to global humanity. Hence, every language that nature has chosen to evolve with the collaboration of humans within a culture holds important treasure for global humanity. The sheer variety of the phonology, morphology, syntax, and semantics of each of the 7000 odd languages of the world speak to the vitality of the principle of multiple realisability in the philosophy of mind. Every language therefore may be perceived as one viable path between a small set of finite symbols and an infinite set of culturally useful texts in a humongous search space. The exceedingly large search space of ways of organising this small finite set of symbols into a system capable of coding infinite expressions of information, knowledge and wisdom has been searched heuristically since the emergence of homo sapiens. Having found some of these viable paths, it therefore behoves us to ensure that we exploit them maximally. This includes using them in domains other than basic human communication as analogical guides when dealing with more complex natural phenomena.

Accumulating and exploiting big data for linguistic diversity

As has been earlier discussed, for most languages of the world, everyday use does not produce documentary footprints that may be accumulated into big data. Having argued that big data has a lot to contribute to linguistic diversity in the information age, we must explore the possibility of generating as much data as possible from everyday use of most languages of the world.

There are highly varying facilities for various languages. Some, like the official languages of linguistically pluralistic nations, do have regularly generated printed materials available in many languages offering possibility of both written language corpora and parallel corpora. For such languages in which government notices and other public information are available, it is important to deliberately harvest all such printed materials, annotate them appropriately and deliberately accumulate them into some publicly accessible repository. Some languages are used in the production of daily newspapers, regular magazines, and other periodicals. Such must be deliberately harvested and accumulated in computer readable form. Some languages are used in judicial proceedings and legislative debates. These too must be deliberately archived. In some cases, elements of judicial proceedings in indigenous languages are translated and recorded in some foreign colonial legacy languages. Efforts should be made to retranslate such too and document them in the source languages. Keen attention should be given to any everyday uses of any language that has the potential of generating relevant documentation. All these should be accessed and curated in relevant formats.

For very many languages, the most likely opportunity for documentation with little or no extra effort is broadcasting. Broadcasting offers opportunities for the gathering of written language corpora, spoken language corpora as well as bilingual corpora. Unfortunately, however, very few broadcast stations appreciate the value of their broadcast materials beyond the broadcasting objectives of information, education, and entertainment. The value of the accumulated materials in their archives as linguistic data may not be realised. Worse still, in some stations, some important and valuable materials are not considered worthy of storage. For example, in most of Africa, news bulletins are prepared in English or some other colonial legacy languages. Versions of the bulletin are then broadcast in some of the indigenous languages of the relevant locality, translated from the colonial legacy foreign language bulletins. In some situations, the newscast in the indigenous language is translated extemporaneously from the colonial legacy language text on air. In other situations, even though the translation is scripted before casting, the script is written in longhand and so is not rendered in computer readable form. In some of such situations, the paper on which the translation was scripted is discarded immediately after the newscast, precluding any possibility of typing the script and thereby rendering it in computer readable form at a later date. This way, valuable materials for written text corpus and parallel corpus that could have been accumulated into big data is lost every day in this unfortunate process of “linguistic gas flaring”.

Present global trends in community broadcasting offer a unique opportunity of unprecedented access to a significant number of the world’s languages used in broadcasting at community level. This offers invaluable opportunities for the documentation of these languages, both as written and spoken language corpora and their accumulation into big data. Even if the broadcast materials emerge essentially extemporaneously, audio recordings of the broadcast could be properly curated as the basis of documentation. In order to avoid the acts of “linguistic gas flaring” as earlier described and thereby ensure that broadcast materials are available for long term retrieval, proper arrangements need to be made to introduce favourable practices that promote the storage and accumulation of broadcast materials and ensure their easy access for researchers.

Arrangement for harvesting broadcast materials as the basis of big data for linguistic

analysis needs to be strictly formalised and implemented within accountable structures. A relevant existing global structure that may be used to facilitate access to these broadcast materials is the World Association of Community Radio Broadcasters (Association Mondiale des Radiodiffuseurs Communautaires - AMARC). The importance of seeking to accumulate big data for linguistic analysis through such an organisation calls for collaborative planning. The potential data is invaluable and incentivisation for measured volumes of data would not be out of place.

The accumulation of written and spoken language materials into big data is of great importance as a natural resource and it qualifies to be catered for within the national language policies of nations in which "linguistic gas flaring" takes place on a daily basis. Proper documentation and aggregation of both printed and broadcast materials in computer readable forms and formats in an easily accessible repository located within a public institution is a matter of paramount importance.

Conclusion

We now live in the information society of the information age brought about by the Information revolution. In this information society in which we now find ourselves, data is a key resource that fuels all activities. Data is generated in simple interactions that result from everyday transactions but in many parts of the world, the data is not harvested and accumulated into big data; the form in which it becomes an asset, and its value becomes accessible. Language is the foundational substrate on which all information and therefore communication depends. Hence, in the information society, language is of prime importance. Linguistic diversity is a gift of nature to humankind and should not be compromised. The wasteful flaring of the linguistic data generated but not harvested in the simple interactions that result from everyday transactions threatens linguistic diversity. Some of the promises of the information age may depend on linguistic diversity in yet unknown ways. We cannot afford to trade off the benefits that may accrue from these promises on the altar of prevailing ignorance.

References

Boroditsky, L. (2018) How language shapes the way we think. https://www.youtube.com/watch?v=RKK7wGAYP6k consulted September 2021

Krauwer, S., (2003) The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap; http://www.elsnet.org/dox/krauwer-specom2003.pdf, consulted September, 2021.

Miller, G. & Spoolman, S. (2012). Environmental Science – Biodiversity Is a Crucial Part of the Earth's Natural Capital. Cengage Learning. https://books.google.com.ng/books?id=NYEJAAAAQBAJ&pg=PA62&redir_esc=y#v=onepage&q&f=-false, consulted September 2021

National Geographic, Biodiversity. https://www.nationalgeographic.org/encyclopedia/biodiversity/, consulted September, 2021.

Skutnabb-Kangas, (2002) why should linguistic diversity be maintained and supported in Europe? Some arguments. Language Policy Division, Directorate of School, Out-of-School and Higher Education DGIV, Council of Europe, Strasbourg

LA DIVERSITAT LINGÜÍSTICA A L'ERA DE LES DADES MASSIVES

Tunde Adegbola
African Languages Technology Initiative (Alt-i)[1]

Introducció

La llengua és un dels dons naturals de la raça humana, i hi ha tantes llengües com bases de distinció entre cultures. Com a patrimoni cultural, la llengua constitueix una base fonamental per expressar diferències culturals. Tot i que la llengua s'utilitza principalment per comunicar, també té un objectiu igual d'important d'identitat. Així doncs, podem definir cultures i comunitats segons la seva geografia o altres realitats espacials, però generalment, la llengua constitueix un patrimoni intangible important segons el qual es defineix i es coneix un poble.

La diversitat és una de les forces de la natura més importants. Es manifesta a nivells vitals per assegurar la subsistència del món natural mitjançant el seu rol de contribuir a la varietat necessària per mantenir l'equilibri a la natura. Impregna tota la natura i es pot observar a diversos nivells al món natural. Segurament la diversitat més òbvia a la natura és la biodiversitat, però la diversitat abiòtica és també igual d'important. La biodiversitat descriu la immensitat d'organismes vius; s'estima que hi ha uns 8 milions de fongs, bacteris, plantes i animals. Segons National Geographic, només se n'ha identificat al voltant d'1,2 milions. El poder de la diversitat deriva del fet que cada element en un ecosistema divers contribueix amb característiques úniques a la integritat de l'ecosistema.

La diversitat lingüística

Els éssers humans som animals socials. Per això, vivim en comunitats. Els humans ens comuniquem com a membres de comunitats, construint-les de manera natural al voltant de les qüestions que comuniquem. I com comuniquem aquestes qüestions? Guiats per l'estructura lingüística intrínseca de la ment humana. Per aquest motiu, les llengües humanes evolucionen de forma natural en resposta a la nostra necessitat d'expressar-nos dins dels nostres entorns i l'estructura de les nostres ments. Tots els humans fem front als mateixos problemes primaris de supervivència i subsistència i, per tant, discutim qüestions essencialment semblants utilitzant el mateix conjunt d'habilitats vocals. Doncs, és d'esperar que vàries llengües humanes manifestin un alt nivell de semblança. Aquestes semblances en les qüestions discutides i les habilitats utilitzades per expressar-les són factors importants responsables de les semblances entre llengües. No obstant això, les pertorbacions aleatòries en la manera d'expressar-nos, a més de les peculiaritats dels nostres entorns diversos, tenen tendència a manifestar-se estructuralment a la nostra llengua, portant el que hauria estat una única llengua en vàries direccions degut a la separació en temps i espai entre dos o més grups de parlants d'una única llengua. Skutnabb-Kangas va observar que els països amb megadiversitats han tingut microentorns més variats per observar, analitzar, descriure i discutir que països amb menys diversitat, i tots aquests coneixements s'han codificat a les seves múltiples llengües (Skutnabb-Kangas 2002). Doncs, tant les semblances com les variacions que es manifesten a les llengües humanes són

1.- http://www.alt-i.org/

producte de característiques que defineixen tant la ment humana com l'entorn humà. Segons Lera Boroditsky (2018), la diversitat lingüística revela la ingenuïtat de la ment humana en la seva invenció de 7.000 universos cognitius, cadascun dels quals correspon a una de les llengües parlades arreu del món.

La diversitat lingüística té l'important objectiu de caracteritzar l'estructura de la ment humana, així com els mètodes mitjançant els quals els humans ens adaptem al nostre entorn de formes críticament diverses. A més, la gran varietat de maneres que tenim els humans d'expressar-nos mitjançant la llengua podria donar informació sobre altres aspectes de diversitat de l'entorn humà. Així doncs, la possibilitat d'una manipulació contemplativa i no-destructiva de la llengua com una realitat ideal enlloc de material podria donar facilitats per estudiar aquests altres aspectes de la diversitat natural com a equivalents de la llengua humana.

El valor intrínsec de la diversitat es nota a tots els nivells on es troba diversitat al món natural. La llengua és un component vital del sistema cognitiu humà i cada element de coneixement que ens trobem els humans es codifica en una llengua o una altra. Per tant, la diversitat lingüística és una de les forces de la natura que recolza una comprensió de les diverses variabilitats, i la unitat, del món natural.

La importància de la biodiversitat és relativament fàcil de creure, s'entén bé i la consciència sobre la mateixa està bastant estesa, però la importància de la diversitat lingüística està molt menys valorada. No obstant això, ens podríem adonar de la similar importància de la diversitat lingüística si prenem nota de que els esforços per mantenir i promoure la biodiversitat s'han de complementar amb esforços per mantenir i promoure també la diversitat lingüística, si s'ha de maximitzar la utilitat dels esforços per mantenir i promoure la biodiversitat. La biodiversitat produeix una àmplia varietat de realitat material que recorre a la llengua per classificar-la i descriure-la. Les classificacions i descripcions atorgades a aquests elements de realitat material mitjançant la llengua solen extreure's dels trets característics rellevants d'aquests materials. Aquests trets característics poden ser latents, però les classificacions i descripcions atorgades serveixen d'indicacions vitals de les seves potencialitats. Similar al que representa el programari pel maquinari, segur que la biodiversitat sense la seva corresponent diversitat lingüística negarà a la comunitat de coneixement global les potencials i innombrables realitats materials ubicades a zones específiques dins l'ecologia de la nostra biosfera.

Factors que van en contra de la diversitat lingüística

La recerca de supervivència dels organismes vius sol col·locar-los en conflicte dins el mateix grup i també entre diferents grups. A vegades, aquests conflictes porten a la dominació, i causen l'exterminació de certs grups, ja sigui de forma material o ideal. A les societats humanes, tal dominació pot manifestar-se en genocidi, resultant en l'aniquilació d'un poble dominat. No obstant, amb més freqüència es manifesta com una recessió gradual de la cultura dominada degut a una imposició de la cultura del grup dominant. La recessió d'una cultura sol anar acompanyada de la recessió de la llengua que ha de donar vida a la cultura. Les experiències colonials a diverses parts del món diuen molt de la recessió de llengües degut a la dominació d'una cultura estrangera. Apart de la colonització, són freqüents diversos nivells d'interaccions entre comunitats veïnes, on una cultura es veu dominada per una altra i la llengua que expressa la cultura dominada en cau víctima.

Més enllà d'aquestes causes tradicionals que posen en perill una llengua, que tendeixen a disminuir la diversitat lingüística, el desenvolupament, el desplegament, i l'explotació de tecnologies de la informació i la comunicació (TIC) modernes s'han convertit recentment en una altra força potent que podria anar en contra de la diversitat lingüística. Des del punt de vista de la intuïció bàsica, com la comunicació depèn de la informació, i alhora la informació de la llengua, s'espera que una tecnologia de la informació i la comunicació treballi per promocionar la causa de la comunicació de la informació a través de la llengua i, d'aquesta manera, contribuir al desenvolupament lingüístic. Tanmateix, les TIC modernes en general tenen la capacitat de o bé promocionar o bé malmetre la diversitat lingüística segons com s'utilitzin. L'objectiu d'aquest article és examinar críticament les tendències contemporànies del desenvolupament de les TIC, especialment el desenvolupament i l'ús de les dades massives en el processament del llenguatge natural (PLN), analitzar els seus possibles efectes negatius sobre la diversitat lingüística, i considerar maneres de mitigar-los.

Les TIC i la diversitat lingüística

Els avenços en tecnologia digital han servit per produir i promocionar les TIC com a eines importants que amplien l'abast de les capacitats comunicatives humanes. Tant la informació com la seva comunicació es recolzen inequívocament en la llengua, i la proliferació de dispositius TIC i el seu ús generalitzat en la comunicació humana de ben segur tindrà efectes significatius sobre les llengües i el seu desenvolupament. Per començar, les llengües per les quals s'han dissenyat dispositius TIC tindran un avantatge sobre de llengües a les quals aquests dispositius encara no responen. Per tant, les condicions del desenvolupament de les TIC respecte a les llengües que encara no s'hi han adaptat tindran conseqüències en el desenvolupament lingüístic i la supervivència de les llengües.

Segurament la comunicació intergeneracional és el factor més important de supervivència lingüística. Mentre els pares parlin als seus fills en la llengua de la seva cultura, aquesta llengua té la possibilitat de sobreviure. No obstant, si els pares es veuen obligats a parlar als fills en una llengua estrangera, la supervivència de la seva pròpia llengua estarà en perill. Hi ha moltes circumstàncies que poden obligar als pares a parlar als fills en una llengua estrangera, però la que té importància per aquest article és l'efecte de les TIC modernes.

Ara vivim a un món on els humans ens comuniquem cada cop més els uns amb els altres a través de màquines. A més, cada cop hi ha més necessitat de comunicació entre humans i màquines. La gent jove, al ser nadius digitals, acullen els dispositius TIC amb entusiasme i, pràcticament, viuen la major part de les seves vides a través d'aquests dispositius al ciberespai. Enlloc de viure completament al món real de massa i volum, ara viuen una part considerable de les seves vides al món virtual de les xarxes socials, interactuant més amb els seus companys a aquest món virtual i menys amb els seus pares al món real, l'únic món que aquests coneixen bé. Això deixa entendre que la llengua dels joves pot estar més determinada per la seva interacció amb els seus iguals a les xarxes socials globals que per interaccions amb els seus progenitors a un espai geogràfic local. Aquest és el primer nivell al qual les TIC poden afectar el desenvolupament lingüístic. Per empitjorar-ho, si els dispositius que utilitzen per interactuar amb els seus iguals no estan ben adaptats a la llengua dels seus pares, de qui suposadament han d'heretar la seva llengua, no només tindran un desavantatge amb els

seus companys, també es veuran afectades negativament les seves interaccions amb els seus progenitors. Aquests joves sempre esperaran que els seus pares puguin comunicar-se amb ells a través de dispositius TIC. Tanmateix, els pares, al ser immigrants digitals, continuaran trobant-se dificultats amb aquests dispositius, amb sort amb una mica d'ajuda i ànims dels fills. Si, no obstant això, aquests dispositius TIC no s'adapten bé a la llengua que els pares esperen llegar als seus fills, es compromet la cadena intergeneracional de supervivència lingüística. Els progenitors d'aquesta cultura es veuen obligats a comunicar-se amb els fills en una llengua estrangera, posant la seva llengua indígena en perill i, d'aquesta manera, dificultant que els pares puguin comunicar matisos importants del seu entorn cultural als seus descendents. Per tant, és lògic que, més enllà dels factors tradicionals que posen en perill la supervivència de les llengües, qualsevol llengua per la qual els dispositius TIC moderns no hagin estat dissenyats està en perill d'extinció.

Hi ha alguns dispositius TIC que són agnòstics lingüísticament, i la taxa d'adopció d'aquestes tecnologies és testimoni de la importància de la llengua en l'adopció de les tecnologies de la informació i la comunicació. Un bon exemple d'una tecnologia de la informació i la comunicació lingüísticament agnòstica és el telèfon com a dispositiu de comunicació basat en la veu. L'alta taxa d'adopció del telèfon en tot el món és un bon testimoni de l'agnosticisme lingüístic del telèfon com a dispositiu basat en la veu. No obstant això, amb la progressiva integració de tecnologies de la comunicació amb tecnologies de la informació, l'abast de comunicació d'informació del "servei telefònic bàsic" (POTS, per les seves sigles en anglès[2]) s'ha ampliat per acomodar molts serveis nous força sorprenents fins tal punt que inclús la comunicació intergeneracional a través d'aquests dispositius ara demana utilitzar més que la mera veu. Comunicar amb aquests tipus de dispositius amb més que la mera veu i aprofitar al màxim les facilitats que ofereixen, al nivell més bàsic, requereix el desenvolupament de configuracions locals, teclats, correctors ortogràfics i de gramàtica. A un nivell més avançat, es necessitaran sistemes de reconeixement de veu, sintetitzadors de veu, traductors automàtics i altres serveis per facilitar la comunicació a través de i amb les màquines.

El desenvolupament de dispositius TIC que s'ajustin a les innombrables llengües a les quals poden necessitar adaptar-se, presenta el problema de l'ou o la gallina per la majoria de les 7.000 llengües parlades al món avui dia. Apart de les moltes llengües d'Europa i algunes de les llengües d'Àsia, la majoria de les altres llengües del món serien considerades llengües amb pocs recursos en el millor dels casos. Una llengua amb escassos recursos en aquest context és una llengua que no té els recursos electrònics que cal fer servir per desenvolupar tecnologia del llenguatge humà pertinent a la llengua en qüestió. Alguns dels recursos més bàsics necessaris pel desenvolupament de tecnologia del llenguatge humà per qualsevol llengua inclouen corpus de llengua escrita, corpus de llengua oral, diccionaris monolingües i diccionaris bilingües, tots en suport informàtic. També són pertinents les col·leccions de terminologia, gramàtiques, etiquetadors, analitzadors morfològics i intèrprets de llenguatge d'ordres a més de sistemes de reconeixement de veu, sintetitzadors de text a veu, estàndards d'anotació, eines d'anotació, eines d'exploració i explotació de corpus, i corpus

2.- Plain old telephone service.

bilingües per entrenar traductors automàtics. Aquests i molts més, anomenats col·lectivament *Basic LAnguage Resource Kit* (BLARK, per les seves sigles en anglès)[3], són els recursos electrònics bàsics en suport informàtic necessaris per construir dispositius TIC que puguin adaptar-se a una llengua determinada (Krauwer, 2003). Malauradament, però, crear el BLARK d'una llengua des de zero requereix la disponibilitat d'alguns dels elements mateixos del BLARK, cosa que porta al ja esmentat problema de l'ou o la gallina, una situació desconcertant on es necessita una gallina per pondre l'ou i alhora la gallina ha de sortir de l'ou per pondre'l. La solució a tal problema implica una estratègia que converteix el famós cercle viciós en una espiral virtuosa. Un procés cíclic iteratiu on s'empeny el moviment circular a un pla bidimensional cap a una tercera dimensió, i d'aquesta manera, es converteix el que hauria restat un cercle en una espiral que inclou alguna millora incremental, encara que sigui mínima durant cada cicle de moviment.

La configuració local és fonamental per al BLARK. Com a terme informàtic, la configuració local especifica l'entorn lingüístic d'un programari a través del qual es defineixen els paràmetres d'interfície de l'usuari com el format de dates i temps, el format de visualització de números, símbols de divises i altres paràmetres específics de la regió del món on s'utilitza una llengua. Per desenvolupar la configuració local pertinent a la cultura d'una llengua determinada, pot ser suficient examinar minuciosament la cultura en qüestió i comparar les seves necessitats de configuració local amb els elements d'algunes configuracions locals ja existents. Amb una mica d'intuïció, és possible desenvolupar d'aquesta manera una configuració local creïble per una cultura i llengua. Per desenvolupar una distribució de teclat eficient per a una llengua, és possible que calgui considerar la distribució estadística dels caràcters definits per l'ortografia de la llengua i, d'aquesta manera, col·locar els caràcters utilitzats amb més freqüència en posicions més fàcilment accessibles pels dits més adequats. Per diverses raons històriques, és possible que això no s'hagi fet amb el disseny de la distribució del teclat QWERTY, però s'hauria d'aprofitar qualsevol oportunitat per beneficiar-se de l'ergonomia. Per obtenir aquesta distribució estadística, necessitem un corpus considerable de text en suport informàtic, i per construir-ne un d'aquestes característiques necessitem una distribució de teclat eficient, i aquí rau el problema de l'ou o la gallina de la situació. L'estratègia per convertir aquest cercle aparentment viciós en l'espiral virtuosa descrita anteriorment és començar amb un corpus provisional modest desenvolupat sobre una distribució de teclat ja existent però ineficient, i utilitzar aquest corpus per desenvolupar un nou disseny que millori repetidament l'eficiència de la distribució existent.

Aquesta estratègia pot funcionar bé per desenvolupar el corpus modest necessari pel desenvolupament d'eines TIC bàsiques com teclats i correctors ortogràfics. Tanmateix, per aprofitar al màxim algunes de les tècniques més modernes ofertes per l'aprenentatge profund[4] en el desenvolupament de sistemes de reconeixement de veu, sintetitzadors de veu, traductors automàtics i altres tecnologies de PLN avançades, cal recórrer a l'ús de dades massives. Aquestes tecnologies necessiten que els ordinadors "aprenguin" mitjançant la immersió en grans quantitats de dades històriques.

3.- Joc bàsic de recursos de llenguatge.
4.- En anglès, deep machine learning.

Les dades massives i la diversitat lingüística

Les dades massives descriuen grans quantitats creixents de dades, acumulades com a resultat de l'emmagatzematge de dades generades a través de processos quotidians normals de transaccions formals i informals. Poden considerar-se un subproducte d'aquests processos perquè la producció de dades no és l'objectiu principal d'aquestes transaccions. Es caracteritzen per les tres "V": volum, velocitat i varietat. Tal com indica el terme, el volum és l'essència de les dades massives. Per necessitat, implica grans quantitats de dades que poden dificultar un emmagatzematge i processament eficients amb eines de gestió de dades tradicionals. Per tant, requereix nous mètodes d'emmagatzematge i processament eficients. La característica de velocitat de les dades massives aborda el fet que, encara que típicament inclouen grans quantitats, encara segueixen creixent a un ritme força ràpid, per tant, es fan encara més voluminoses. A més, tot i que alguns continguts de les dades massives poden estar estructurades, no sempre ho estan necessàriament, perquè les dades massives inclouen una àmplia varietat d'estructures, a més de formats imprevisibles. Un exemple típic de dades massives és l'acumulació de les dades generades a les borses de tot el món. Per suposat, l'objectiu principal d'una borsa no és la generació de dades, però les dades es generen igualment, en grans quantitats, a un ritme accelerat i amb una àmplia varietat d'estructures, o una manca d'estructura, en tot tipus de formats imprevisibles. Per exemple, diuen que la borsa de Nova York produeix prop d'un terabyte de dades en un sol dia d'operacions (Groenfeldt, 2013).

El terme dades massives no es refereix només a la àmplia varietat de dades acumulades a alta velocitat en grans quantitats. També descriu les noves infraestructures dins les quals s'emmagatzemen les dades, a més de les eines i tècniques utilitzades per processar i analitzar aquestes dades voluminoses. El seu valor deriva del fet que el coneixement es troba amagat en patrons, i els patrons només esdevenen obvis amb quantitats significatives de dades. Per tant, en virtut del seu volum, velocitat i varietat, les dades massives tenen la capacitat de revelar patrons fins ara no identificats que guarden coneixements fins ara no disponibles i, en molts casos, coneixements als quals els ordinadors tampoc tindrien accés de cap altra manera. Aquest és el motiu pel qual les dades massives juguen i segueixen jugant un rol molt important en el desenvolupament de la intel·ligència artificial.

Recentment, les dades massives han esdevingut una font molt significativa de dades lingüístiques. Donat que moltes activitats humanes ara tenen lloc, o almenys passen per l'Internet, les dades acumulades com a subproducte d'aquestes activitats constitueixen dades útils que creixen en volum fins a convertir-se en dades massives, principalment degut a la velocitat i varietat a les quals es generen. Els butlletins d'informació, els diaris, i les revistes ara es troben normalment a Internet. La documentació d'activitats de governança, a més de processos judicials i debats legislatius, informes de reunions, publicacions acadèmiques i opinions públiques i privades ara arriben a Internet amb certa facilitat, convertint-lo en un repositori vital i convenient. A nivells menys formals, tant els grups de xat professionals com socials, i moltes altres activitats socials, han elevat l'Internet a un camp on molta gent, sobretot jove, ara viu una part significativa de les seves vides. Per tant, tots els texts generats a partir d'aquestes activitats esdevenen fonts naturals de dades lingüístiques, proporcionant una profunditat de percepció sense precedents de la ment humana com a font d'aquestes dades i la llengua humana com a vehicle del pensament humà.

Malauradament, la majoria de les llengües parlades avui dia al món no s'utilitzen en àrees que acumulin les dades lingüístiques generades per l'ús convertint-les en dades massives. Les dades lingüístiques es generen sense voler a través de diversos processos quotidians que impliquen l'ús de les llengües. Per algunes llengües, la majoria de texts generats a través d'aquestes senzilles interaccions quotidianes deixen petjades documentals a Internet i, d'aquesta manera, ofereixen l'oportunitat d'acumular grans quantitats de dades segons les condicions de volum, velocitat i varietat que promouen la seva acumulació i conversió en dades massives. No obstant això, la majoria de llengües no s'utilitzen en àrees que promoguin aquesta acumulació, la qual cosa deixa escasses o nul·les oportunitats per a què aquestes llengües produeixin texts que es puguin acumular i convertir en dades massives. Segur que les llengües que entren dins d'aquesta categoria es veuran desafavorides segons els imperatius de l'era de la informació.

Abans de l'era de la informació, era suficient utilitzar una llengua en transaccions quotidianes per considerar-la una llengua viva. Tanmateix, a l'era de la informació on ens trobem ara, donada la necessitat dels humans de comunicar-nos amb les màquines i a través d'elles i el valor que aquestes han aportat a les dades massives, per a què una llengua es consideri una llengua viva activa, cal que l'ús d'aquesta llengua produeixi petjades documentals que puguin acumular-se i convertir-se en dades massives utilitzades per avançar i donar vida a l'ús de la llengua. Això esdevé un gran repte per la gran majoria de les prop de 7.000 llengües parlades avui dia al món. Concretament, el repte rau en com impulsar l'ús d'aquestes llengües en àrees on el seu ús no equivalgui a una "combustió de gasos" lingüística. La combustió de gasos és un procés de la indústria petroquímica on es cremen deliberadament gasos combustibles útils com el metà, el butà i el propà, produïts com a subproductes, com a estratègia d'eliminació. Dins el context d'ús lingüístic, comparem les dades lingüístiques desaprofitades i generades durant interaccions humanes quotidianes amb la crema de gas valuós. Aquests elements de dades lingüístiques que permetem que s'esvaeixin, quan es genera un text sense documentar en interaccions humanes quotidianes, constitueixen recursos lingüístics valuosos que esdevenen, però, productes residuals perquè no han estat aprofitats. Considerant la importància de les dades massives a l'era de la informació, s'han d'instaurar estratègies que garanteixin la captura i acumulació d'aquests recursos valuosos, convertint-los en dades lingüístiques útils.

Un problema local amb conseqüències globals

Cada organisme dins la biosfera juga un rol important en l'ecologia de la mateixa, i sabem que aquests rols tenen una transcendència global. Podem extrapolar aquest fet al món lingüístic. En aquest sentit, és raonable assumir que una llengua pot codificar elements importants de coneixement que potser no estan disponibles en altres llengües. Això pot deure's al fet que es poden observar certs fenòmens naturals amb més facilitat a la localitat on es parla aquesta llengua. Basant-nos en la gran diversitat de llengües, totes elles intentant complir els mateixos objectius de comunicació i identitat, podem atrevir-nos a conjecturar que cada llengua té la capacitat de fer servir una estratègia de codificació d'informació única i altament eficient. Aquestes estratègies de codificació d'informació úniques i extremadament eficients poden oferir maneres de codificar informació general o especialitzada de formes molt productives però fins ara desconegudes. El llarg període de temps

durant el qual han evolucionat les llengües humanes i l'èxit de la informàtica moderna en implementar algoritmes genètics insinuen la viabilitat d'aquesta conjectura. Per exemple, la gran varietat de morfologia de diverses llengües suggereix que els humans hem estat molt creatius en l'evolució de les nostres llengües. Hem aconseguit trobar maneres d'utilitzar un conjunt relativament limitat de símbols lingüístics per representar el nombre infinit de pensaments que passen pels nostres caps. La gran varietat de disposicions creatives d'aquest relativament petit conjunt de símbols lingüístics que converteixen combinacions de fons en fonemes, fonemes en morfemes, morfemes en paraules, paraules en frases i frases en discursos per aconseguir la codificació d'un conjunt infinit de pensaments continua essent una gesta, encara que pugui ser comuna. Als experts globals els deuen quedar moltes lliçons per aprendre d'aquesta proesa natural de la humanitat.

Les correspondències analògiques entre diversos fenòmens naturals molt diferents presenten oportunitats per desenvolupar estratègies creatives per codificar informació. Es poden utilitzar fenòmens evidents i, per tant, fàcilment comprensibles, per comprendre fenòmens més complexos però analògics. Els últims avenços en bioinformàtica han demostrat el valor d'una perspectiva de processament d'informació a la biologia molecular. L'ús de la intel·ligència artificial per tractar el problema de plegament de proteïnes, especialment la feina d'Alphafold de DeepMind, insinua una forta base lingüística de la genòmica. La llengua natural podria proporcionar equivalents fàcilment comprensibles d'alguns dels patrons, estructures i processos genòmics complexos. El PNL pot oferir algoritmes comprovats per abordar els problemes més complexos que sorgeixen a la seqüenciació genòmica. Si les llengües que demostren de forma més evident l'aplicació d'alguns d'aquests algoritmes s'extingeixen abans de descobrir-se aquests algoritmes, la humanitat perdria un element de coneixement molt important. Aquests tipus d'explicacions analògiques no es poden limitar només a la bioinformàtica.

A la ciència cognitiva, per exemple, la llengua juga un rol fonamental en la teorització de l'estructura de la ment humana, perquè la llengua és una producció intangible i abundant de la ment. Per aquest motiu, per tant, l'estudi de la fonologia, morfologia, sintaxi i pragmàtica de la producció lingüística de la ment humana porta a conèixer l'estructura i el funcionament de la mateixa. Malauradament, tot el que avui sabem de la ciència cognitiva podria venir de prop de 30 o de ben segur menys de 50 llengües força similars de les 7.000 llengües del món. Des d'un punt de vista estadístic, això constitueix un mostreig de conveniència, el qual, segons la teoria estadística, no es pot generalitzar. Per tant, existeix una necessitat d'ampliar radicalment la base de llengües disponibles als estudis de ciència cognitiva i altres camps relacionats. Podríem estar investigant la ment humana a través d'unes poques finestres fosques i petites, mentre existeixen altres finestres segurament més lluminoses i àmplies, però que encara romanen inexplorades.

Per això, té molt més valor, més enllà dels sentiments merament romàntics però vàlids, la diversitat lingüística i el valor econòmic que ofereix a la subsistència local. El valor de les lliçons que ens podrien donar les estratègies de codificació d'informació d'algunes llengües naturals obscures segueix essent força desconegut. De ben segur que la comunitat global surt perdent amb la desaparició de qualsevol llengua individual, encara més amb la pèrdua de la diversitat lingüística. L'estructura de les dades o les solucions algorítmiques d'alguns problemes encara pendents podrien haver-se desenvolupat genèticament i utilitzat en alguna llengua humana desconeguda.

La responsabilitat principal del desenvolupament d'una llengua rau en els parlants de la llengua. Primer, és la responsabilitat dels parlants d'una llengua identificar àrees pertinents i útils per les seves vides i subsistència, i utilitzar la seva llengua en aquestes esferes. No obstant això, cada llengua natural és un regal de la natura, no només pels seus parlants, però també per tota la humanitat. Per tant, cada llengua que la natura ha decidit evolucionar amb la col·laboració dels humans dins una cultura guarda un tresor essencial per tota la humanitat. La gran diversitat de la fonologia, morfologia, sintaxi i semàntica de cadascuna de les prop de 7.000 llengües del món és mostra de l'energia del principi de realització múltiple a la filosofia de la ment. Per tant, cada llengua pot ésser considerada un camí viable entre un petit conjunt de símbols finits i un conjunt infinit de texts útils culturalment en un espai de cerca enorme. S'ha cercat heurísticament dins d'aquest espai excessivament gran maneres d'organitzar aquest petit conjunt finit de símbols en un sistema capaç de codificar infinites expressions d'informació des de l'aparició de l'homo sapiens. Després de trobar alguns d'aquests camins viables, tenim el deure d'assegurar-nos que els explotem al màxim. Això inclou utilitzar-los en altres dominis apart de la comunicació humana bàsica com a guies analògiques per abordar fenòmens naturals més complexos.

Acumular i explotar les dades massives per a la diversitat lingüística

Com ja hem comentat, per a la majoria de llengües del món, l'ús quotidià no produeix petjades documentals que es puguin acumular i convertir en dades massives. Després d'argumentar que les dades massives poden contribuir molt a la diversitat lingüística en l'era de la informació, cal explorar la possibilitat de generar el màxim de dades possibles a partir de l'ús quotidià de la majoria de les llengües del món.

Existeix una gran varietat de característiques per diverses llengües. Algunes, com les llengües oficials de nacions lingüísticament plurals, sí ofereixen materials impresos generats freqüentment i disponibles en moltes llengües, cosa que ofereix la possibilitat de crear tant corpus lingüístics escrits com paral·lels. Per llengües on s'ofereixen comunicacions governamentals i altra informació pública, és important recollir tots aquests materials impresos, anotar-los apropiadament i acumular-los expressament en algun tipus de repositori d'accés públic. Algunes llengües s'utilitzen per produir diaris, revistes periòdiques i d'altres butlletins. Cal recollir-los i acumular-los expressament en suport informàtic. Algunes llengües s'utilitzen en processos judicials i debats legislatius. També cal arxivar-los deliberadament. En alguns casos, es tradueixen i registren en algunes llengües estrangeres amb llegat colonial elements de processos judicials en llengües indígenes. Ens hem d'esforçar a re-traduir-los i documentar-los en les llengües de partida. Cal prestar especial atenció als usos quotidians de qualsevol llengua que tingui la capacitat de generar documentació rellevant. S'ha d'accedir a i conservar tota la documentació en formats apropiats.

Per moltes llengües, l'oportunitat més probable de documentació amb poc o zero esforç addicional és l'emissió televisiva i la radiodifusió. Aquest tipus de comunicació dona oportunitats per recollir corpus lingüístics escrits, orals i bilingües. Malauradament, molts pocs mitjans de comunicació comprenen el valor dels seus materials de comunicació més enllà dels seus objectius d'informació, educació i entreteniment. Poden no adonar-se del valor dels materials acumulats als seus arxius com a dades lingüístiques. Pitjor encara, alguns mitjans no consideren

alguns materials importants i valuosos prou dignes de guardar. Per exemple, a gran part d'Àfrica, els butlletins de notícies es preparen en anglès o alguna altra llengua colonial. Llavors, s'emeten versions dels butlletins en algunes de les llengües indígenes de la localitat pertinent, traduïdes a partir dels butlletins en la llengua colonial estrangera. En algunes situacions, les notícies en la llengua indígena es tradueixen extemporàniament a partir del text en la llengua colonial en directe. En altres situacions, encara que la traducció estigui redactada d'abans, el guió està manuscrit i, per tant, no es troba en suport informàtic. En algunes d'aquestes situacions, es llença el paper on s'ha redactat la traducció tan bon punt acaben les notícies, impedint qualsevol possibilitat de mecanografiar el guió i convertir-lo en suport informàtic més tard. D'aquesta manera, materials valuosos per corpus de texts escrits i corpus paral·lels que podrien acumular-se i convertir-se en dades massives es perden a diari en aquest desafortunat procés de "combustió lingüística".

Les tendències globals actuals a la comunicació comunitària ofereixen una oportunitat única d'accés sense precedents a un nombre significatiu de les llengües del món utilitzades a la comunicació a nivell comunitari. Això ofereix oportunitats incalculables per a la documentació d'aquestes llengües, tant a corpus lingüístics escrits com orals, i la seva acumulació i conversió en dades massives. Encara que els materials de comunicació sorgeixin essencialment de forma extemporània, les gravacions d'àudio de l'emissió podrien conservar-se adequadament com a base de documentació. Per evitar els actes de "combustió lingüística" descrits anteriorment i així assegurar que els materials de comunicació estan disponibles a llarg termini, s'ha de dur a terme preparatius per introduir pràctiques favorables que promoguin l'emmagatzematge i l'acumulació de materials de comunicació i garanteixin el seu fàcil accés per a investigadors.

S'ha de formalitzar estrictament i implementar la recollida de materials de comunicació com a base de dades massives pel seu anàlisi lingüístic dins d'estructures responsables. Una estructura global existent i pertinent que es podria fer servir per facilitar l'accés a aquests materials de comunicació és l'Associació Mundial de Ràdios Comunitàries (AMARC, per les seves sigles en francès[5]). La importància d'intentar acumular dades massives pel seu anàlisi lingüístic a través d'aquesta organització requereix una planificació col·laborativa. Les dades potencials són incalculables i no estaria fora de lloc incentivar l'acumulació mesurable de dades.

L'acumulació de materials lingüístics escrits i orals i la seva conversió en dades massives és de gran importància com a recurs natural, i podrien encarregar-se'n les polítiques lingüístiques nacionals de països on la "combustió lingüística" té lloc a diari. És de suma importància una bona documentació i agrupació de materials emesos i impresos en suport informàtic en un repositori de fàcil accés ubicat dins d'una institució pública.

Conclusions

Ara vivim a la societat de la informació, a l'era de la informació, fruit de la revolució de la informació. A la societat de la informació on ens trobem actualment, les dades són un recurs clau que alimenta totes les activitats. Les dades es generen a interaccions senzilles a partir de transaccions quotidia-

5.- Association Mondiale Des Radiodiffuseurs Communautaires.

nes, però a moltes parts del món, les dades no es recullen ni s'acumulen per convertir-se en dades massives; la manera en què es converteixen en patrimoni i el seu valor esdevé accessible. La llengua és el substrat fonamental del qual depèn tota la informació i, per tant, la comunicació. Així doncs, a la societat de la informació, la llengua és de vital importància. La diversitat lingüística és un regal de la natura a la humanitat i no s'hauria de comprometre. El malbaratament de dades lingüístiques generades però no recollides en interaccions senzilles a partir de transaccions quotidianes posa en perill la diversitat lingüística. Algunes de les promeses de l'era de la informació podrien dependre de la diversitat lingüística de formes encara desconegudes. No ens podem permetre sacrificar els beneficis que podrien acumular-se a partir d'aquestes promeses a l'altar de la ignorància predominant.

Referències

Boroditsky, L. (2018) How language shapes the way we think. https://www.youtube.com/watch?v=RKK7wGAYP6k consultat el setembre de 2021

Krauwer, S., (2003) The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap; http://www.elsnet.org/dox/krauwer-specom2003.pdf, consultat el setembre de 2021.

Miller, G. & Spoolman, S. (2012). Environmental Science – Biodiversity Is a Crucial Part of the Earth's Natural Capital. Cengage Learning. https://books.google.com.ng/books?id=NYEJAAAAQBAJ&pg=PA62&redir_esc=y#v=onepage&q&f=false, consultat el setembre de 2021

National Geographic, Biodiversity. https://www.nationalgeographic.org/encyclopedia/biodiversity/, consultat el setembre de 2021.

Skutnabb-Kangas, (2002) why should linguistic diversity be maintained and supported in Europe? Some arguments. Language Policy Division, Directorate of School, Out-of-School and Higher Education DGIV, Council of Europe, Strasbourg

THE INDIGENOUS LANGUAGES TECHNOLOGY (ILT) PROJECT AT THE NATIONAL RESEARCH COUNCIL OF CANADA, AND ITS CONTEXT

Roland Kuhn

INTRODUCTION

Recent decades have seen the creation of hundreds of Indigenous language revitalization and reclamation projects across Canada. Most are located inside Indigenous communities; a few are centred in universities. They range in scale from the ambitious, multi-language efforts of the First Peoples' Cultural Council (FPCC), which is located in and funded by the Province of British Columbia and is active there and elsewhere in Canada (https://fpcc.ca/), to unfunded one- or two-person volunteer efforts in remote communities.

The project that I have the honour to lead described in this paper – and on this periodically updated website: https://nrc.canada.ca/en/research-development/research-collaboration/programs/canadian-indigenous-languages-technology-project – is managed inside the National Research Council of Canada (NRC), which is the primary research and technology organization of the Government of Canada. However, the reader should not conclude that "settler" (non-Indigenous) governments play a key role in Indigenous language revitalization in Canada. All successful language revitalization projects that I am aware of are Indigenous-run or collaborate closely with Indigenous language activists.

Our Indigenous Languages Technology (ILT) project at NRC (henceforth, "NRC-ILT") is **not** itself a language revitalization project: it builds tools that people working on language revitalization sometimes find useful. We – the NRC-ILT team – are like lighting technicians or stagehands in a theatre: we are not the actors (we do not appear on stage), nor are we in charge of the production. Those roles are filled by Indigenous language activists and their communities. Language revitalization would go on without us. At times, however, our technical help may make it go slightly better. Everything we have accomplished has been done in collaboration with Indigenous stakeholders.

Disclaimer: this paper is one person's partial, subjective perspective on the technologies being applied to language revitalization in Canada, along with a description of the NRC-ILT project. It is by no means an exhaustive survey of the field; it leaves out many, probably most, organizations active in language revitalization, and their accomplishments. This article is an extended version of (Kuhn et al., 2020), which means it is deeply indebted to the contributions of my co-authors on that paper.

INDIGENOUS LANGUAGES IN CANADA: HISTORY AND DEMOGRAPHICS

Figure 1 shows the range of Indigenous language families in what was later to become Canada, at the time of first contact with Europeans. Each family was and is different from the others, in terms of phonetics, vocabulary, and syntax – in some cases, as different as the Germanic, Turkic, and Sino-Tibetan families in Eurasia are from each other. Languages that were geographical neighbours, even if from different families, sometimes borrowed words and aspects of phonetics or syntax from each other.

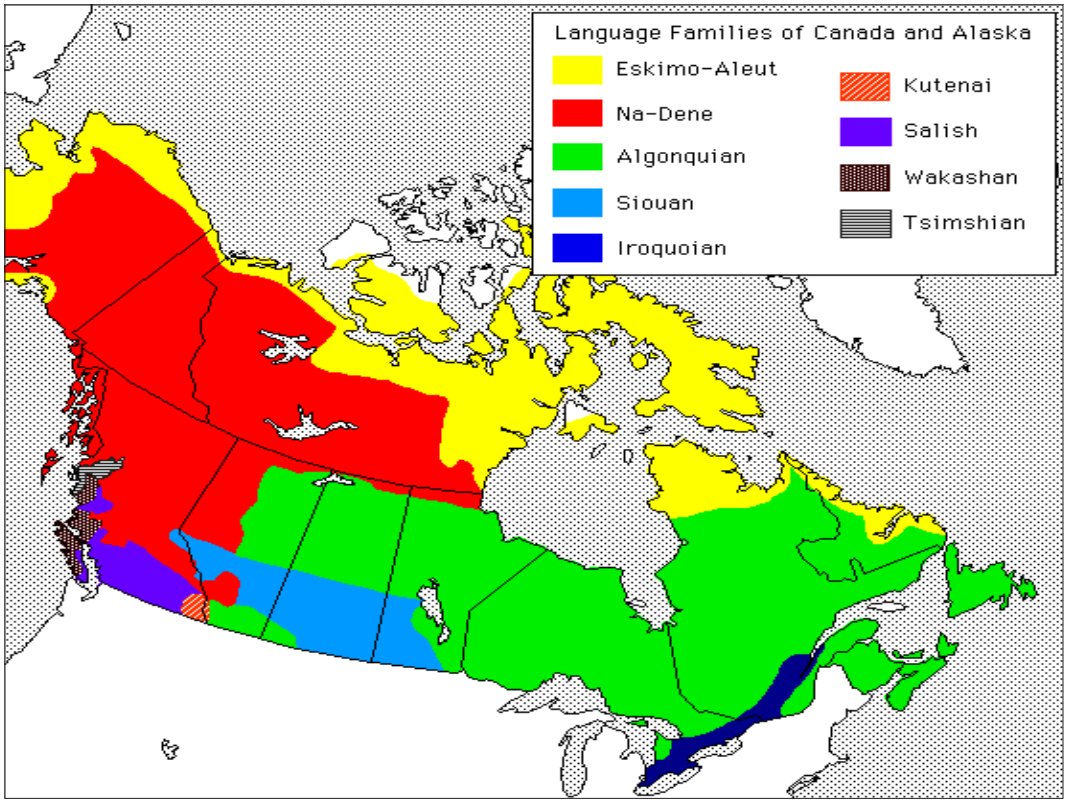


Figure 1. Indigenous languages in Canada during first contact with Europeans.
(Map from Matthew Dryer, Buffalo University)

The terminology for Indigenous communities in Canada often employs three labels: Inuit, Métis, and First Nations. The Inuit are a cohesive ethnic group, as are the Métis. However, "First Nations" groups together all other Indigenous peoples in Canada, a diverse set of ethnicities.

Most Indigenous languages spoken in Canada are polysynthetic: a single word often expresses a complex meaning that would require a full sentence in non-polysynthetic languages. A word is also typically made up of more morphemes than in other languages. An average English word has about 3 morphemes; an average Mohawk word has in the range 5-6. Mohawk (Iroquoian language family), Cree (Algonquian family), and Inuktitut (Eskimo-Aleut family) are all polysynthetic, though they are from otherwise dissimilar families.

After contact between Indigenous peoples and Europeans, creoles with aspects of both Indigenous and European languages arose; most have since disappeared. For instance, contacts between Dutch traders and Mohawks in the Hudson valley gave rise to a Dutch-Mohawk creole. The Chinook Jargon creole in its final form (it survived into the 20th century) had elements from West Coast Indigenous languages, Russian, Hawaiian, and Chinese.

The Métis emerged as a distinct nation of mixed Indigenous and European ancestry in the early 19th century (Bakker, 1997; Rosen & Souter, 2009). Bungi, a creole spoken by the Scottish Red River Métis in present-day Manitoba, had elements from Orkney Scottish English, Scottish Gaelic, French, Cree and Anishinaabemowin (Ojibwe). However, most of the Métis people spoke Michif,

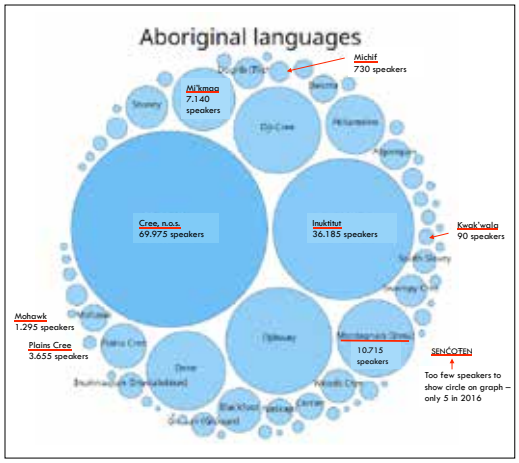


Figure 2. Number of speakers by language (2016 census).
https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/dv-vd/lang/index-eng.cfm

a contact language. Michif survives and is the focus of strong language revitalization efforts. Roughly, it may be viewed as combining mostly French nouns with verbs from Plains Cree (from the Algonquian family); it is mostly polysynthetic (Davies, Santos, and Souter, 2021; Sammons, 2019).

After Canadian Confederation in 1867, Indigenous languages were targeted by government policies that sought to eradicate them and traditional Indigenous customs, with the goal of assimilating Indigenous people into "white" society. The Indian Act discouraged, and often made illegal, gathering for cultural practices and speaking ancestral languages. Many Indigenous children were forcibly removed from their communities and placed in compulsory boarding schools ("residential schools") by the federal government. These schools were run by churches; some by the Catholic Church, others by Protestant denominations. Multiple forms of psychological, physical and sexual abuse that occurred in these schools are well-documented. According to the Truth and Reconciliation Commission of Canada, the residential school system was "created for the purpose of separating Aboriginal children from their families, in order to minimize and weaken family ties and cultural linkages" (Government of Canada (2015), preface). Furthermore, during a period in the mid-20th century, many Indigenous children were taken from their birth families so that they could be adopted by non-Indigenous families, in the notorious Sixties Scoop (Fachinger, 2019).

Despite this history of systematic discrimination, Indigenous communities have resisted assimilation and persisted in preserving and teaching their languages. The benefits of learning an ancestral language have been correlated with improved well-being in youth (Chandler and Lalonde, 1998; Hallett et al., 2007).

On the basis of personal observation in the course of the NRC-ILT project, I believe that many Indigenous languages spoken in Canada are experiencing a renaissance. Young people in Indigenous communities – who often have grown up speaking only English or (in Quebec) French – are enrolling in record numbers in courses that will enable them to speak with each other in their ancestral languages. I've met some extraordinary Indigenous individuals – almost all poorly paid, and some unpaid volunteers – who teach these languages. These language activists form an elite: a self-selected group of people who combine idealism with high intelligence and a capacity for hard work. Teaching a polysynthetic language to people who originally spoke only English or French is not for the faint of heart.

Figure 2 shows the number of speakers of each Indigenous language in 2016, according to the Canadian census (Statistics Canada, 2017). Languages underlined in red are those the ILT-NRC project has interacted with in some way. The numbers shown should be taken with a grain of salt; experts disagree with some of them. However, this figure gives

a rough sense of the demographic weights of these languages. Plains Cree (Algonquian family) has the most speakers; Inuktitut (Eskimo-Aleut family) ranks second. Inuktitut is, uniquely, an important government language (in Nunavut Territory). The existence of a government bureaucracy that partly functions in Inuktitut means that far more text and speech resources are available for this language than for any other Indigenous language in Canada. The disparity with Cree, which has far more speakers but far fewer linguistic resources, because no large bureaucracy uses it, is striking.

The reader should not conclude that most of the "mid-sized" or "small" languages shown have dismal long-term prospects. The figure cannot show the relative vigour of revitalization efforts for each language. For instance, the smallest language shown, SENĆOŦEN, which only had 5 fluent, elderly speakers in 2016, has a vigorous, well-managed language revitalization program that is teaching many young people in the community the language – the next census will count far more SENĆOŦEN speakers (Montler, 2018; First Voices Portal, 2021).

Similarly, the Mohawk language (Kanyen'kéha), with a modest 1,295 speakers in **Figure 2**, is experiencing strong growth, thanks to good language schools in several Mohawk communities. The Kanyen'kéha school for adults at the Six Nations of the Grand River community in southwestern Ontario (Onkwawenna Kentyohkwa; see https://onkwawenna.info) is admired by Indigenous educators across Canada for its success in graduating many fluent speakers after these students have undergone two years of intensive immersion (the language had almost died out at Six Nations, though not in all Mohawk communities). Some graduates are marrying each other, speak Kanyen'kéha at home, and are bringing up children for whom it is their first language. Energy, teaching expertise, and community cohesion have a large positive impact on language revitalization – they may be better predictors of a language's future than the current number of fluent speakers.

TECHNOLOGIES FOR LANGUAGE REVITALIZATION

"Indigenous people must take the lead in developing the next wave of responsive and responsible language technologies … Fortunately, Indigenous people around the world are exercising technological leadership as they claim spaces for their languages in the digital realm" (Brinklow, 2021).

Indigenous languages spoken in Canada are **heterogeneous**, in their linguistic properties (10 unrelated language families) and in the number of speakers per language. Indigenous language activists in Canada are keenly aware of the remarkable achievements of their peers in Aotearoa – i.e., New Zealand, where Māori has been an official language since 1987 (King, 2013) – and in Hawai'i (Asensio, 2019). However, language revitalization in Canada is an even more complex challenge. Each language here has different needs, ranging from those of severely endangered languages – where the priority tends to be recording a small number of Elders who still speak the language fluently – to those of Inuktitut, a language of government in Nunavut that is widely spoken, where the priority is to encourage its wider use (e.g., in medical settings) and to improve education in the language.

Two recent articles (Brinklow et al. 2019, Brinklow 2021) discuss language technologies for revitalizing and reclaiming Indigenous languages. Quoting the linguist Francis Tyers, Brinklow points out that because big software compa-

nies are profit-hungry, commercial technologies focus on the needs of "small numbers of rich people, or large numbers of poor people." Indigenous languages in Canada are spoken by small numbers (by world standards) of relatively poor people – they don't matter much to big companies. Insofar as companies are interested in Indigenous languages, they often want to retain IP in linguistic data, thus threatening Indigenous data sovereignty – Keegan (2019) gives an example from Aotearoa that illustrates this danger.

There are also technical issues with commercial language technologies:

- None of the "big" languages (English, Chinese, Japanese, Arabic, etc.) for which these technologies have been developed are polysynthetic. Standard technologies applied to polysynthetic languages run into difficulties – for instance, the "out of vocabulary" problem. Even with a vocabulary for automatic speech recognition (ASR) of 1.3 million Inuktitut words, Gupta and Boulianne (2020a) found that more than 60% of the words in held-out Inuktitut stories were **not** in that vocabulary (see below). In a comparable experiment carried out with English texts, one would expect this "out of vocabulary" rate to be much, much lower than 60% – almost certainly less than 5%, possibly lower than 1%.
- Indigenous languages in Canada lack the massive quantities of training data on which machine learning relies. NRC-ILT team members have become familiar with questions from technically unsophisticated, non-Indigenous people like: "Why don't you create software for translating between Cree (or another Indigenous language) and English?"

Our two-fold response:

1. Apart from the Inuit, no Indigenous community we've spoken has shown much interest in machine translation (MT) between their ancestral language and English (or French, in Quebec). Communities are typically more interested in tools to encourage learning and use of their ancestral language.
2. The majority of Indigenous languages in Canada have little parallel data – often, it's limited to a few books of the Bible translated by missionaries. The existence of dialects in most of these languages exacerbates this data sparsity problem. (For Cree, our team and our Indigenous collaborators **might** be able to scrape together at most 100,000 Plains Cree sentences in parallel with English – not nearly enough to train an MT system).

We encounter misunderstanding of this situation even among natural language processing experts. Some experts have hinted that we are technically backward because we use rule-based approaches in most of our work. After all, other teams have successfully applied machine learning to low-resource languages … The term "low-resource language" is a barrier to understanding. It is applied to languages like Tamil, with 75 million speakers, most of them literate in the language, and a history of written texts that goes back thousands of years. It is ridiculous to use the same term to describe the "biggest" Indigenous language in Canada, Plains Cree, with 75,000 speakers and few written texts. Most NRC-ILT subprojects deal with "**extremely** low-resource" languages.

Inuktitut is an exception to this "extremely low-resource" situation (see below).

Littell, Kazantseva et al. (2018) outline technologies for language revitalization in Canada. In the list below taken from this paper, technologies in **bold** were created by NRC-ILT with Indigenous collaborators; underlined technologies were created outside NRC with help from NRC-ILT funding:

- **Implementing digital fonts and layouts for Indigenous languages with uncommon character sets;**
- **Predictive text;**
- **Conversion between different orthographies for the same language;**
- Spell-checking *(no NRC-ILT involvement);*
- **Paradigm generation for teaching complex morphology;**
- **Approximate search;**
- **Machine translation;**
- Speech technologies:
  - Automatic speech recognition;
  - Keyword search through speech data;
  - **Speech-text alignment;**
  - **Text to speech** *(preliminary NRC-ILT work)*;
- Image technologies;
- Computer-aided language learning (CALL).

Littell, Kazantseva, et al. (2018) allude briefly to four other important technological activities:

- Building tools for recording and annotating the speech of fluent Elders
- Digitizing old magnetic and audio tapes of Elders' speech
- Curating speech archives
- Making digital dictionaries.

HISTORY AND COLLABORATIVE APPROACH OF THE PROJECT

An academic overview of NRC-ILT is (Kuhn, Davis, Désilets, Joanis et al., 2020); a detailed technical report is at https://nrc-publications.canada.ca/eng/view/object/?id=d4f10144-c711-43c5-b80b-5ace7df5e68b. Regular updates are at https://nrc.canada.ca/en/research-development/research-collaboration/programs/canadian-indigenous-languages-technology-project.

The project began with the March 2017 budget of the Government of Canada, which provided $6 million (CAD) to NRC to develop, in collaboration with Indigenous stakeholders, software technologies to support Indigenous languages. Our part of NRC was thought to be qualified for this work because of experience with technologies for major world languages – e.g., we had worked on MT from Chinese and Arabic into English, funded by US DARPA's GALE and BOLT programs. There had also been a small-scale NRC initiative focused on Inuktitut during 2003-2012 (Martin et al., 2003; Farley, 2012).

The NRC-ILT team tries to fulfill its mandate by developing close, respectful relationships with Indigenous communities and trying to break with the long, painful history of extractive research practices (Keegan, 2019; Brinklow et al., 2019). We believe in the "empowerment" philosophy, whereby research is carried out collaboratively, with equal emphasis on the agenda of the researcher and of the community (Czaykowska-Higgins, 2009).

This has meant asking Indigenous language activists which software tools would be useful to **them**, rather than offering technologies based on intriguing research themes. Hence, the rather disjointed collection of technologies described below: we've responded to different needs from different communities. We were guided by an Advisory Committee made up of Indigenous language revitalization experts. Their counsel has been invaluable. NRC-ILT never claimed ownership of Indigenous language data collected with project funding. Where possible, we implemented a "kiln and pottery" model: language-independent open-source software (OSS) (the kiln) creates language-specific software (the pots) which will be owned by the community.

Our initial funding was more generous than anticipated, so we funded work outside as well as inside NRC. ILT-NRC has three "rings":

Owennatékha (Brian Maracle): founder & head, Onkwawenna Kentyohwa immersion school.Suggested **Kawennón:nis,** & collaborated in its creation.

Akwiratékha Martin: language instructor in Kahnawá:ke – collaborating with NRC-ILT team to create Eastern version of **Kawennón:nis.**

Kawennón:nis - verb conjugator for Kanyen'kéha

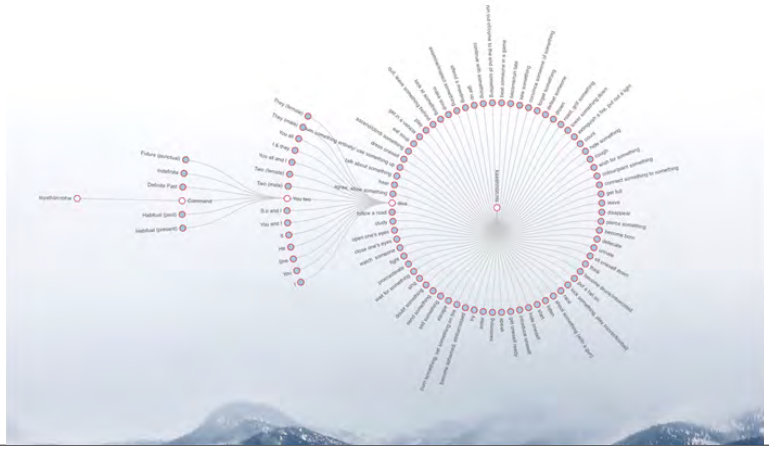


Figure 3. Collaborations on Kanyen'kéha (Mohawk).

- An inner ring of work carried out at NRC in collaboration with Indigenous stakeholders.
- A middle ring of work related to automatic speech recognition (ASR) carried out by the Centre de Recherche Informatique de Montréal (CRIM). CRIM has an illustrious record of ground-breaking research in ASR and related technologies. CRIM delivered experimental ASR results for two important Indigenous languages, and also OSS that speeds up a key stage in annotating recorded speech.
- An outer ring of other subprojects carefully selected by the ILT-NRC team and its Indigenous Advisory Committee for funding. Most of these subprojects were managed by Indigenous organizations; all employed an almost entirely Indigenous work force.

INNER RING: TECHNOLOGIES DEVELOPED INSIDE NRC

VERB CONJUGATORS

The first ILT-NRC subproject arose with a proposal from Owennatékha (Brian Maracle), the director of the Onkwawenna Kentyohkwa (Our Language Society) school in Six Nations of the Grand River in southwestern Ontario: https://onkwawenna.info/. This school is renowned for producing fluent speakers of Kanyen'kéha (Mohawk). Owennatékha suggested that the NRC-ILT team create Kawennón:nis (the WordMaker) – a tool that helps students master the complex verbal system of Kanyen'kéha. There are so many possible verb conjugations for even the most common verb stems in Kanyen'kéha that a physical reference tool for them would be impossibly large; howev-

er, it's feasible to compute these conjugations in software (Kazantseva et al, 2018). Kawennón:nis, which is implemented for the Western dialect of Kanyen'kéha, underwent extensive user testing with students and teachers at the school before being released. It has been received enthusiastically by them. It is designed to match the curriculum, and uses the visual metaphors employed by the school's teachers: https://kawennonnis.ca/wordmaker.

Subsequently, ILT-NRC built general frameworks for building verb conjugators for polysynthetic languages. These are being applied to create verb conjugators for Eastern Kanyen'kéha spoken in Kahnawá:ke (Quebec), and for two languages unrelated to Kanyen'kéha: Algonquin Anishinaabemowin as spoken at Kitigan Zibi (near Maniwaki, Quebec), and Michif.

Figure 3 shows two of our Mohawk collaborators. Both teach Kanyen'kéha: Owennatékha teaches the Western dialect, and Akwiratékha the Eastern dialect. (Other Mohawk educators also helped with building and testing of Kawennón:nis). The figure also shows the Kawennón:nis user interface.

Heather Souter of the Prairies to Woodlands Indigenous Language Revitalization Circle is currently encouraging students in her courses to test a beta version of the Michif verb conjugator. Related work is described in (Davies, Santos, and Souter, 2021).

We plan to roll out verb conjugators for several other languages. Furthermore, Dr. Patrick Littell is leading an effort to develop a user-friendly, spreadsheet-based framework called "Gramble" that may make it far easier for language activists to build conjugators **on their own** (without NRC expertise).

READALONG STUDIO

This is the "surprise hit" of NRC-ILT: we didn't anticipate how many teachers would be interested in having us add a key functionality to pre-existing audio books and videos in their languages. This functionality was pioneered by Prof. Marie-Odile Junker's team at Carleton University.

The functionality is illustrated in **Figure 4**; it is very simple. As an audio book or video with Indigenous speech (here, Atikamekw) is played, the word being spoken is highlighted in accompanying text. In the figure, the word "atisokasotcik" is being spoken, and is thus highlighted. If the listener – a student or teacher – wishes to focus on pronunciation of a word in the text, they click on it, and the pronunciation will be played back to them; they can slow the playback down. We (the Carleton U. and NRC-ILT teams) call audio books or videos with this functionality "ReadAlongs".

Many communities have educational books or videos with speech in their languages, with a transcription into text of the spoken (or sung) words. To turn these books or videos into ReadAlongs, words in the speech must be aligned with words in the text. The Carleton team was aligning ReadAlongs by hand when we began to collaborate with them. Our contribution was to automate this process, using software we developed called "ReadAlong Studio". Delasie Torkornoo of the Carleton team has been helping us improve the ReadAlong Studio code. Our two teams have produced and sent back to Indigenous educators ReadAlongs in the Algonquin, Atikamekw, Southern East Cree, Northern East Cree, Gitksan, Inuktitut, Kwak'wala, Kanyen'kéha (Mohawk), Seneca, and SENĆOŦEN languages. Every couple of weeks, we receive new requests to transform teaching materials in this manner. Indigenous

Figure 4. READALONG STUDIO – A SURPRISE HIT!

educators tell us that the ReadAlongs have high pedagogical value.

KEYBOARDS, ORTHOGRAPHY CONVERSION, & PREDICTIVE TEXT

NRC-ILT has released OSS that implements keyboards for some poorly served writing systems, and that converts between writing systems for some languages. Eddie Antonio Santos of the team worked with Saulteaux, Makah, and Plains Cree. Along with University of Alberta linguist Arok Wolvengrey, he also persuaded Google to change its syllabic keyboard for Plains Cree on Chromebooks (the previous version was appropriate for East Cree, but not Western Cree).

NRC-ILT has also released code for text prediction for mobile devices. In theory, the software can implement text prediction for any language. However, it is difficult for non-experts to use. So far, we have only implemented text prediction for the SENĆOŦEN language. However, since SENĆOŦEN orthography is difficult, members of that community have told us they are very happy with this functionality – it makes entering text on their devices much easier.

INUKTUT

All the work in this section benefited from collaboration with the Pirurvik Centre, an Inuit-run culture and language centre (https://www.pirurvik.ca/).

The related languages or dialects spoken by Inuit people in Nunavut are called collectively “Inuktut”. Most NRC-ILT work on Inuktut has involved its Inuktitut version, spoken on Baffin Island. This work differs from our work on other languages in two ways: 1. We didn’t build educational tools, but **office tools that assist writing or reading Inuktut text** (or translating between Inuktut and English); 2. We worked on **machine translation (MT)** between Inuktut and English, using machine

learning. These differences are because there exist a bureaucracy and a legislative body that function partly in Inuktut: these organizations require text to be written in Inuktut, and supply a growing amount of bilingual Inuktut-English text that can be used to train MT systems.

A small-scale effort at NRC during 2003-2012 to build software tools for Inuktut yielded valuable results – most importantly, a morphological analyzer. A key contributor was Benoît Farley, who subsequently retired from NRC but who maintains a repository for these tools (Farley, 2012).

Another NRC output during this period was successive releases of parallel English-Inuktut text corpora based on proceedings of the Nunavut Legislative Assembly, the "Nunavut Hansard" (henceforth, "NH"). The first of these, NH 1.0, was released to the research community in 2003 (Martin et al., 2003); it covered 155 days of proceedings of the Nunavut Assembly, and comprised 3,432,212 English tokens and 1,586,423 Inuktitut tokens.

When NRC-ILT restarted work on Inuktut in 2017, many more years of unaligned NH proceedings were available. We **created and released a new aligned corpus**, NH 3.0. It comprises 8,068,977 Inuktitut tokens and 17,330,271 English tokens; it covers the proceedings of 687 debate days from April 1, 1999 to June 8, 2017 (Joanis et al., 2020).

The size of NH 3.0 makes it feasible to train **machine translation** (MT) systems. We helped the organizers of WMT 2020, an annual, international evaluation of MT systems, to set up evaluations for MT between English and Inuktitut in both directions. We also (along with Microsoft) paid for human evaluators fluent in Inuktitut to assess translations of test data into Inuktitut from the competing systems. The extreme polysynthetic nature of Inuktitut, unusual among the world's languages, attracted many research teams to this MT challenge.

We built our own MT systems that competed in this evaluation (Knowles et al., 2020). The overall results are here: http://www.statmt.org/wmt20/translation-task.html. We hope that continuing interest in MT research for this language pair will eventually yield practical tools for Inuktitut.

We continue to improve the **office tools for Inuktitut** that NRC released in 2003-2012, and to build new ones:

- we are improving Benoît Farley's morphological analyzer.
- we are reviving WeBInuk, a web-based NRC tool that searches for Inuktitut translations of English words. It displays parallel English-Inuktitut sentences where the English sentence contains the search word. WeBInuk was heavily used by Inuktut translators in 2007-2017, then allowed to disappear (NRC management had other priorities).
- we are building a lemmatizer that inputs an Inuktitut word and outputs the most frequent word with the same root. This is a key step in supporting monolingual search in Inuktitut, or search in the Inuktitut —> English direction. It is much harder to build a search engine for users who enter search terms in Inuktitut than for English or French, because of polysynthesis. (Whatever Inuktitut word the user enters has probably never existed before in the history of the language, and will probably never exist again). Fuzzy search is necessary. **Figure 5** illustrates progress we've made on this and related tools.

PRELIMINARY TEXT TO SPEECH (TTS)

The curriculum at Onkwawenna Kentyohkwa focuses on mastering the complex ver-

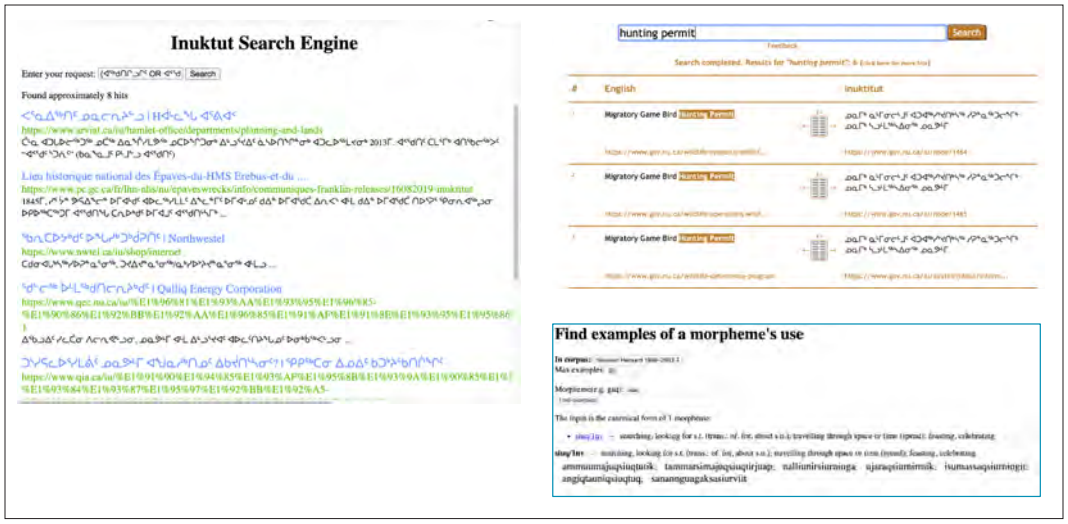


Figure 5. Inuktut Search Engine, WeBInuk (Concordancer), and Morpheme Example Search.

bal morphology of Kanyen'kéha. This is why NRC-ILT and the school jointly developed Kawennón:nis (above).

Kawennón:nis has one limitation: students can generate complex words in text, but do not necessarily know how to pronounce them. The goal of the school is to produce fluent *speakers*, not just *readers/writers*, so there is a gap between what Kawennón:nis produces (text) and what students need most (speech to imitate). Yet just as it would not be feasible to compose and print Kawennón:nis in book form because of its enormous implicit vocabulary, it would likewise not be feasible to **record** all forms. This suggests implementation of TTS – can we train a TTS system on a recorded, phonetically balanced subset of the vocabulary?

We built a concatenative TTS prototype. It rearranges subword units from 852 recordings of conjugations from fluent speakers into complete coverage of the first 122,966 conjugations in Kawennón:nis. The teachers at the school declared the quality of the output speech to be better than acceptable. We have also recently developed a neural TTS model capable of producing acceptable-quality synthesized speech from as few as 30 minutes of audio. Owennatékha recently declared TTS to be in his top three technological priorities for Kanyen'kéha.

Kawennón:nis illustrates the challenges facing Indigenous curriculum developers: complex languages, few speakers to teach the language, few recorded materials, and few speakers to record new materials, and thus large parts of the language that students never get to hear. TTS enables resource-strapped organizations to make the most of their most precious, limited resource – fluent speakers' time – by leveraging a small number of recordings to cover a much larger domain.

MIDDLE RING: SPEECH TECHNOLOGIES DEVELOPED BY CRIM

Annotation and transcription are the biggest part of the workload in most language documentation and conservation projects. Cox et al. (2019) discuss the "transcription bottleneck": speech (especially of Elders) is being recorded in Indigenous languages much faster than it is being transcribed, or even an-

notated (it should be called the "annotation bottleneck" instead). This is true of minority languages across the world. Annotation and especially transcription take time, and are tedious. It is tempting to record many hours of speech and plan to carry out these tedious steps "later on" – they are often postponed indefinitely …

This has been going on for decades, long before digital recording media were available. All over Canada, there are thousands of hours of speech in Indigenous languages on magnetic tapes or audio cassettes that are potentially valuable resources for language revitalization, but that have become inaccessible in practice because nobody knows what is in them. They are scattered across communities and university departments (I have heard stories of magnetic tapes with important recordings being found, unlabeled, on the shelves of a deceased professor's garage). Nobody has the time to listen to most such recordings; in some tragic cases, nobody is left who still understands the language that was recorded. Older recordings have typically never been copied over into a digital medium.

CORE RESEARCH

We funded the Centre de recherche informatique de Montréal (CRIM) to research automatic speech recognition (ASR) for a small number of Indigenous languages spoken in Canada. Full-vocabulary ASR for such languages is unrealistic in the short term, because the large amounts of parallel speech-transcription data required to train ASR systems do not exist. There was not enough funding available to create such data. Thus, the goal was to create imperfect ASR systems trained on small amounts of parallel data. Such systems have been shown, for other languages, to make "audio search" (also termed "keyword spotting" or "spoken term detection") feasible. When ASR accuracy is low, it may still be possible for users to find certain words or phrases in speech with audio search techniques.

ASR experiments at CRIM have focused on Inuktitut and East Cree. The CRIM researchers arranged for nearly 81 hours of Inuktitut speech, and 102 hours of East Cree speech, to be transcribed; they split these parallel data into training and test sets. Their subsequent work on Inuktitut and East Cree ASR is described in (Gupta and Boulianne, 2020a) and (Gupta and Boulianne, 2020b), respectively. The work benefited from collaboration with Pirurvik Centre (https://www.pirurvik.ca/) and the Canadian Broadcasting Corporation (https://www.cbc.ca/).

A key question for the CRIM team was what units to use for ASR. Inuktitut is highly polysynthetic; thus, the language has a very high "out of vocabulary" rate (60%) – many words are only ever spoken once. An ASR system whose vocabulary is made of words will misrecognize most words in new Inuktitut speech data. The CRIM researchers tried morphemes, syllables and phonemes as the basic ASR unit, and obtained the best results with syllables. But even with syllable units, about 71% of Inuktitut words in the test data were misrecognized.

In unpublished work, these researchers nevertheless managed to get reasonably good keyword-spotting performance for Inuktitut *(personal communication)*. The best-performing approach was conventional: search through a phoneme lattice generated by syllable-based ASR, with a weighted confusion lattice providing robustness to ASR errors.

In CRIM experiments, East Cree had a significantly lower "out of vocabulary" rate than Inuktitut, though much smaller amounts of textual data were available for building an East

Cree vocabulary. With a 30,000-word vocabulary obtained from text in two genres – transcripts of video stories and Bible translations – the "out of vocabulary" rate for held-out text data was 25% for video stories and 9% for Bible data. These numbers made words a plausible unit for East Cree ASR. In ASR experiments on held-out East Cree speech data, 70% of words in the test video stories were misrecognized, and 25% of the words in the test Bible data. No keyword-spotting experiments were conducted for East Cree.

The CRIM researchers obtained extremely promising experimental results (low rates of misrecognition) for both Inuktitut and East Cree when the ASR system was speaker-**dependent**. The poor results quoted above are speaker-**independent.** For Indigenous language revitalization, however, where many hours of speech are often collected from each of a small number of fluent Elders, speaker-**dependent** ASR would be practically useful. Consider a mode of work in which (e.g.) a couple of hours of speech from an Elder are transcribed by hand by a human expert, as is done now. Then these two transcribed hours would be used to train an ASR system adapted to the speech of this particular Elder, which could produce a first-draft transcription of the many remaining hours of speech from the same Elder. This would alleviate the "transcription bottleneck".

PRODUCTIVITY TOOLS

The CRIM team also released tools that make early stages of processing recorded speech easier. These were packaged as Web services on CRIM's VESTA platform (http://vesta.crim.ca) and are available there or through an ELAN extension – ELAN is an annotation tool at https://archive.mpi.nl/tla/elan/ (Wittenburg et al., 2006). The tools enable segmenting speech files into speech vs. non-speech (silence, noise, music, etc), language retrieval (finding segments spoken in a particular language, with 32 languages being identifiable), speaker retrieval (finding segments spoken by a particular speaker), multichannel voice activity detection (detecting segments containing speech separately for each track in a multichannel, multi-microphone recording), and other useful capabilities.

Drs. Chris Cox and Olivia Sammons (of Carleton University and First Nations University respectively), used these CRIM tools recently while collecting Michif speech data. They report a 4-to-5-fold speedup for early stages of annotation *(personal communication)*.

OUTER RING: OTHER WORK FUNDED BY NRC-ILT

This section outlines Indigenous language revitalization subprojects to which NRC-ILT funding contributed – at least partially – but which were managed outside NRC.

ONLINE COURSES AND GAMES

Online language learning tools must be updated frequently because the underlying software has become obsolete. For instance, Adobe stopped supporting Flash Player in early 2021. NRC-ILT funded remedial measures for several subprojects.

Learning Platforms at Carleton University:
Prof. Marie-Odile Junker and her Carleton University team have been collaborating for years with Indigenous partners to develop online lessons for Algonquian languages. She uses a participatory-action framework to work with communities. Her team's contributions are too numerous to list here – see https://www.marieodile-junker.ca/. Prof. Junker holds the Governor General's Innovation Award (2017).

NRC-ILT funding began to flow to the pre-existing Carleton project in 2018. Software changes had partially stranded educational language tools on Carleton's web-based platform for Innu (not to be confused with Inuktut, an unrelated language) and East Cree. The funding helped pay for a technological update (removal of dependence on Flash and other outdated technologies) and for the expertise of two Indigenous language experts who contributed large amounts of new pedagogical content. Prof. Junker's team continues to develop the platform with other funding (Innu: https://lessons.innu-aimun.ca/; East Cree: https://lessons.eastcree.org/).

7000 Languages:
Some Indigenous communities wished to develop online courses for their languages. We provided modest amounts of funding to 7000 Languages, a non-profit that creates free language learning software, to jointly create courses with these communities (whom we also funded). Three of these courses are online at the 7000 Languages site, www.7000.org: for Kwak'wala, Michif, and Mi'kmaq.

Computer-assisted Language Learning (CALL) at the University of Alberta:
This subproject supports the Y-dialect of Cree (Plains Cree). The U. Alberta team, together with members of Cree-speaking communities, has been creating an adaptive CALL system called "CreeTutor". Cree-language content has also been developed: 13 personal stories from 8 Elders have been recorded, transcribed, and translated. CreeTutor will soon go live. A key element of CreeTutor is *Sound Hunters,* a receptive phonemic awareness activity (Lothian et al., 2020). The main creator of *Sound Hunters,* Delaney Lothian (Cree-Métis), is a U. Alberta MSc student and a part-time NRC-ILT member. She is working on a Michif version of *Sound Hunters*.

FirstVoices:
The First Peoples' Cultural Council (FPCC) of British Columbia has a strong record of providing state-of-the-art technologies, training and technical support to Indigenous language activists (mainly but not exclusively in the province), within its FirstVoices program (http://www.fpcc.ca/about-us/). The Language Tutor, which enables communities to build language lessons, is part of FirstVoices and had become technologically "stranded". ILT-NRC funding helped to redress this situation.

On the Path of the Elders:
"On the Path of the Elders" is a role-playing game designed to acquaint players with historical and cultural topics related to the James Bay region. When this free online game was released in 2007, it was widely praised for its innovative use of historical resources. It suffered loss of functionality due to changes in the underlying software industry. NRC-ILT supplied funding to update the software and add Swampy Cree content. The "On the Path of the Elders" website is now mobile-friendly and compatible with Chrome, FireFox and Internet Explorer browsers: www.pathoftheelders.com.

CAPACITY-BUILDING WORKSHOPS

Two subprojects funded by NRC-ILT provided training for people who document Indigenous languages. NRC-ILT also provided modest financial assistance to the language activist Caroline Running Wolf; she organized four successful online sessions for language activists, the "Community Workshops for Indigenous Language Technology" (CWILTs) in Dec. 2020 – July 2021.

Yukon Native Language Centre (YNLC):
Eight Indigenous languages (Gwich'in, Hän, Kaska, Northern Tutchone, Southern Tutchone, Tagish, Tlingit, and Upper Tanana) are spoken in the Yukon: almost 15% of Indig-

enous languages in Canada. Helped by funding from NRC-ILT, YNLC supported 12 trainees in a 10-month program during which they acquired practical skills in developing, disseminating, and preserving digital materials for these languages in their communities. The Covid-19 pandemic disrupted some of the planned activities in this subproject, which was centred on in-person activities, but competent management by YNLC staff made it very successful nevertheless.

- In Workshop 1 (October 2019), trainees practiced how to use the video equipment and how to carry out interviews with fluent speakers. They were also trained in file management.
- In Workshop 2 (November 2019), attendees were trained to use the ELAN and SayMore software suites for transcribing, annotating, and translating speech recordings.
- Workshop 3 (February 2020), began with review and practice of skills learned earlier. It then covered sharing and repurposing of videos, and using ELAN materials as a resource for language learning and teaching. On the last day, students showed their finished videos.

The 12 trainees created 548 minutes of documentation and mastered skills that will enable them to record Elders in their communities on a continuing basis.

Indigitization:

This is a partnership between the University of British Columbia, the Musqueam Archives, and the Heiltsuk Cultural Education Centre that focuses on digitization of speech data. While the YNLC subproject trained members of Indigenous communities to collect **new** data, the continuing Indigitization subproject trains people to convert **old** data into digital, accessible formats. It helps make Indigenous communities autonomous by familiarizing them with hardware and software options for digitization. For communities that already have digitized content, the subproject offers training in transcription of speech into Indigenous languages, translation of the transcriptions into English or French, and management of the resulting archives.

Indigitization began in January 2020, and underwent re-planning because of the COVID-19 crisis (in-person workshops became inadvisable). More resources were devoted to creation of instructional resources than in the original plan: "We have a unique opportunity to write experience-based, practical guides that will help small, often underfunded, Indigenous organizations to start managing their collections in a structured manner" *(personal communication from Gerry Lawson, Indigitization project lead)*.

The text content for 65 of these new guides has been developed; the guides are currently undergoing editing and graphic design. This subproject plans to lend hardware to communities when necessary.

CWILTS:

CWILTs ("Community Workshops for Indigenous Language Technology") is a series of video sessions for Indigenous language activists organized by Caroline Running Wolf of Apsáalooke Nation (Crow), each featuring lightning talks centered on a language revitalization theme followed by conversations in breakout rooms. Her goal is to build a community of practice. Many CWILTS attendees have kept in touch with each other – there has been progress towards this goal.

- Session 1 (Dec. 7, 2020): 26 attendees. Ms. Running Wolf and Sara Child (Kwakwaka'wakw community) introduced CWILTs; Delaney Lothian (Cree, Métis) presented *Sound Hunters*, her game for training learners to recognize Cree phonemes and syllabics (above); Eddie Antonio Santos presented Western Cree syllabics conversion software (above).

- Session 2 (Jan. 7, 2021): 33 attendees. Stephanie Tenasco (Kitigan Zibi Anishinabe) shared her experience of creating language learning videos; Patrick Littell and Aidan Pine (NRC) presented ReadAlong Studio.
- Session 3 (Jan. 28, 2021): 80 attendees (most were from Canada but there were 13 from USA, 2 from Norway, 1 from Mexico, 1 from Bangladesh, and 2 from Morocco). Theme: "Strengthening Language through Technology: A Context for Capacity Building". Aaron Plahn and Justin Bambrick presented their corpus collection and capacity-building initiative for Tŝilhqot'in: https://www.youtube.com/watch?v=L-eX-WFHnCSc&feature=youtu.be .
- Session 4 (July 14, 2021): number of attendees not recorded (at least 40). Sean Smith, Krista Dempster, and Dustin Hill of YNLC described capacity building in their subproject.

SINGLE-LANGUAGE SUBPROJECTS

Cree text digitization at Blue Quills University:
Blue Quills was the first university in Canada to be fully owned and operated by First Nations people. This subproject digitized and indexed the largest known collection of Cree syllabics text: approximately 40,000 pages from monthly newsletters called kihcitwāw miteh (Sacred Heart) produced by the Catholic Church between 1906 and 1978. The corpus will soon be available at http://language-unbq.bluequills.ca.

Kanyen'kéha data collection at TTO:
This effort is led by Nathan Thanyehténhas Brinklow, a language teacher at Queen's University in Ontario, in collaboration with the Tsi Tyónnheht Onkwawén:na Language and Cultural Centre (TTO) on Tyendinaga Mohawk Territory. The goal is creation of text and audio corpora to support research on ASR for Kanyen'kéha. This subproject has collected 112,420 Kanyen'kéha written words and 26 hours of Kanyen'kéha speech, covering genres such as scripts and audio from movies and TV shows, translated books of the Bible, and recordings of Elders.

Kwak'wala data collection:
This effort also involves corpus collection followed by ASR research. Kwak'wala (Wakashan family) is spoken by 18 Kwakwaka'wakw Nations whose traditional territory is on northern Vancouver Island, nearby islands, and the adjacent mainland. NRC-ILT funding supported a partnership among two community-based teams from three different Kwakwaka'wakw communities, a technical team, and three university-based linguists (two of whom are Kwakwaka'wakw). This subproject has compiled 25 hours of machine-readable Kwak'wala audio data, consisting of conversational speech, pedagogical materials, and storytelling by Elders. It has also identified over a hundred hours of further Kwak'wala speech recordings in various locations.

Michif talking dictionary:
This subproject mobilized and made accessible an out-of-print Michif dictionary called "The Michif Dictionary: Turtle Mountain Dictionary Chippewa Cree", first published in 1983. Partly funded by NRC-ILT, and with assistance from Michif first-language speakers, computational linguists, and others, the team developed a digital, spoken version of this resource. The subproject produced over 181 hours of high-quality audio recordings of the dictionary from four speakers; the dictionary includes 15,422 entries. All 350 pages of Michif lexical entries and example sentences have been recorded by at least one speaker, with some entries being recorded by two or more speakers. The link to the dictionary: https://dictionary.michif.org/home.

The final report for this subproject refers to the new CRIM technologies described above: "Each recording was automatically segmented into pause-delimited utterances using a Deep Neural Network (DNN) voice activity detection service that was developed within the VESTA-ELAN project by the Computer Research Institute of Montréal (CRIM). This auto-segmentation saved an immeasurable amount of time in the annotation process" *(communication from Heather Souter and Olivia Sammons, subproject co-leads).*

Nsyilxcn:
The Nsyilxcn language (Interior Salish family) was traditionally spoken in the Okanagan Valley in present-day British Columbia. It is critically endangered, with a dozen highly fluent Elders remaining. This subproject began in December 2019; it was carried out by Syilx Language House (SLH), a community-based organization, with some funding from NRC-ILT. Seven hours of fluent Elder stories in Nsyilxcn were recorded, transcribed (with a glossary provided in English) and archived. All recordings are shared at www.thelanguagehouse.ca. 4 trainees and 10 volunteer learners were trained to fluency as part of the subproject.

SENĆOŦEN:
SENĆOŦEN (Coast Salish family) was spoken just north of present-day Victoria, British Columbia. It is the most severely endangered language NRC-ILT worked with: until recently, there were only five fluent speakers left. However, the community is engaged in a vigorous language revitalization effort, led by the WSÁNEĆ School Board. During 1981-1991, the linguist Dr. Timothy Montler recorded Elsie Claxton, the last monolingual SENĆOŦEN speaker, telling stories (Montler, 2018). NRC-ILT paid for two Elders to work with Dr. Montler to transcribe these recordings, and translate them into English. This material will be an invaluable resource for teaching of the language.

Tŝilhqot'in:
Tŝilhqot'in (Dene, i.e. "Athabaskan" family) was traditionally spoken in the southern interior of present-day British Columbia. Funding from NRC-ILT enabled the Tŝilhqot'in National Government (TNG) to build on an already impressive language revitalization effort. TNG had already recorded 35,000 audio clips, developed a linguistic database, published verb paradigms in a "Digital Verb Book", and created a website with diverse learning tools: www.tsilhqotinlanguage.ca. With our funding, TNG recorded a dozen hours of speech from 20 speakers; 20 hours of speech were digitized, transcribed, and aligned with the transcriptions. Labeling of 46,000 audio clips was completed. A list of 26,200 English words and phrases to be translated has been drawn up; many of these have been translated into Tŝilhqotin. Four community members were trained in IT-oriented roles.

A superb video on this community's language revitalization efforts from CWILTs is at https://www.youtube.com/watch?v=L-eXW-FHnCSc&feature=youtu.be .

FUTURE WORK AND CONCLUSION

Brinklow (2021) states that it is possible for communities to collaborate productively with some non-Indigenous partners on Indigenous Language Technology (ILT):

"Many communities in Canada are collaborating with non-profit partners for the development of ILT outside of a profit-driven ecosystem … These innovative partnerships include universities, communities, research institutions, governments, and others … In the Canadian context, these partnerships (ironically, often funded by the original colonizing governments) are producing ground-breaking language technologies that are innovative by any standard. "

This Indigenous author has kind words for an NRC-ILT subproject: "One pioneering partnership has developed between Onkwawenna Kentyohkhwa at Six Nations in Ontario and the National Research Council (NRC) of Canada. The partnership is between the NRC-ILT project team and local teachers to develop a verb-conjugator for Kanyen'kéha … While the initial development was done with a specific language (Mohawk), the underlying tool was created to work with any language in a 'first deep, then broad' approach to design and development … The partnership with the NRC adds value to the project at the national level because a broad approach to development is not the responsibility of an individual language community."

What are the next steps?

Our priorities are:

- Extending technologies we have already developed to new languages.
- Making these technologies much easier to use. It's not enough to release open-source software (OSS); the "empowerment" approach implies that such OSS should be easy to use for people inside Indigenous communities. It should not require an advanced degree in computer science or computational linguistics. An example of our current efforts in this direction: work inside NRC-ILT led by Dr. Patrick Littell on "Gramble" – a spreadsheet-based software framework that, we hope, will make it much easier to create verb conjugators.
- A new focus for NRC-ILT will be text-to-speech (TTS). It is preferable for students to learn a language from fluent human speakers. However, given the shortage of teachers for most Indigenous languages, TTS may sometimes be an acceptable second-best. An NRC-ILT team member, Aidan Pine, already created a prototype TTS system for Kanyen'kéha (see above); recently, he has built experimental TTS systems for two other Indigenous languages spoken in Canada, Gitksan and SENĆOŦEN, that yielded surprisingly good voice quality.

The prospect for many Indigenous languages spoken in Canada is brighter than it has been for decades – thanks to the efforts of Indigenous peoples themselves. The NRC-ILT project shows that it is possible for non-Indigenous organizations, even governmental ones, to develop technology in collaboration with Indigenous stakeholders that is useful to these language revitalization efforts. However, non-Indigenous organizations active in this area must remember that their role is a secondary one, and that their goal should be empowerment of Indigenous communities.

BIBLIOGRAPHY

Rubèn Fernández Asensio (editor). 2019. "Old kava in new gourds: language revitalization and schooling in Hawaii". In *Linguapax Review*. https://www.linguapax.org/wp-content/uploads/2020/02/linguapax19-1-1.pdf

Peter Bakker. 1997. "A language of our own". Oxford & New York: Oxford University Press.

Nathan Thanyehténhas Brinklow, Patrick Littell, Delaney Lothian, Aidan Pine, and Heather Souter. 2019. "Indigenous Language Technologies and Language Reclamation in Canada". In *Collection of research papers of the 1st international conference on language technologies for all*. European Language Resources Association (ELRA). https://lt4all.elra.info/proceedings/lt4all2019/pdf/2019.lt4all-1.100.pdf.

Nathan Thanyehténhas Brinklow. 2021. "Indigenous Language Technologies: Anti-Colonial Oases in a Colonizing (Digital) World". In *WINHEC: International Journal*

of Indigenous Education Scholarship, pp. 239-266; Special Issue on Indigenous Language Revitalization. Innovation, Reflection and Future Directions. http://dx.doi.org/10.18357/wj1202120288.

Michael J. Chandler and Christopher Lalonde. 1998. "Cultural continuity as a hedge against suicide in Canada's First Nations". In *Transcultural Psychiatry*, 35(4):191–219.

Christopher Cox, Gilles Boulianne, and Jahangir Alam. 2019. "Taking aim at the transcription bottleneck: Integrating speech technology into language documentation and conservation". [Conference Presentation]. *International Conference on Language Documentation & Conservation*. https://scholarspace.manoa.hawaii.edu/bitstream/10125/44841/44841.pdf

Ewa Czaykowska-Higgins. 2009. "Research models, community engagement, and linguistic fieldwork: Reflections on working within Canadian Indigenous communities". In *Language documentation & conservation*, 3(1):182–215.

Fineen Davies, Eddie A. Santos, and Heather Souter. 2021. "On the Computational Modeling of Michif Verbal Morphology". *EACL*. https://aclanthology.org/2021.eacl-main.226.pdf.

Petra Fachinger. 2019. "Colonial violence in sixties scoop narratives: from In Search of April Raintree to A Matter of Conscience". In *Studies in American Indian Literatures*, 31(1-2):115.

Benoît Farley. 2012. "The Uqailaut project". [Website]. http://www. inuktitutcomputing.ca.

First Voices Portal. 2021. "SENĆOŦEN Home Page". Downloaded from https://www.firstvoices.com/explore/FV/sections/Data/THE%20SEN%C4%86O%C5%A6EN%20LANGUAGE/SEN%C4%86O%C5%A6EN/SEN%C4%86O%C5%A6EN on Oct. 1, 2021.

Government of Canada. 2015. "Final report of the Truth and Reconciliation Commission". http://nctr.ca/reports.php.

Vishwa Gupta and Gilles Boulianne. 2020a. "Automatic transcription challenges for Inuktitut, a low-resource polysynthetic language". In *Proceedings of LREC*. https://aclanthology.org/2020.lrec-1.307.pdf.

Vishwa Gupta and Gilles Boulianne. 2020b. "Speech transcription challenges for resource constrained Indigenous language Cree". In *Proceedings of 1st Joint SLTU-CCURL Workshop*. https://aclanthology.org/2020.sltu-1.51/

Darcy Hallett, Michael J. Chandler, and Christopher E. Lalonde. 2007. "Aboriginal language knowledge and youth suicide". In *Cognitive Development*, 22(3):392–399.

Eric Joanis, Rebecca Knowles, Roland Kuhn, Samuel Larkin et al. 2020. "The Nunavut Hansard Inuktitut–English parallel corpus 3.0 with preliminary machine translation results". In *Proceedings of LREC-2020*. https://aclanthology.org/2020.lrec-1.312.pdf .

Anna Kazantseva, Owennatékha Brian Maracle, Ronkwe'tiyohstha Josiah Maracle, and Aidan Pine. 2018. "Kawennón:nis: the wordmaker for Kanyen'kéha". In *COLING Workshop on Computational Modeling of Polysynthetic Languages*. https://aclanthology.org/W18-4806.pdf

Te Taka Keegan. 2019. Issues with Māori sovereignty over Māori language data. [Video Presentation]. http://video.web.gov.bc.ca/public/fpcc/letlanguageslive.html.

Jeannette King. 2013. "Te Kōhanga Reo: Māori Language Revitalization". In L. Hinton and K. Hale (editors), *The Green Book of Language Revitalization in Practice*. Brill.

Rebecca Knowles, Darlene Stewart, Samuel Larkin, and Patrick Littell. 2020. "NRC Systems for the 2020 Inuktitut-English

News Translation Task". In *Proceedings of WMT*. https://aclanthology.org/2020.wmt-1.13/.

Roland Kuhn, Fineen Davis, Alain Désilets, Eric Joanis et al. 2020. "The Indigenous Languages Technology project at NRC Canada: An empowerment-oriented approach to developing language software". In *Proceedings of COLING 2020*. https://www.aclweb.org/anthology/2020.coling-main.516.pdf.

Patrick Littell, Anna Kazantseva, Roland Kuhn, Aidan Pine, Antti Arppe, Christopher Cox, and Marie-Odile Junker. 2018. "Indigenous language technologies in Canada: Assessment, challenges, and successes". In *27th International Conference on Computational Linguistics (COLING 2018)*, Santa Fe, New Mexico, USA. https://www.aclweb.org/anthology/C18-1222.pdf

Delaney Lothian, Gokce Akcayir, Anaka Sparrow, Owen Mcleod, and Carrie Demmans Epp. 2020. "SoundHunters: Increasing Learner Phonological Awareness in Plains Cree". In *Artificial Intelligence in Education*. doi: 10.1007/978-3-030-52237-7_28.

Joel Martin, Howard Johnson, Benoît Farley, and Anna Maclachlan. 2003. "Aligning and using an English-Inuktitut parallel corpus". In *Proceedings of the HLT-NAACL 2003 Workshop on Building and using parallel texts*, Volume 3, pages 115–118. Association for Computational Linguistics.

Timothy Montler. 2018. "SENCOTEN: a dictionary of the Saanich language". University of Washington Press, ISBN: 9780295743851.

Nicole Rosen and Heather Souter. 2009. "Language revitalization in a multilingual community: the case of Michif." In *1st International Conference on Language Documentation and Conservation (ICLDC)*, Honolulu.

Statistics Canada. 2017. "Proportion of mother tongue responses for various regions in Canada, 2016 census". https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/dv-vd/lang/index-eng.cfm.

Olivia N. Sammons. 2019. "Nominal classification in Michif". *Ph.D. thesis, University of Alberta*. https://era.library.ualberta.ca/items/903b7ee2-4f3f-4959-b4f0-8231205d5d67/view/9a967b05-b223-4716-97de-6da32e6931fe/Sammons_Olivia_N_201901_PhD.pdf

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. "ELAN: A professional framework for multimodality research". In *European Language Resources Association (ELRA)*.

EL PROJECTE *INDIGENOUS LANGUAGES TECHNOLOGY (ILT)* DEL CONSELL NACIONAL D'INVESTIGACIÓ DE CANADÀ, I EL SEU CONTEXT

Roland Kuhn

INTRODUCCIÓ

En les últimes dècades, s'han creat centenars de projectes de revitalització i recuperació de llengües indígenes per tot Canadà. La majoria es troben a comunitats indígenes; d'altres se centren a universitats. Varien en dimensió des de les ambicioses iniciatives multilingües del First Peoples' Cultural Council (FPCC)[1] que es troba dins de i és finançat per la Província de Columbia Britànica, i és actiu allà i a altres regions de Canadà (https://fpcc.ca/), fins a iniciatives d'un o dos voluntaris sense finançament a comunitats remotes.

El projecte que tinc l'honor de liderar descrit en aquest article —i a aquesta pàgina web que s'actualitza periòdicament: https://nrc.canada.ca/en/research-development/research-collaboration/programs/canadian-indigenous-languages-technology-project— es gestiona des del Consell Nacional d'Investigació de Canadà (NRC, per les seves sigles en anglès, *National Research Council of Canada*), que és la principal organització d'investigació i tecnologia del Govern de Canadà. Tanmateix, el lector no hauria de concloure que els governs "colons" (no indígenes) juguen un paper clau a la revitalització de les llengües indígenes a Canadà. Tots els projectes exitosos de revitalització lingüística que em consten els gestionen les comunitats indígenes o hi col·laboren estretament activistes de llengües indígenes.

El nostre projecte *Indigenous Languages Technology* (ILT)[2] al NRC (d'ara endavant, "NRC-ILT") **no** constitueix un projecte de revitalització lingüística: construeix eines que a vegades resulten útils per a les persones que treballen en la revitalització lingüística. L'equip del NRC-ILT som com els tècnics d'il·luminació o tramoistes al teatre: no som els actors (no sortim a l'escenari), ni som els encarregats de la producció. Aquests papers els juguen els activistes de llengües indígenes i les seves comunitats. La revitalització lingüística seguiria endavant sense nosaltres. Hi ha moments, però, on la nostra ajuda tècnica pot facilitar lleugerament la feina. Tot el que hem aconseguit ha estat gràcies a la col·laboració amb les parts interessades indígenes.

Avís: aquest article és l'opinió subjectiva i parcial d'un individu sobre les tecnologies que s'estan aplicant a la revitalització lingüística a Canadà, amb una descripció del projecte NRC-ILT. No pretén ser un estudi exhaustiu del camp; exclou moltes, probablement la majoria de les organitzacions actives en la revitalització lingüística, i els seus èxits. Aquest article és una versió ampliada de (Kuhn et al., 2020), per la qual cosa em sento profundament en deute amb les contribucions dels meus coautors en aquell treball.

1.- Consell cultural dels primers pobles.
2.- Projecte de tecnologies per a llengües indígenes.

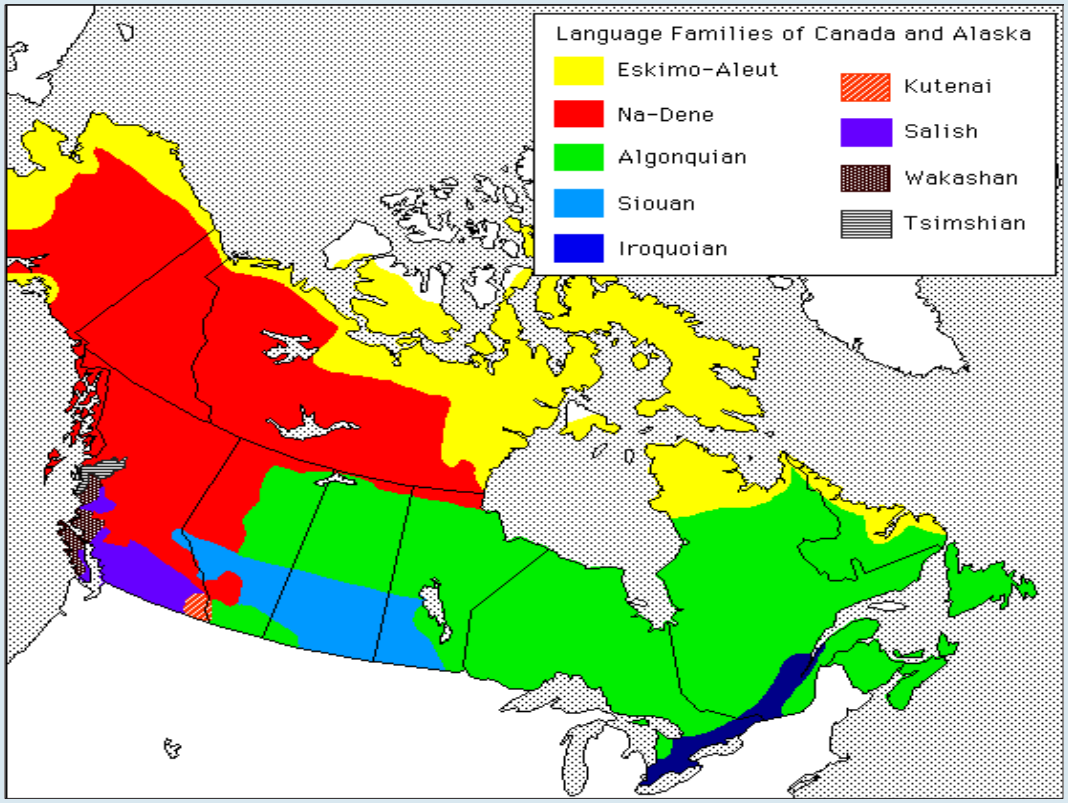


Figura 1. Llengües indígenes a Canadà durant el primer contacte amb europeus .
(Mapa de Matthew Dryer, Universitat de Buffalo)

LES LLENGÜES INDÍGENES A CANADÀ: HISTÒRIA I DEMOGRAFIA

La **Figura 1** mostra la diversitat de famílies de llengües indígenes al que després va esdevenir Canadà, en el moment del primer contacte amb els europeus. Cada família era i és diferent de les altres, en quant a fonètica, vocabulari i sintaxi —en alguns casos, tan diferent com les famílies de llengües germàniques, turques i sinotibetanes d'Euràsia ho són entre elles. Les llengües que eren veïnes geogràfiques, malgrat ser de diferents famílies, a vegades manllevaven paraules i aspectes fonètics o sintàctics entre sí.

Les comunitats indígenes a Canadà sovint es denominen amb tres termes: Inuit, Métis, i Primeres Nacions. Els inuit són un grup ètnic cohesiu, al igual que els métis. No obstant, el terme "Primeres Nacions" agrupa a tots els altres pobles indígenes a Canadà, un conjunt divers d'etnicitats.

La majoria de les llengües indígenes parlades a Canadà són polisintètiques: una sola paraula sovint expressa un significat complex que necessitaria una frase sencera en llengües no polisintètiques. Normalment, una paraula també consta de més morfemes que en altres llengües. Una paraula anglesa consta d'uns 3 morfemes de mitjana; una paraula en mohawk consta d'entre 5-6 morfemes de mitjana. Les llengües mohawk (família iroquesa), cree (família algonquina), i inuktitut (família esquimoaleutiana) són totes polisintètiques, tot i pertànyer a famílies diferents en altres aspectes.

Després del contacte entre els pobles indígenes i els europeus, van sorgir llengües criolles que combinaven aspectes tant de les

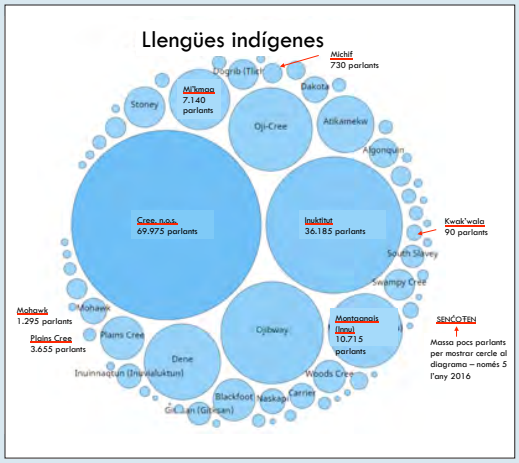


Figura 2. Nombre de parlants per llengua (cens de l'any 2016)
https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/dv-vd/lang/index-eng.cfm

llengües indígenes com europees; la majoria d'elles han desaparegut. Per exemple, el contacte entre comerciants holandesos i els mohawk a la vall del riu Hudson va donar lloc a un crioll holandès-mohawk. L'argot chinook a la seva última forma (va sobreviure fins el segle XX) contenia elements de llengües indígenes de la costa oest, rus, hawaià i xinès.

Els métis van sorgir com a nació pròpia d'una mescla d'avantpassats indígenes i europeus a principis del segle XIX (Bakker, 1997; Rosen & Souter, 2009). El bungi, un crioll parlat pels Métis de Red River, en l'actual Manitoba, constava d'elements de l'anglès escocès d'Orkney, el gaèlic escocès, el francès, el cree i l'anishinaabemowin (ojibwa). Tanmateix, la majoria dels métis parlaven el michif, una llengua de contacte. El michif encara sobreviu i és el protagonista de moltes iniciatives de revitalització lingüística. A grans trets, podríem dir que combina substantius principalment francesos amb verbs en cree de la plana (de la família algonquina); és majoritàriament polisintètic (Davies, Santos, i Souter, 2021; Sammons, 2019).

Després de la Confederació de Canadà l'any 1867, les polítiques governamentals es van enfocar en erradicar les llengües indígenes i les costums indígenes tradicionals, amb l'objectiu d'integrar els pobles indígenes a la societat "blanca". La *Indian Act*, la llei relativa als indígenes, dissuadia i sovint declarava il·legals les trobades culturals i parlar llengües ancestrals. Molts infants indígenes van ésser llevats a la força de les seves comunitats i van ésser col·locats a internats obligatoris ("escoles residencials") pel govern federal. Les esglésies eren les encarregades d'aquestes escoles; algunes les gestionava l'Església catòlica, d'altres, esglésies protestants. S'han documentat múltiples formes d'abús psicològic, físic i sexual que tenien lloc a aquestes escoles. Segons la Comissió canadenca de la Veritat i la Reconciliació[3], el sistema d'escoles residencials es va crear amb l'objectiu de separar els infants aborígens de les seves famílies, per minimitzar i debilitar els lligams familiars i les connexions culturals (Government of Canada (2015), prefaci). A més, durant un temps a mitjans del segle XX, es van separar molts infants indígenes de les seves famílies d'origen per ésser adoptats per famílies no indígenes, denominat *Sixties Scoop* (Fachinger, 2019).

Tot i aquesta història de discriminació sistemàtica, les comunitats indígenes s'han resistit a integrar-se i han persistit en conservar i ensenyar les seves llengües. Els avantatges d'aprendre una llengua ancestral s'han vinculat amb un millor benestar entre els joves (Chandler i Lalonde, 1998; Hallett et al., 2007).

Segons la meva observació personal durant el projecte NRC-ILT, penso que moltes de les llengües indígenes parlades a Canadà

3.- Truth and Reconciliation Commission of Canada.

estan revifant. Quantitats històriques de joves de comunitats indígenes —que sovint s'han criat parlant només anglès o francès (al Quebec)— s'estan inscrivint a cursos que els permetran parlar-se en les seves llengües ancestrals. He conegut a individus indígenes extraordinaris —gairebé tots mal pagats, i alguns voluntaris no retribuïts— que ensenyen aquestes llengües. Aquests activistes lingüístics formen una elit: un grup de persones que combina l'idealisme amb una gran intel·ligència i capacitat per esforçar-se. Ensenyar una llengua polisintètica a persones que originàriament només parlaven anglès o francès no és apte per a dèbils.

La **Figura 2** mostra el nombre de parlants de cada llengua indígena l'any 2016, segons el cens de Canadà (Statistics Canada, 2017). Les llengües subratllades en vermell són amb les quals el projecte NRC-ILT ha interactuat d'alguna manera. Les xifres que es mostren no s'han de prendre al peu de la lletra; hi ha experts que no estan d'acord amb algunes d'elles. Tanmateix, aquesta figura ens dona una idea aproximada del pes demogràfic d'aquestes llengües. La llengua cree de la plana (família algonquina) és la que té més parlants; l'inuktitut (família esquimoaleutiana) ocupa la segona posició. L'inuktitut és, excepcionalment, una important llengua de govern (al territori Nunavut). L'existència de burocràcia governamental que funciona parcialment en inuktitut significa que existeixen molts més recursos escrits i orals per aquesta llengua que per qualsevol altra llengua indígena a Canadà. La disparitat amb la llengua cree, que té molts més parlants però molts menys recursos lingüístics, perquè no hi ha cap òrgan de govern que l'utilitzi, és notable.

El lector no hauria de concloure que la majoria de les llengües "mitjanes" o "petites" mostrades tenen escasses possibilitats a llarg termini. La figura no pot mostrar l'energia relativa dels esforços de revitalització per a cada llengua. Per exemple, la llengua més petita mostrada, saanich (SENĆOŦEN), que només tenia 5 Ancians que dominessin la llengua l'any 2016, compta amb un programa de revitalització lingüística ben gestionat i rigorós que està ensenyant la llengua a molts joves de la comunitat —el següent cens comptarà amb molts més parlants de SENĆOŦEN (Montler, 2018; First Voices Portal, 2021).

De forma similar, la llengua mohawk (kanyen'kéha), amb 1.295 parlants a la **Figura 2**, està experimentant un creixement considerable, gràcies a bones escoles d'idiomes a vàries comunitats mohawk. L'escola de kanyen'kéha per adults a la comunitat Six Nations of the Grand River al sud-oest d'Ontàrio (Onkwawenna Kentyohkwa; vegeu https://onkwawenna.info) és admirada per educadors indígenes a tot Canadà pel seu èxit en graduar molts estudiants que parlen la llengua amb fluïdesa després de dos anys d'immersió intensiva (gairebé s'havia extingit la llengua a Six Nations, tot i que no a totes les comunitats mohawk). Alguns graduats es casen, parlen kanyen'kéha a casa, i crien els seus fills en aquesta llengua, que esdevé la seva llengua materna. L'energia, l'experiència educativa i la cohesió comunitària tenen un gran impacte positiu en la revitalització lingüística —fins i tot poden ser millors indicadors del futur d'una llengua que el nombre actual de parlants que la dominen.

TECNOLOGIES PER A LA REVITALITZACIÓ LINGÜÍSTICA

Cal que els pobles indígenes liderin el desenvolupament de la següent onada de tecnologies lingüístiques responsives i responsables... Per sort, els pobles indígenes de tot el món exerceixen un lideratge tecnològic a mesura que guanyen terreny digital per a les seves llengües (Brinklow, 2021).

Les llengües indígenes parlades a Canadà són **heterogènies** en quant a propietats lingüístiques (10 famílies lingüístiques inconnexes) i en el nombre de parlants per llengua. Els activistes de llengües indígenes a Canadà són molt conscients del notable èxit dels seus iguals a Aotearoa —és a dir, Nova Zelanda, on el maori ha estat llengua oficial des de l'any 1987 (King, 2013)— i a Hawaii (Asensio, 2019). Tanmateix, la revitalització lingüística a Canadà és un repte encara més complex. A Canadà, cada llengua té necessitats diferents, des de les de llengües en perill greu —on la prioritat acostuma a ser documentar un nombre petit d'Ancians que parlen la llengua amb fluïdesa— fins a les de l'inuktitut, una llengua governamental a Nunavut que es parla extensament, on la prioritat és fomentar el seu ús més ampli (p. ex., en entorns mèdics) i millorar l'educació en aquesta llengua.

Dos articles recents (Brinklow et al. 2019, Brinklow 2021) parlen de tecnologies lingüístiques per revitalitzar i recuperar llengües indígenes. Citant al lingüista Francis Tyers, Brinklow senyala que, donat que les grans empreses de programari tenen ànsies de beneficis, les tecnologies comercials se centren en les necessitats "d'un petit nombre de rics, o d'un gran nombre de pobres." Les llengües indígenes a Canadà són parlades per un petit nombre (segons els estàndards mundials) de persones relativament pobres —no els importen gaire a les grans empreses. En la mesura que les empreses sí que estan interessades en les llengües indígenes, sovint volen conservar l'IP de dades lingüístiques i, per tant, amenacen la sobirania de les dades indígenes —Keegan (2019) dona un exemple d'Aotearoa que il·lustra aquest perill.

També existeixen problemes tècnics amb tecnologies lingüístiques comercials:

- Cap de les "grans" llengües (anglès, xinès, japonès, àrab, etc.) per les quals s'han desenvolupat aquestes tecnologies són polisintètiques. Les tecnologies estàndard aplicades a llengües polisintètiques s'enfronten a dificultats —per exemple, el problema de "fora de vocabulari" (OOV, per les seves sigles en anglès, *out of vocabulary*). Fins i tot amb un vocabulari per al reconeixement automàtic de la parla (RAP) d'1,3 milions de paraules inuktitut, Gupta i Boulianne (2020a) van trobar que més del 60% de paraules d'històries inuktitut retingudes **no** estaven incloses en aquell vocabulari (vegeu a continuació). A un experiment comparable realitzat amb texts anglesos, seria d'esperar que aquesta taxa "fora de vocabulari" fos molt més baixa que el 60% —pràcticament segur menys del 5%, possiblement per sota de l'1%.
- Les llengües indígenes a Canadà no compten amb les quantitats massives de dades d'entrenament de les quals depèn l'aprenentatge automàtic. Els membres de l'equip del NRC-ILT estem acostumats a preguntes de persones no indígenes amb pocs coneixements tècnics com: "Perquè no creeu programari per traduir entre el cree (o una altra llengua indígena) i l'anglès?"

La nostra resposta consta de dues vessants:

1. Apart dels inuit, cap altra comunitat indígena amb les quals hem parlat han mostrat gaire interès en la traducció automàtica (TA) entre la seva llengua ancestral i l'anglès (o el francès, al Quebec). Les comunitats solen estar més interessades en eines que fomentin l'aprenentatge i l'ús de la seva llengua ancestral.
2. La majoria de les llengües indígenes a Canadà tenen poques dades paral·leles —sovint es limiten a alguns llibres de la Bíblia traduïts per missioners. L'existència de dialectes a la majoria d'aquestes llengües agreuja aquest problema d'escassetat de dades. (Per la llengua cree, el nostre equip i els nostres col·laboradors indígenes

potser podrien arreplegar com a molt 100.000 frases en cree de la plana en paral·lel amb l'anglès; això no seria suficient per entrenar un sistema de TA).

Ens trobem amb malentesos sobre aquesta situació fins i tot entre experts en processament del llenguatge natural. Alguns experts han insinuat que estem endarrerits tècnicament perquè utilitzem plantejaments basats en regles a la majoria de les nostres tasques. Al cap i a la fi, d'altres equips han aconseguit aplicar exitosament l'aprenentatge automàtic a llengües amb pocs recursos... El terme "llengua amb pocs recursos" pot dificultar la comprensió. S'aplica a llengües com el tàmil, amb 75 milions de parlants, la majoria d'ells alfabetitzats en la seva llengua, amb una història de texts escrits que es remunten milers d'anys. És ridícul emprar el mateix terme per descriure la llengua indígena "més gran" a Canadà, cree de la plana, amb 75.000 parlants i pocs texts escrits. La majoria dels subprojectes del NRC-ILT tracten llengües amb "recursos **extremadament** escassos".

L'inuktitut és una excepció a aquesta situació de "recursos extremadament escassos" (vegeu a continuació).

Littell, Kazantseva et al. (2018) resumeix les tecnologies per a la revitalització lingüística a Canadà. A la següent llista extreta d'aquest article, les tecnologies en **negreta** van ésser creades pel NRC-ILT amb col·laboradors indígenes; les tecnologies subratllades van ésser creades de forma externa al NRC amb l'ajuda de finançament del NRC-ILT:

- **Implementació de fonts i formats digitals per a llengües indígenes amb conjunts de caràcters poc comuns;**
- **Text predictiu;**
- **Conversió entre diferents ortografies per la mateixa llengua;**
- Correcció ortogràfica *(sense participació del NRC-ILT)*;
- **Generació de paradigmes per ensenyar morfologia complexa;**
- **Cerca aproximada;**
- **Traducció automàtica;**
- Tecnologies de veu:
  - Reconeixement automàtic de la parla;
  - Cerca de paraules clau a través de dades de veu;
  - **Alineació de veu a text**;
  - **Conversió de text a veu** *(tasca preliminar del NRC-ILT)*;
- Tecnologies d'imatge;
- Aprenentatge de llengües assistit per ordinador (CALL, per les seves sigles en anglès, *computer-aided language learning*).

Littell, Kazantseva, et al. (2018) al·ludeix breument a quatre altres activitats tecnològiques importants:

- Creació d'eines per documentar i anotar discursos d'Ancians que parlen la llengua amb fluïdesa;
- Digitalització de cintes magnètiques i d'àudio de discursos d'Ancians;
- Conservació d'arxius de discursos;
- Elaboració de diccionaris digitals.

LA HISTÒRIA I L'ENFOCAMENT COL·LABORADOR DEL PROJECTE

Trobareu un resum acadèmic del NRC-ILT a (Kuhn, Davis, Désilets, Joanis et al., 2020), i un informe tècnic detallat a https://nrc-publications.canada.ca/eng/view/object/?id=d4f10144-c711-43c5-b80b-5ace7df5e68b. Les actualitzacions periòdiques es troben a https://nrc.canada.ca/en/research-development/research-collaboration/programs/canadian-indigenous-languages-technology-project.

El projecte va començar amb el pressupost de març de l'any 2017 del Govern de Canadà, que proporcionava 6 milions de dòlars canadencs (CAD) al NRC per

Owennatékha (Brian Maracle): fundador i director, escola d'immersió Onkwawenna Kentyohwa.Va proposar **Kawennón:nis,** i va col·laborar en la seva creació.

Akwiratékha Martin: professor de kanyen'kéha a Kahnawake – col·labora amb l'equip de NRC-ILT per crear la versió oriental de **Kawennón:nis.**

Kawennón:nis - conjugador verbal de kanyen'kéha

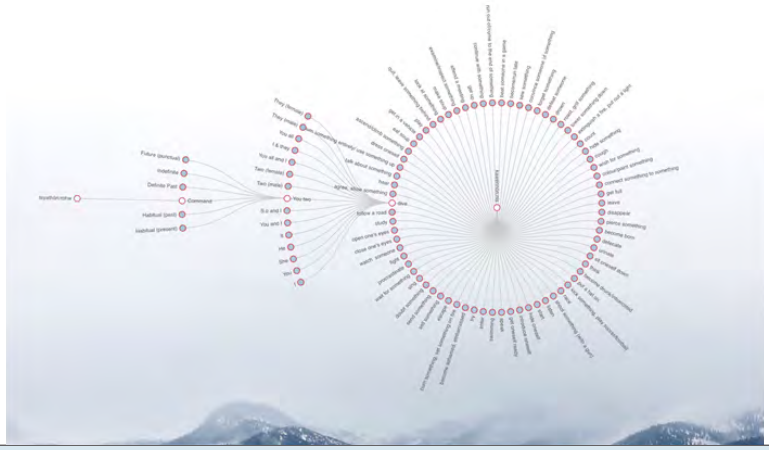


Figura 3. Col·laboracions en kanyen'kéha (mohawk).

desenvolupar, amb la col·laboració de parts interessades indígenes, tecnologies de programari per donar suport a llengües indígenes. Es va considerar que la nostra secció del NRC estava capacitada per fer aquesta tasca gràcies a la nostra experiència amb tecnologies per a grans llengües mundials —p. ex. havíem treballat en la TA del xinès i l'àrab a l'anglès, amb finançament dels programes GALE[4] i BOLT[5] de l'Agència d'Investigacions de Projectes Avançats de Defensa (DARPA, per les seves sigles en anglès, *Defense Advanced Research Projects Agency*) dels Estats Units. El NRC també havia dut a terme una iniciativa a petita escala centrada en l'inuktitut durant els anys 2003-2012 (Martin et al., 2003; Farley, 2012).

L'equip del NRC-ILT procura complir amb el seu mandat desenvolupant relacions estretes i respectuoses amb comunitats indígenes i intentant trencar amb la llarga i dolorosa història de pràctiques de recerca extractives (Keegan, 2019; Brinklow et al., 2019). Creiem en la filosofia de "l'empoderament", on la recerca es fa de manera col·laborativa, amb el mateix èmfasi en les intencions de l'investigador i la comunitat (Czaykowska-Higgins, 2009).

Això ha significat preguntar als activistes de llengües indígenes quines eines de programari trobarien **ells** útils, enlloc d'oferir tecnologies basant-se en temes de recerca interessants. D'aquí la col·lecció

4.- Global Autonomous Language Exploitation.
5.- Broad Operational Language Translation.

una mica inconnexa de tecnologies que es descriuen a continuació: hem donat resposta a les diferents necessitats de diferents comunitats. Ens va guiar un comitè consultiu format per experts en la revitalització de llengües indígenes. El seu consell ha estat de valor incalculable. El NRC-ILT no s'ha declarat mai propietari de les dades de llengües indígenes recollides amb el finançament del projecte. Quan ha estat possible, hem implementat un model de "forn i ceràmica": el programari de codi obert (OSS per les seves sigles en anglès, *open-source software*) independent de la llengua (el forn) crea programari específic per la llengua (la ceràmica) que serà propietat de la comunitat.

El finançament inicial va ésser més generós del que esperàvem, així que vam finançar tasques dins i fora del NRC. El NRC-ILT té tres "cercles":

- Un cercle intern de feina duta a terme al NRC amb la col·laboració de parts interessades indígenes.
- Un cercle mitjà de feina relacionada amb el reconeixement automàtic de la parla (RAP) duta a terme pel Centre de Recerca Informàtica de Montreal (CRIM per les seves sigles en francès, *Centre de Recherche Informatique de Montréal*). El CRIM té una il·lustre trajectòria de recerca pionera en RAP i tecnologies relacionades. El CRIM va entregar resultats experimentals de RAP per a dues llengües indígenes importants, i també l'OSS que accelera una fase clau en l'anotació de discursos gravats.
- Un cercle extern d'altres subprojectes curosament seleccionats per l'equip del NRC-ILT i el seu comitè consultiu indígena per rebre finançament. La majoria d'aquests subprojectes estaven gestionats per organitzacions indígenes; tots ells comptaven amb una plantilla gairebé completament indígena.

CERCLE INTERN: TECNOLOGIES DESENVOLUPADES DINS DEL NRC

CONJUGADORS VERBALS

El primer subprojecte del NRC-ILT va sorgir d'una proposta d'Owennatékha (Brian Maracle), el director de l'escola Onkwawenna Kentyohkwa (Our Language Society) a Six Nations of the Grand River al sud-oest d'Ontàrio: https://onkwawenna.info/. Aquesta escola és coneguda per produir parlants que dominen el kanyen'kéha (mohawk). Owennatékha va proposar a l'equip del NRC-ILT de crear Kawennón:nis (the WordMaker) —una eina que ajuda a estudiants a dominar el complex sistema verbal del kanyen'kéha. Fins i tot pels temes verbals més comuns en kanyen'kéha existeixen tantes possibles conjugacions verbals que una eina de referència física seria impossiblement gran; no obstant això, és factible computar aquestes conjugacions mitjançant un programari (Kazantseva et al, 2018). Kawennón:nis, implementat pel dialecte occidental del kanyen'kéha, es va sotmetre a extenses proves d'usuari amb estudiants i professors a l'escola abans del seu llançament. L'han rebut amb entusiasme. Està dissenyat per complementar el pla d'estudis, i utilitza les metàfores visuals que empren els professors de l'escola: https://kawennonnis.ca/wordmaker.

Posteriorment, el NRC-ILT va elaborar sistemes genèrics per crear conjugadors verbals per a llengües polisintètiques. S'estan aplicant per crear conjugadors verbals pel kanyen'kéha oriental parlat a Kahnawake (Quebec), i per dues llengües no relacionades al kanyen'kéha: La llengua algonquina anishinaabemowin parlada a Kitigan Zibi (a prop de Maniwaki, Quebec), i el michif.

Figura 4. READALONG STUDIO – UN ÈXIT INESPERAT!

La **Figura 3** mostra dos dels nostres col·laboradors mohawk. Tots dos ensenyen kanyen'kéha: Owennatékha ensenya el dialecte occidental, i Akwiratékha el dialecte oriental. (D'altres educadors mohawk també han ajudat a crear i provar el Kawennón:nis). La figura també mostra la interfície d'usuari de Kawennón:nis.

Heather Souter del Prairies to Woodlands Indigenous Language Revitalization Circle actualment anima als estudiants dels seus cursos a provar la versió beta del conjugador verbal de michif. Davies, Santos, i Souter (2021) descriuen obres relacionades.

Tenim previst llançar conjugadors verbals per a vàries altres llengües. A més, el Dr. Patrick Littell dirigeix una iniciativa per desenvolupar un sistema fàcil d'utilitzar basat en fulls de càlcul anomenat "Gramble" que podria facilitar la feina d'activistes lingüístics per crear conjugadors **de forma independent** (sense els coneixements del NRC).

READALONG STUDIO

Aquest és "l'èxit inesperat" del NRC-ILT: no anticipàvem la quantitat de professors que estarien interessats en afegir una funcionalitat clau a audiollibres i vídeos preexistents en les seves llengües. Aquesta funcionalitat ha estat promoguda per l'equip de la Prof. Marie-Odile Junker a la Universitat de Carleton.

La funcionalitat s'il·lustra a la **Figura 4**; és molt senzilla. A mesura que s'escolta un audiollibre o es veu un vídeo amb discurs indígena (aquí, atikamekw), la paraula parlada apareix destacada al text acompanyant. A la figura, s'està escoltant la paraula "atisokasotcik", i per això apareix destacada. Si l'oient —un estudiant o professor— vol centrar-se en la pronunciació d'una paraula dins del text, fa clic a sobre la paraula, i n'escoltarà la pronúncia; també pot alentir la reproducció. Nosaltres (els equips de la Universitat de Carleton i del NRC-ILT)

anomenem els audiollibres o vídeos amb aquesta funcionalitat “ReadAlongs” (llibres o vídeos amb lectura gravada).

Moltes comunitats tenen llibres o vídeos educatius amb discursos en les seves llengües, amb una transcripció en text de les paraules parlades (o cantades). Per poder afegir la lectura gravada a aquests llibres o vídeos, cal que les paraules del discurs s’alineïn amb les paraules del text. L’equip de Carleton estava fent aquestes alineacions a mà quan vam començar a col·laborar junts. La nostra contribució va ésser automatitzar aquest procés, utilitzant un programari que vam crear anomenat “ReadAlong Studio”. Delasie Torkornoo de l’equip de Carleton ens ha estat ajudant a millorar el codi de ReadAlong Studio. Ambdós equips han produït i enviat ReadAlongs a educadors indígenes en les llengües algonquines, atikamekw, cree sud-oriental, cree nord-oriental, gitksan, inuktitut, kwak’wala, kanyen’kéha (mohawk), seneca i SENĆOŦEN. Cada quinze dies rebem noves sol·licituds per convertir materials educatius d’aquesta manera. Els educadors indígenes ens comenten que els ReadAlongs tenen un alt valor pedagògic.

TECLATS, CONVERSIÓ ORTOGRÀFICA, I TEXT PREDICTIU

El NRC-ILT ha llançat un OSS que implementa teclats per a alguns sistemes d’escriptura mal atesos, i que converteix entre sistemes d’escriptura per a algunes llengües. Eddie Antonio Santos de l’equip va treballar amb les llengües saulteaux, makah i cree de la plana. Juntament amb el lingüista Arok Wolvengrey de la Universitat d’Alberta, també va persuadir a Google per canviar el seu teclat sil·làbic pel cree de la plana als Chromebooks (la versió anterior era adequada pel cree oriental, però no pel cree occidental).

El NRC-ILT també ha llançat codi per a la predicció de text per a dispositius mòbils. En teoria, el programari pot implementar la predicció de text per a qualsevol llengua. No obstant això, és difícil d’utilitzar per a usuaris no experts. Fins ara, només hem implementat la predicció de text per a la llengua SENĆOŦEN. Tanmateix, donat que l’ortografia del SENĆOŦEN és difícil, els membres d’aquesta comunitat ens han comentat que estan molt contents amb aquesta funcionalitat —fa que sigui molt més fàcil introduir text als seus dispositius.

INUKTUT

Tota la feina feta a aquesta secció ha estat gràcies a la col·laboració amb el Pirurvik Centre, un centre cultural i lingüístic gestionat pels inuit (https://www.pirurvik.ca/).

Les llengües o dialectes relacionades i parlades pels inuit a Nunavut s’anomenen col·lectivament “inuktut”. La majoria de la feina del NRC-ILT amb l’inuktut ha estat relacionada amb la versió inuktitut, parlada a Baffin Island. Aquesta feina difereix del nostre treball amb altres llengües de dues maneres: 1. No vam crear eines educatives, sinó **eines ofimàtiques que ajuden a escriure o llegir text inuktut** (o traduir entre l’inuktut i l’anglès); 2. Vam treballar en la **traducció automàtica (TA)** entre l’inuktut i l’anglès, mitjançant l’aprenentatge automàtic. Aquestes diferències existeixen perquè hi ha una burocràcia i un òrgan legislatiu que funcionen parcialment en inuktut: aquestes organitzacions necessiten texts escrits en inuktut, i proporcionen una quantitat creixent de texts bilingües inuktut-anglès que es poden utilitzar per entrenar sistemes de TA.

Una iniciativa a petita escala al NRC durant els anys 2003-2012 per crear eines per l’inuktut va donar resultats valuosos —el més important, un analitzador morfològic.

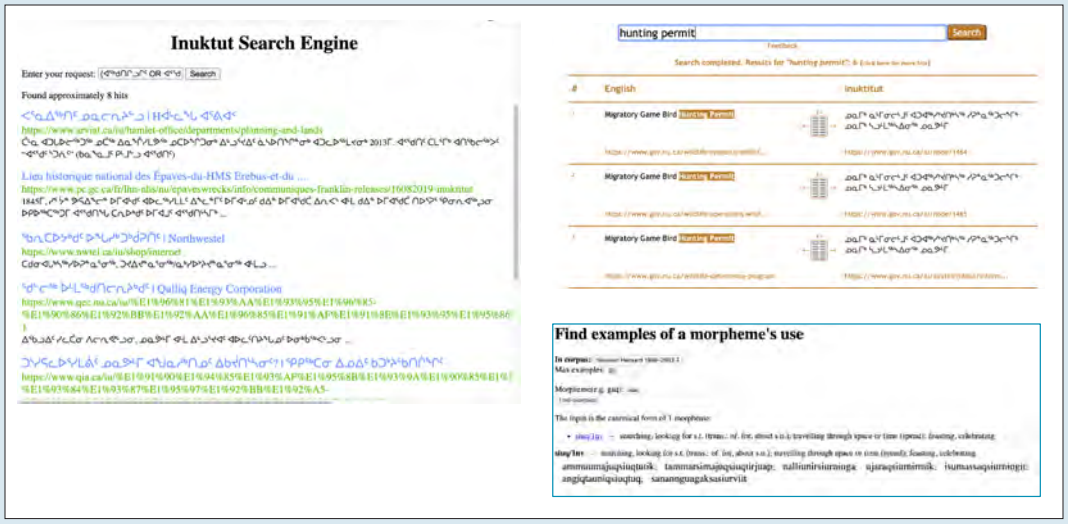


Figura 5. Cercador d'inuktut, WeBInuk (programa de concordança), i exemple de cerca per morfema

Un col·laborador clau va ser Benoît Farley, que tot i haver-se retirat del NRC manté un repositori per a aquestes eines (Farley, 2012).

Una altra producció del NRC durant aquest període va ser la publicació consecutiva de corpus paral·lels de texts anglès-inuktut basat en les actes de l'Assemblea Legislativa de Nunavut, el "Nunavut Hansard" (d'ara endavant, "NH"). El primer d'aquests, el NH 1.0, es va cedir a la comunitat d'investigadors l'any 2003 (Martin et al., 2003); cobria actes de 155 dies de l'Assemblea de Nunavut, i constava de 3.432.212 símbols anglesos i 1.586.423 símbols inuktitut.

Quan el NRC-ILT va reprendre el seu treball amb l'inuktut l'any 2017, es disposava d'actes de molts més anys de NH sense alinear. Vam **crear i publicar un nou corpus alineat**, el NH 3.0. Consta de 8.068.977 símbols inuktitut i 17.330.271 símbols anglesos; comprèn les actes de 687 dies de debat des de l'1 d'abril de 1999 fins el 8 de juny de 2017 (Joanis et al., 2020).

La dimensió del NH 3.0 fa que sigui factible entrenar sistemes de **traducció automàtica** (TA). Vam ajudar als organitzadors del WMT 2020, una avaluació internacional anual de sistemes de TA, a establir avaluacions per a TA entre l'anglès i l'inuktitut en ambdues direccions. Juntament amb Microsoft, també vam invertir en avaluadors humans que dominaven l'inuktitut perquè valoressin les traduccions de dades de prova cap a l'inuktitut de part de sistemes de la competència. La naturalesa extremadament polisintètica de l'inuktitut, insòlit d'entre les llengües del món, va atraure molts equips d'investigació a aquest repte de TA.

Vam crear els nostres propis sistemes de TA que van competir a aquesta avaluació (Knowles et al., 2020). Els resultats globals es troben aquí: http://www.statmt.org/wmt20/translation-task.html. Esperem que es mantingui l'interès per la recerca en TA en aquest parell lingüístic i que finalment resulti en eines pràctiques per l'inuktitut.

Seguim millorant les eines ofimàtiques per l'inuktitut que el NRC va llançar els anys 2003-2012, i seguim creant-ne de noves:

- Estem millorant l'analitzador morfològic de Benoît Farley.
- Estem recuperant el projecte WeBInuk, una eina del NRC basada en la web que cerca

traduccions inuktitut de paraules angleses. Mostra frases en anglès i inuktitut on la frase anglesa conté la paraula clau cercada. Hi havia traductors inuktut que utilitzaven molt WeBInuk els anys 2007-2017, però la direcció del NRC tenia altres prioritats i es va permetre que desaparegués.

- Estem creant un lematitzador que a l'importar una paraula inuktitut exporta la paraula més comuna amb la mateixa arrel. Aquest és un pas clau per donar suport a la cerca monolingüe en inuktitut, o per cercar en la direcció inuktitut —> anglès. És molt més difícil crear un cercador per a usuaris que introdueixen paraules de cerca en inuktitut que en anglès o francès, degut al seu caràcter polisintètic. (És probable que qualsevol paraula inuktitut que introdueixi l'usuari no hagi existit mai en la història de la llengua, i segurament no tornarà a existir mai més). Cal fer cerques difuses. La **Figura 5** il·lustra els avenços que hem fet amb aquesta i d'altres eines relacionades.

CONVERSIÓ PRELIMINAR DE TEXT A VEU (TTS PER LES SEVES SIGLES EN ANGLÈS, *TEXT TO SPEECH*)

El pla d'estudis a Onkwawenna Kentyohkwa se centra en dominar la complexa morfologia verbal del kanyen'kéha. Per aquest motiu, el NRC-ILT i l'escola van desenvolupar conjuntament Kawennón:nis (descrit anteriorment).

Kawennón:nis té una limitació: els estudiants poden generar paraules complexes en format text, però potser no saben com pronunciar-les. L'objectiu de l'escola és produir *parlants* que dominin la llengua, no només *lectors/escriptors*, així que no concorda el que produeix Kawennón:nis (text) amb el que més necessiten els estudiants (parla per imitar-la). Tot i això, de la mateixa manera que no seria factible escriure i publicar Kawennón:nis en format llibre degut al seu enorme vocabulari implícit, seria igual de poc factible **gravar** totes les formes. Això suggereix la implementació de TTS; podem entrenar un sistema de TTS sobre un subconjunt gravat i equilibrat fonèticament del vocabulari?

Vam crear un prototip de TTS concatenatiu. Reordena unitats de subparaules de 852 gravacions de conjugacions de parlants que dominen la llengua per cobrir completament les primeres 122.966 conjugacions a Kawennón:nis. Els professors de l'escola van declarar que la qualitat del discurs de sortida era més que acceptable. Recentment, també hem desenvolupat un model de TSS neuronal capaç de produir parla sintetitzada de qualitat acceptable a partir de tan sols 30 minuts d'àudio. Owennatékha va declarar fa poc que el TTS és una de les seves tres prioritats tecnològiques per al kanyen'kéha.

Kawennón:nis il·lustra els reptes als quals s'enfronten els elaboradors de plans d'estudis indígenes: llengües complexes, pocs parlants per ensenyar la llengua, pocs materials gravats, i pocs parlants per gravar nous materials i, per tant, grans parts de la llengua als quals els estudiants no poden accedir ni escoltar. El TTS permet que organitzacions amb pocs recursos puguin aprofitar al màxim el seu recurs més preciós i limitat —el temps dels parlants que dominen la llengua— al fer ús d'un petit nombre de gravacions per cobrir un domini molt més ampli.

CERCLE MITJÀ: TECNOLOGIES DE VEU DESENVOLUPADES PEL CRIM

L'anotació i la transcripció són la càrrega més gran a la majoria de projectes de documentació i conservació lingüística. Cox et al. (2019) parlen del "coll d'ampolla de la transcripció": s'estan gravant discursos (especialment d'Ancians) en llengües

indígenes de forma més ràpida del que s'estan transcrivint, o fins i tot anotant (s'hauria d'anomenar "el coll d'ampolla de l'anotació"). Això passa a llengües minoritàries arreu del món. L'anotació i especialment la transcripció porten temps, i són tasques tedioses. Resulta temptador gravar moltes hores de discursos i planejar dur a terme aquests passos tediosos "més endavant", però sovint es posposen indefinidament...

Això passa des de fa dècades, molt abans de que es disposés de mitjans de gravació digital. Per tot Canadà, existeixen milers d'hores de discursos en llengües indígenes a cintes magnètiques o cassets d'àudio que són recursos potencialment valuosos per a la revitalització lingüística, però que han esdevingut inaccessibles a la pràctica perquè ningú sap què contenen. Estan repartides per comunitats i departaments universitaris (he sentit històries de cintes magnètiques amb gravacions importants que s'han trobat sense etiquetar a les prestatgeries del garatge d'un professor difunt). Ningú té temps d'escoltar la majoria d'aquest tipus de gravacions; en alguns casos tràgics, no queda ningú que entengui la llengua gravada. Sovint les gravacions més antigues no s'han copiat a mitjans digitals.

RECERCA PRINCIPAL

Vam fundar el Centre de Recerca Informàtica de Montreal (CRIM per les seves sigles en francès, *Centre de Recherche Informatique de Montréal*) per fer recerca sobre el reconeixement automàtic de la parla (RAP) per un petit nombre de llengües indígenes parlades a Canadà. Un RAP amb vocabulari complet per aquestes llengües no és realista a curt termini, perquè no existeixen les grans quantitats de dades de transcripció de veu paral·leles necessàries per entrenar els sistemes de RAP. No disposem de suficient finançament per crear aquestes dades. Per tant, l'objectiu era crear sistemes de RAP imperfectes entrenats amb quantitats petites de dades paral·leles. S'ha demostrat que aquests tipus de sistemes han fet que la "cerca d'àudio" (també anomenada "reconeixement de paraules clau" o "detecció de termes parlats") sigui factible per a d'altres llengües. Quan l'exactitud del RAP és baixa, encara és possible que els usuaris trobin certes paraules o frases en discursos mitjançant tècniques de cerca d'àudio.

Els experiments de RAP al CRIM s'han centrat en l'inuktitut i el cree oriental. Els investigadors del CRIM van gestionar la transcripció de gairebé 81 hores de discurs inuktitut, i 102 hores de discurs cree oriental; van dividir aquestes dades paral·leles en conjunts de prova i d'entrenament. La seva feina posterior en el RAP de inuktitut i cree oriental es descriu a (Gupta i Boulianne, 2020a) i (Gupta i Boulianne, 2020b), respectivament. La feina va ser possible gràcies a la col·laboració amb el Pirurvik Centre (https://www.pirurvik.ca/) i la Canadian Broadcasting Corporation (https://www.cbc.ca/).

Una qüestió clau per l'equip de CRIM era quines unitats utilitzar pel RAP. L'inuktitut és molt polisintètic; per tant, la llengua té una taxa molt alta de "fora de vocabulari" (60%) — moltes paraules només s'expressen un cop. Un sistema de RAP amb un vocabulari format per paraules no reconeixeria bé la majoria de les paraules a dades de veu inuktitut noves. Els investigadors del CRIM van provar-ho amb morfemes, síl·labes i fonemes com a unitat bàsica de RAP, i van obtenir els millors resultats amb síl·labes. Però fins i tot amb unitats sil·làbiques, al voltant d'un 71% de les paraules inuktitut utilitzades a les dades de prova no es van reconèixer correctament.

No obstant això, a obres no publicades, aquests investigadors van aconseguir un rendiment de reconeixement de paraules clau

relativament bo per l'inuktitut (*comunicació personal*). L'enfocament amb el millor rendiment era convencional: cercar a través d'una matriu de fonemes generada per RAP basat en síl·labes, amb una matriu de confusió ponderada que aporta robustesa als errors de RAP.

Als experiments del CRIM, el cree oriental té una taxa de "fora de vocabulari" significativament més baix que l'inuktitut, tot i que es disposava de quantitats molt més petites de dades textuals per crear un vocabulari de cree oriental. Amb un vocabulari de 30.000 paraules obtingut de textos en dos gèneres —transcripcions de vídeos i traduccions de la Bíblia—, la taxa de "fora de vocabulari" per a dades de text retingudes era del 25% pels vídeos i del 9% per les dades bíbliques. Aquestes xifres van fer que les paraules fossin una unitat factible pel RAP en cree oriental. A experiments de RAP sobre dades de veu retingudes de cree oriental, el 70% de les paraules als vídeos de prova no es van reconèixer correctament, així com el 25% de les paraules a les dades bíbliques de prova. No es van realitzar experiments de reconeixement de paraules clau pel cree oriental.

Els investigadors del CRIM van obtenir resultats experimentals molt prometedors (taxes baixes de reconeixement erroni) tant per l'inuktitut com el cree oriental quan el sistema de RAP **depenia** del parlant. Els pobres resultats anteriorment citats s**ó**n **independents** del parlant. No obstant, per a la revitalitzaci**ó** de lleng**ü**es ind**í**genes, on sovint es recullen moltes hores de discurs d'un petit nombre d'Ancians que parlen la llengua amb flu**ï**desa, un sistema de RAP que **depèn** del parlant seria **ú**til a nivell pr**à**ctic. Considerem una manera de treballar on (p.ex.) un expert hum**à** transcriu un parell d'hores del discurs d'un Anci**à**, com es fa actualment. Llavors, aquestes dues hores transcrites s'utilitzarien per entrenar un sistema de RAP adaptat al discurs d'aquest Anci**à** en particular, la qual cosa podria produir un esborrany de la transcripci**ó** de les nombroses hores restants de discurs de la mateixa persona. Aix**ò** alleujaria el "coll d'ampolla de la transcripci**ó**".

EINES DE PRODUCTIVITAT

L'equip del CRIM també va publicar eines que faciliten les fases inicials del processament de discursos gravats. Aquestes eines estaven empaquetades com a serveis web a la plataforma VESTA del CRIM (http://vesta.crim.ca) i estan disponibles allà o a través d'una extensió ELAN —ELAN és una eina d'anotació a https://archive.mpi.nl/tla/elan/ (Wittenburg et al., 2006). Les eines permeten la segmentació d'arxius de veu, en veu i no veu (silenci, soroll, música, etc.), la recuperació lingüística (trobar segments parlats en una llengua concreta, podent-se identificar 32 llengües), la recuperació per parlant (trobar segments parlats per un parlant concret), la detecció d'activitat de veu multicanal (detectar segments que contenen veu de forma separada per a cada pista d'una gravació multicanal amb més d'un micròfon), i d'altres capacitats útils.

Els Drs. Chris Cox i Olivia Sammons (de la Universitat de Carleton i la Universitat First Nations, respectivament), van fer servir aquestes eines CRIM recentment mentre recollien dades de veu michif. Informen d'una acceleració quadruplicada o quintuplicada per a les fases inicials d'anotació (*comunicació personal*).

CERCLE EXTERN: ALTRES TASQUES FINANÇADES PEL NRC-ILT

Aquesta secció defineix els subprojectes de revitalització de llengües indígenes a les

quals ha col·laborat el finançament del NRC-ILT, almenys parcialment, però que han estat gestionats des de fora del NRC.

CURSOS I JOCS EN LÍNIA

S'han d'actualitzar periòdicament les eines d'aprenentatge lingüístic en línia perquè el programari subjacent ha esdevingut obsolet. Per exemple, Adobe va deixar d'oferir suport pel Flash Player a principis de l'any 2021. El NRC-ILT va finançar mesures correctives per a diversos subprojectes.

Plataformes educatives a la Universitat de Carleton:
La Prof. Marie-Odile Junker i el seu equip a la Universitat de Carleton porten anys col·laborant amb companys indígenes per desenvolupar lliçons en línia per a llengües algonquines. Utilitza un sistema d'acció participativa per treballar amb les comunitats. Les contribucions del seu equip són massa nombroses per enumerar-les aquí —vegeu https://www.marieodilejunker.ca/. La Prof. Junker va rebre el *Governor General's Innovation Award* (2017).

El finançament del NRC-ILT va començar a ajudar el projecte ja existent a Carleton l'any 2018. Els canvis de programari havien deixat parcialment penjades les eines educatives lingüístiques a la plataforma web de Carleton per l'innu (no confondre amb l'inuktut, una llengua independent) i el cree oriental. El finançament va ajudar a pagar una actualització tecnològica (eliminació de la dependència de Flash i d'altres tecnologies obsoletes) i l'experiència de dos experts en llengües indígenes que van contribuir amb grans quantitats de continguts pedagògics. L'equip de la Prof. Junker segueix desenvolupant la plataforma amb altres fonts de finançament (innu: https://lessons.innu-aimun.ca/; cree oriental: https://lessons.eastcree.org/).

7000 Languages:
Algunes comunitats indígenes volien desenvolupar cursos en línia per a les seves llengües. Vam proporcionar un finançament modest a 7000 Languages, una organització sense ànim de lucre que crea programari educatiu lingüístic, per crear cursos juntament amb aquestes comunitats (a les quals també vam finançar). Tres d'aquests cursos es troben en línia a la web de 7000 Languages, www.7000.org: per al kwak'wala, el michif, i el mi'kmaq.

Aprenentatge de llengües assistit per ordinador (CALL) a la Universitat d'Alberta:
Aquest subprojecte dona suport al dialecte Y del cree (cree de la plana). L'equip de la Universitat d'Alberta, juntament amb membres de comunitats que parlen cree, ha estat creant un sistema CALL adaptatiu anomenat "CreeTutor". També s'han creat continguts en la llengua cree: s'han gravat, transcrit i traduït 13 històries personals de 8 Ancians. CreeTutor es publicarà aviat. Un element clau de CreeTutor és *Sound Hunters* [caçadors de sons], un exercici de consciència fonèmica receptiva (Lothian et al., 2020). La principal creadora de *Sound Hunters,* Delaney Lothian (cree-métis), és una estudiant de llicenciatura en ciències a la Universitat d'Alberta i membre a temps parcial del NRC-ILT. Està treballant en la versió michif de *Sound Hunters*.

FirstVoices:
El First Peoples' Cultural Council (FPCC) a la Columbia Britànica té una sòlida trajectòria proporcionant tecnologies d'última generació, formació i suport tècnic a activistes de llengües indígenes (principalment però no exclusivament a la província), dins del programa FirstVoices (http://www.fpcc.ca/about-us/). Language Tutor [tutor lingüístic], que permet a les comunitats crear lliçons de llengua, forma

part de FirstVoices i s'havia quedat "obsolet" a nivell tecnològic. El finançament del NRC-ILT va ajudar a reconduir la situació.

On the Path of the Elders:
"On the Path of the Elders" [Al camí dels Ancians] és un joc de rol dissenyat per donar a conèixer temes històrics i culturals relacionats amb la regió de James Bay. Quan aquest joc en línia gratuït es va llançar l'any 2007, va rebre molts elogis pel seu ús innovador de recursos històrics. Va perdre funcionalitats degut a canvis en la indústria del programari subjacent. El NRC-ILT va proporcionar finançament per actualitzar el programari i afegir continguts en cree swampy. La pàgina web de "On the Path of the Elders" ara és compatible amb mòbils i amb els navegadors Chrome, Firefox i Internet Explorer: www.pathoftheelders.com.

TALLERS DE DESENVOLUPAMENT D'HABILITATS

Dos subprojectes finançats pel NRC-ILT van proporcionar formació per a persones que documenten llengües indígenes. El NRC-ILT també va proporcionar un finançament modest a l'activista lingüística Caroline Running Wolf; va organitzar quatre exitoses sessions en línia per a activistes lingüístics, els "Community Workshops for Indigenous Language Technology" (CWILTs, tallers comunitaris de tecnologies per a llengües indígenes) del desembre de 2020 al juliol de 2021.

Yukon Native Language Centre (YNLC)[6]:
A Yukon es parlen vuit llengües indígenes (gwich'in, hän, kaska, tutchone septentrional, tutchone meridional, tagish, tlingit, i alt tanana): gairebé el 15% de les llengües indígenes a Canadà. Amb l'ajuda del finançament del NRC-ILT, el YNLC va donar suport a 12 estudiants en un programa de 10 mesos durant el qual van adquirir habilitats pràctiques per elaborar, difondre i conservar materials digitals per aquestes llengües a les seves comunitats. La pandèmia del COVID-19 va interrompre algunes de les activitats planejades a aquest subprojecte, que estava centrat en activitats presencials, però la gestió competent del personal del YNLC va fer que fos un èxit igualment.

- Al Taller 1 (octubre de 2019), els estudiants van practicar com utilitzar equips de vídeo i com fer entrevistes amb parlants que dominen la llengua. També van rebre formació en gestió d'arxius.
- Al Taller 2 (novembre de 2019), van rebre formació sobre l'ús dels paquets de programari d'ELAN i SayMore per transcriure, anotar i traduir gravacions de veu.
- El Taller 3 (febrer de 2020) va començar amb una revisió i posada en pràctica de les habilitats apreses fins al moment. També va incloure compartir i reutilitzar vídeos, i utilitzar materials ELAN com a recurs per a l'aprenentatge i l'ensenyament de llengües. L'últim dia, els estudiants van mostrar els seus vídeos acabats.

Els 12 estudiants van crear 548 minuts de documentació i van arribar a dominar habilitats que els permetran gravar als Ancians dins de les seves comunitats de forma continuada.

Indigitization[7]:
Aquesta és una col·laboració entre la Universitat de la Columbia Britànica, Musqueam Archives, i el centre d'educació

6.- Centre de llengües natives de Yukon.
7.- El nom del projecte és un joc de paraules que combina els conceptes de *indigenization* (indigenització) i *digitization* (digitalització).

cultural Heiltsuk Cultural Education Centre que se centra en la digitalització de dades de veu. Mentre que el subprojecte YNLC formava a membres de les comunitats indígenes per recollir dades **noves**, el subprojecte continuat Indigitization forma a persones per convertir dades **antigues** a formats digitals accessibles. Ajuda a que les comunitats indígenes guanyin autonomia al familiaritzar-los amb les opcions de programari i maquinari per a la digitalització. Per a comunitats que ja tenen continguts digitalitzats, el subprojecte ofereix formació de transcripció de discursos a llengües indígenes, la traducció de transcripcions a l'anglès i al francès, i la gestió dels arxius resultants.

El projecte Indigitization va començar al gener de 2020, i es va haver de reorganitzar degut a la crisis del COVID-19 (els tallers presencials van esdevenir poc aconsellables). Es va dedicar més recursos a la creació de recursos instructius que al pla original: "Tenim una oportunitat única per redactar guies pràctiques basades en l'experiència que ajudaran a organitzacions indígenes petites, sovint poc finançades, a començar a gestionar les seves col·leccions de forma estructurada" (*comunicació personal amb Gerry Lawson, cap del projecte Indigitization*).

S'han elaborat els continguts de text de 65 d'aquestes noves guies; actualment s'està realitzant l'edició i el disseny gràfic. Aquest subprojecte pensa prestar maquinari a les comunitats quan sigui necessari.

CWILTs:

Els "Community Workshops for Indigenous Language Technology" (CWILTs, tallers comunitaris de tecnologies per a llengües indígenes) són una sèrie de sessions en vídeo per a activistes de llengües indígenes organitzat per Caroline Running Wolf de Apsáalooke Nation (crow), cadascuna de les quals inclou xerrades ràpides centrades en el tema de la revitalització lingüística, seguit per converses en grups reduïts. El seu objectiu és crear una comunitat de pràctica. Molts participants als CWILTs han mantingut el contacte entre ells —s'ha avançat cap a aquest objectiu.

- Sessió 1 (7 de desembre de 2020): 26 participants. Running Wolf i Sara Child (comunitat kwakwaka'wakw) van presentar els CWILTs; Delaney Lothian (cree, métis) va presentar *Sound Hunters*, el seu joc per a estudiants per reconèixer fonemes i síl·labes cree (esmentat anteriorment); Eddie Antonio Santos va presentar el programa de conversió sil·làbica de cree occidental (esmentat anteriorment).
- Sessió 2 (7 de gener de 2021): 33 participants. Stephanie Tenasco (Kitigan Zibi Anishinabe) va compartir la seva experiència creant vídeos d'aprenentatge lingüístic; Patrick Littell i Aidan Pine (NRC) van presentar ReadAlong Studio.
- Sessió 3 (28 de gener de 2021): 80 participants (la majoria eren de Canadà, però n'hi havia 13 d'EUA, 2 de Noruega, 1 de Mèxic, 1 de Bangladesh, i 2 del Marroc). Tema: *"Strengthening Language through Technology: A Context for Capacity Building"* [Enfortir la llengua mitjançant la tecnologia: un context per al desenvolupament d'habilitats]. Aaron Plahn i Justin Bambrick van presentar la seva iniciativa de col·lecció de corpus i desenvolupament d'habilitats per al tŝilhqot'in:
- https://www.youtube.com/watch?v=L-eXWFHnCSc&feature=youtu.be.
- Sessió 4 (14 de juliol de 2021): no es va gravar el nombre de participants (almenys 40). Sean Smith, Krista Dempster, i Dustin Hill de YNLC van descriure el desenvolupament d'habilitats al seu subprojecte.

SUBPROJECTES MONOLINGÜES

La digitalització de text cree a la Blue Quills University:
La Blue Quills va ésser la primera universitat a Canadà propietat de i dirigida per pobles de Primeres Nacions. Aquest subprojecte va digitalitzar i indexar la col·lecció més gran de textos sil·làbics cree: aproximadament 40.000 pàgines de butlletins mensuals, anomenats kihcitwāw miteh (Sacred Heart [sagrat cor]) produïts per l'església catòlica entre els anys 1906 i 1978. El corpus estarà disponible aviat a http://language-unbq.bluequills.ca.

Col·lecció de dades kanyen'kéha a TTO:
Aquesta tasca està liderada per Nathan Thanyehténhas Brinklow, un professor de llengua a la Queen's University a Ontàrio, amb la col·laboració del centre cultural i lingüístic Tsi Tyónnheht Onkwawén:na Language and Cultural Centre (TTO) al territori Tyendinaga Mohawk. L'objectiu és la creació de corpus de text i àudio per donar suport a la recerca de RAP per al kanyen'kéha. Aquest subprojecte ha recollit 112.420 paraules escrites en kanyen'kéha i 26 hores d'àudio, i inclou gèneres com guions i àudio de pel·lícules i sèries de televisió, llibres traduïts de la Bíblia, i gravacions d'Ancians.

Col·lecció de dades kwak'wala:
Aquesta iniciativa també implica la col·lecció de corpus, seguit de recerca de RAP. La llengua kwak'wala (família wakash) es parla a 18 nacions kwakwaka'wakw, el territori tradicional dels quals es troba al nord de la illa de Vancouver, a illes properes i als territoris continentals adjacents. El finançament del NRC-ILT va donar suport a una col·laboració entre dos equips comunitaris de tres comunitats kwakwaka'wakw diferents, un equip tècnic, i tres lingüistes universitaris (dos dels quals són kwakwaka'wakw). Aquest subprojecte ha acumulat 25 hores de dades d'àudio kwak'wala llegibles a través de l'ordinador, que consisteixen en converses, materials pedagògics, i narracions d'Ancians. També ha identificat més de cent hores d'altres gravacions de veu kwak'wala a diverses ubicacions.

Diccionari oral michif:
Aquest subprojecte va mobilitzar i fer accessible un diccionari michif descatalogat anomenat "The Michif Dictionary: Turtle Mountain Dictionary Chippewa Cree", publicat per primer cop l'any 1983. Finançat parcialment pel NRC-ILT, i amb l'assistència de parlants nadius de michif, lingüistes computacionals, i d'altres, l'equip va desenvolupar una versió digital parlada d'aquest recurs. El subprojecte va produir més de 181 hores de gravacions d'àudio d'alta qualitat del diccionari de la mà de quatre parlants; el diccionari inclou 15.422 entrades. Almenys un parlant ha gravat totes les 350 pàgines d'entrades lèxiques michif amb frases d'exemple, i algunes entrades les han gravat dos o més parlants. L'enllaç al diccionari: https://dictionary.michif.org/home.

L'informe final d'aquest subprojecte fa referència a les noves tecnologies CRIM descrites anteriorment:

Cada gravació es va segmentar automàticament en expressions delimitades per pauses mitjançant un servei de detecció d'activitat de veu a través d'una xarxa neuronal profunda (DNN per les seves sigles en anglès, *deep neural network*) desenvolupat dins del projecte VESTA-ELAN pel CRIM. Aquesta auto-segmentació va estalviar una quantitat de temps incalculable al procés d'anotació (*comunicació de Heather Souter i Olivia Sammons, coordinadores del subprojecte*).

Nsyilxcn:
La llengua okanagan (nsyilxcn, família salish interior) era parlada tradicionalment a la vall d'Okanagan en l'actual Columbia Britànica.

És una llengua en perill crític, amb només una dotzena d'Ancians que la parlen amb fluïdesa. Aquest subprojecte va començar al desembre de 2019; el va dur a terme Syilx Language House (SLH), una organització comunitària, amb finançament parcial del NRC-ILT. Es van gravar, transcriure (amb un glossari en anglès) i arxivar set hores de narracions d'Ancians que dominen l'okanagan. Totes les gravacions es troben a www.thelanguagehouse.ca. Dins del subprojecte, es va formar a 4 persones en període de pràctiques i 10 estudiants voluntaris per adquirir fluïdesa.

SENĆOŦEN:

La llengua saanich (SENĆOŦEN, família salish de la costa) es parlava al nord de l'actual Victoria, Columbia Britànica. És la llengua en perill més greu amb la qual ha treballat el NRC-ILT: fins fa poc, només quedaven cinc persones que parlessin la llengua amb fluïdesa. No obstant això, la comunitat està compromesa amb una iniciativa vigorosa de revitalització lingüística, liderada pel WSÁNEĆ School Board. Durant els anys 1981-1991, el lingüista Dr. Timothy Montler va gravar l'Elsie Claxton, la última parlant monolingüe de SENĆOŦEN, narrant històries (Montler, 2018). El NRC-ILT va pagar a dos Ancians perquè treballessin amb el Dr. Montler per transcriure aquestes gravacions, i traduir-les a l'anglès. Aquest material serà un recurs de valor incalculable per ensenyar la llengua.

Tŝilhqot'in:

La llengua tŝilhqot'in (família atapascana o dené) es parlava tradicionalment a la zona interior meridional de l'actual Columbia Britànica. El finançament per part del NRC-ILT va permetre al Govern Nacional de Tŝilhqot'in (TNG per les seves sigles en anglès, *Tŝilhqot'in National Government*) ampliar una iniciativa de revitalització lingüística impressionant. El TNG ja havia gravat 35.000 fragments d'àudio, elaborat una base de dades lingüístiques, publicat paradigmes verbals a un "llibre verbal digital", i creat una pàgina web amb diverses eines educatives: www.tsilhqotinlanguage.ca. Amb el nostre finançament, el TNG va gravar una dotzena d'hores de discursos de 20 parlants; es van digitalitzar, transcriure i alinear les transcripcions de 20 hores de discursos. Es va completar l'etiquetatge de 46.000 fragments d'àudio. S'ha elaborat una llista de 26.200 paraules i frases angleses a traduir; moltes d'aquestes s'han traduït al tŝilhqot'in. Es va formar a quatre membres de la comunitat en rols informàtics.

Es pot trobar un excel·lent vídeo dels esforços de revitalització lingüística d'aquesta comunitat als CWILTs a https://www.youtube.com/watch?v=L-eXWFHnCSc&feature=youtu.be.

FEINA FUTURA I CONCLUSIONS

Brinklow (2021) declara que la col·laboració productiva amb col·laboradors no indígenes en tecnologies per a llengües indígenes (ILT) és possible:

"Moltes comunitats a Canadà estan col·laborant amb cooperants sense ànim de lucre per desenvolupar ILT fora d'un ecosistema lucratiu... Aquestes col·laboracions innovadores inclouen universitats, comunitats, institucions de recerca, governs, i d'altres... Al context canadenc, aquestes col·laboracions (irònicament, sovint finançades pels mateixos governs colonitzadors) estan produint tecnologies lingüístiques pioneres que són innovadores des de qualsevol punt de vista."

Aquest autor indígena té elogis per un subprojecte del NRC-ILT: "Ha sorgit una

col·laboració pionera entre Onkwawenna Kentyohkhwa a Six Nations, a Ontàrio, i el NRC. La col·laboració és entre l'equip del projecte NRC-ILT i professors locals per desenvolupar un conjugador verbal pel kanyen'kéha... Tot i que el desenvolupament inicial es va fer amb una llengua concreta (mohawk), l'eina subjacent es va crear per treballar amb qualsevol llengua amb una perspectiva 'primer profunda i després àmplia' al disseny i al desenvolupament... La col·laboració amb el NRC aporta valor al projecte a nivell nacional perquè un enfocament ampli no és responsabilitat d'una comunitat lingüística individual."

Quins són els següents passos?

Les nostres prioritats són:

- Ampliar les tecnologies que ja hem desenvolupat a llengües noves.
- Facilitar l'ús d'aquestes tecnologies. No és suficient llançar programari de codi obert (OSS); la perspectiva "d'empoderament" significa que aquest OSS hauria de ser fàcil d'utilitzar per a persones dins de comunitats indígenes. No hauria de requerir un màster en informàtica o lingüística computacional. Un exemple dels nostres esforços actuals en aquesta direcció: feina dins del NRC-ILT liderat pel Dr. Patrick Littell sobre "Gramble", un sistema de programari basat en fulls de càlcul que, esperem, facilitarà la creació de conjugadors de verbs.
- Un nou enfocament del NRC-ILT serà la conversió de text a veu (TTS). És preferible que els estudiants aprenguin una llengua a partir de parlants humans que la dominin. Tanmateix, donada l'escassetat de professors per a la majoria de llengües indígenes, la TTS podria a vegades esdevenir la segona millor opció acceptable. Un membre de l'equip del NRC-ILT, Aidan Pine, ja va crear un sistema de TTS prototip pel kanyen'kéha (mencionat anteriorment); recentment, ha creat sistemes de TTS experimentals per a dues altres llengües indígenes parlades a Canadà, el gitksan i el SENĆOŦEN, que van obtenir una qualitat de veu sorprenentment bona.

Les perspectives de futur per a moltes llengües indígenes parlades a Canadà són més bones que en les últimes dècades, gràcies als esforços dels mateixos pobles indígenes. El projecte NRC-ILT mostra que és possible que organitzacions no indígenes, fins i tot governamentals, desenvolupin tecnologies en col·laboració amb les parts interessades indígenes que resultin útils per aquestes iniciatives de revitalització lingüística. Tanmateix, cal que les organitzacions no indígenes que hi actuen recordin que el seu paper és secundari, i que el seu objectiu hauria de ser empoderar les comunitats indígenes.

REFERÈNCIES

Rubèn Fernández Asensio (editor). 2019. "*Kava* vell en carbasses noves: revitalització lingüística i escolarització a Hawaii". *Linguapax Review*. https://www.linguapax.org/wp-content/uploads/2020/02/linguapax19-1-1.pdf

Peter Bakker. 1997. "A language of our own". Oxford & New York: Oxford University Press.

Nathan Thanyehténhas Brinklow, Patrick Littell, Delaney Lothian, Aidan Pine, and Heather Souter. 2019. "Indigenous Language Technologies and Language Reclamation in Canada". *Collection of research papers of the 1st international conference on language technologies for all*. European Language Resources Association (ELRA). https://lt4all.elra.info/proceedings/lt4all2019/pdf/2019.lt4all-1.100.pdf.

Nathan Thanyehténhas Brinklow. 2021. "Indigenous Language Technologies: Anti-

Colonial Oases in a Colonizing (Digital) World". *WINHEC: International Journal of Indigenous Education Scholarship,* pp. 239-266; Special Issue on Indigenous Language Revitalization. Innovation, Reflection and Future Directions. http://dx.doi.org/10.18357/wj1202120288.

Michael J. Chandler and Christopher Lalonde. 1998. "Cultural continuity as a hedge against suicide in Canada's First Nations". *Transcultural Psychiatry*, 35(4):191–219.

Christopher Cox, Gilles Boulianne, and Jahangir Alam. 2019. "Taking aim at the transcription bottleneck: Integrating speech technology into language documentation and conservation". [Presentació de conferència]. *International Conference on Language Documentation & Conservation*. https://scholarspace.manoa.hawaii.edu/bitstream/10125/44841/44841.pdf

Ewa Czaykowska-Higgins. 2009. "Research models, community engagement, and linguistic fieldwork: Reflections on working within Canadian Indigenous communities". *Language documentation & conservation*, 3(1):182–215.

Fineen Davies, Eddie A. Santos, and Heather Souter. 2021. "On the Computational Modeling of Michif Verbal Morphology". *EACL*. https://aclanthology.org/2021.eacl-main.226.pdf.

Petra Fachinger. 2019. "Colonial violence in sixties scoop narratives: from In Search of April Raintree to A Matter of Conscience". *Studies in American Indian Literatures*, 31(1-2):115.

Benoît Farley. 2012. "The Uqailaut project". [Lloc web]. http://www.inuktitutcomputing.ca.

First Voices Portal. 2021. "SENĆOŦEN Home Page". Descarregat de https://www.firstvoices.com/explore/FV/sections/Data/THE%20SEN%C4%86O%C5%A6EN%20LANGUAGESEN%C4%86O%C5%A6EN/SEN%C4%86O%C5%A6EN l'1 d'octubre del 2021.

Government of Canada. 2015. "Final report of the Truth and Reconciliation Commission". http://nctr.ca/reports.php.

Vishwa Gupta and Gilles Boulianne. 2020a. "Automatic transcription challenges for Inuktitut, a low-resource polysynthetic language". *Proceedings of LREC*. https://aclanthology.org/2020.lrec-1.307.pdf.

Vishwa Gupta and Gilles Boulianne. 2020b. "Speech transcription challenges for resource constrained Indigenous language Cree". *Proceedings of 1st Joint SLTU-CCURL Workshop*. https://aclanthology.org/2020.sltu-1.51/

Darcy Hallett, Michael J. Chandler, and Christopher E. Lalonde. 2007. "Aboriginal language knowledge and youth suicide". *Cognitive Development*, 22(3):392–399.

Eric Joanis, Rebecca Knowles, Roland Kuhn, Samuel Larkin et al. 2020. "The Nunavut Hansard Inuktitut–English parallel corpus 3.0 with preliminary machine translation results". *Proceedings of LREC-2020*. https://aclanthology.org/2020.lrec-1.312.pdf.

Anna Kazantseva, Owennatékha Brian Maracle, Ronkwe'tiyohstha Josiah Maracle, and Aidan Pine. 2018. "Kawennón:nis: the wordmaker for Kanyen'kéha". *COLING Workshop on Computational Modeling of Polysynthetic Languages*. https://aclanthology.org/W18-4806.pdf

Te Taka Keegan. 2019. Issues with Māori sovereignty over Māori language data. [Presentació vídeo]. http://video.web.gov.bc.ca/public/fpcc/letlanguageslive.html.

Jeannette King. 2013. "Te Kōhanga Reo: Māori Language Revitalization". L. Hinton and K. Hale (editors), *The Green Book of Language Revitalization in Practice*. Brill.

Rebecca Knowles, Darlene Stewart, Samuel Larkin, and Patrick Littell. 2020. "NRC

Systems for the 2020 Inuktitut-English News Translation Task". *Proceedings of WMT*. https://aclanthology.org/2020.wmt-1.13/.

Roland Kuhn, Fineen Davis, Alain Désilets, Eric Joanis et al. 2020. "The Indigenous Languages Technology project at NRC Canada: An empowerment-oriented approach to developing language software". *Proceedings of COLING 2020*. https://www.aclweb.org/anthology/2020.coling-main.516.pdf.

Patrick Littell, Anna Kazantseva, Roland Kuhn, Aidan Pine, Antti Arppe, Christopher Cox, and Marie-Odile Junker. 2018. "Indigenous language technologies in Canada: Assessment, challenges, and successes". *27th International Conference on Computational Linguistics (COLING 2018)*, Santa Fe, New Mexico, USA. https://www.aclweb.org/anthology/C18-1222.pdf

Delaney Lothian, Gokce Akcayir, Anaka Sparrow, Owen Mcleod, and Carrie Demmans Epp. 2020. "SoundHunters: Increasing Learner Phonological Awareness in Plains Cree". *Artificial Intelligence in Education*. doi: 10.1007/978-3-030-52237-7_28.

Joel Martin, Howard Johnson, Benoît Farley, and Anna Maclachlan. 2003. "Aligning and using an English-Inuktitut parallel corpus". *Proceedings of the HLT-NAACL 2003 Workshop on Building and using parallel texts*, Volume 3, pages 115–118. Association for Computational Linguistics.

Timothy Montler. 2018. "SENĆOŦEN: a dictionary of the Saanich language". University of Washington Press, ISBN: 9780295743851.

Nicole Rosen and Heather Souter. 2009. "Language revitalization in a multilingual community: the case of Michif." *1st International Conference on Language Documentation and Conservation (ICLDC)*, Honolulu.

Statistics Canada. 2017. "Proportion of mother tongue responses for various regions in Canada, 2016 census". https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/dv-vd/lang/index-eng.cfm.

Olivia N. Sammons. 2019. "Nominal classification in Michif". *Ph.D. thesis, University of Alberta*. https://era.library.ualberta.ca/items/903b7ee2-4f3f-4959-b4f0-8231205d5d67/view/9a967b05-b223-4716-97de-6da32e6931fe/Sammons_Olivia_N_201901_PhD.pdf

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. "ELAN: A professional framework for multimodality research". *European Language Resources Association (ELRA)*.

OPENSPEAKS: TRANSFORMING LEARNING OF TECH INNOVATIONS IN LOW-RESOURCE LANGUAGE DOCUMENTATIONS TO OPEN EDUCATIONAL RESOURCES

Subhashish Panigrahi
subhashish@theofdn.org

Subhashish is a researcher and film-maker who works in the intersections of social justice, language documentation and technology. In the past he has led community-catalysing roles across Asia-Pacific at not-for-profits including Wikimedia Foundation, Centre for Internet and Society, Mozilla and Internet Society. As a 2018 National Geographic Explorer, he has served many endangered language communities by documenting their languages. Subhashish is the founder of the OpenSpeaks project and co-founded the O Foundation.

ABSTRACT

The OpenSpeaks project is a toolkit for archivists who are creating audio-visual documentations of low or limited resource-languages. The project is available online at https://en.wikiversity.org/wiki/OpenSpeaks/. In this article, I examine some of the technological innovations that are in use for the protection of endangered, indigenous and other under-resourced languages, and detail how such innovations are evaluated for creating Open Educational Resources for archivists. Among other observations, I emphasize that language revival and archival might seem like technological processes from a surface-level view, but they are tied directly to the social, economic and political hierarchy and hence should always be addressed using social justice viewpoints.

1. INTRODUCTION

1.1. ENDANGERED LANGUAGES

The gradual decline in use of a language by native speakers, especially the young ones, lead to language endangerment. To outline the premise of language endangerment, UNESCO explains, “A language is in danger when its speakers cease to use it, use it in an increasingly reduced number of communicative domains, and cease to pass it on from one generation to the next. That is, there are no new speakers, adults or children.” (UNESCO Ad Hoc Expert Group on Endangered Languages, 2003). Linguists have created different frameworks and methodologies to identify the endangerment of languages. As linguist Michael Krauss had noted, the rate of language extinction is extraordinarily high whereas the survival of languages is very poor. The name “endangered language” is framed after endangered biological species and is often understood the same way such a species goes extinct. Using this conceptualisation, Krauss also created the distinction between endangered languages and the “moribund languages”, the latter includes languages that are no longer spoken by children of a speaker community. (Hale et al., 1992) The UNESCO Atlas of the World’s Languages in Danger has a comprehensive list of 2,464 languages with a varied degree of endangerment: from “vulnerable” (spoken by most children in closed environments such as home) to “extinct” (no living speakers left), through “definitely endangered” (not spoken by children), “severely endangered”

(spoken only by the oldest generations but parents might not speak with children), and "critically endangered" (sparsely spoken by semi-speakers who may be speaking sparsely) (Moseley, 2010). Educational resources in OpenSpeaks have been enriched through multiple language documentation projects: documentation of the Baleswari dialect of Odia (a poorly documented dialect of a rather dominant language) in 2015, prior to the creation of OpenSpeaks; and a 2018 archival project with support from the National Geographic Society to document the moribund Kusunda language (Panigrahi, 2019/2019) of western Nepal and the indigenous languages Bonda (Rajpal, 2018) and Ho (Panigrahi, 2019/2019) in the eastern Indian state of Odisha.

1.2 ROLE OF TECHNOLOGY IN PROTECTING, DOCUMENTING AND REVITALIZING LANGUAGES

Languages in general, and endangered and other low-resource languages particularly, are impacted in a range of ways as technology affects societies around the world. There is a significant increase in terms of access to digital technology, especially because of the pervasiveness of smartphones and decreasing cost of the mobile data. However, while internet penetration is growing at a remarkably steeping rate across the world, with 51% people already having access to the internet by 2019 (International Telecommunication Union, 2021), only 14% of the total population in low-income countries could afford to use the internet in 2017 (World Bank Group, 2017). Internet affordability is further affected across the ethnicity and gender lines. The Web Foundation estimated that over a decade, low and low-middle income countries have lost about US$1 trillion in Gross Domestic Product (GDP) as women could not access and participate online because of a number of barriers. Additionally, the "digital gender gap" has dropped from 30.9% to 30.4% since 2011 (Pulgarín & Woodhouse, 2021). As of October 2021, there are active Wikipedias in 312 languages (4.8% of the 6,500 languages as recognised by UNESCO) from around the world (Wikimedia contributors / Wikimedia Foundation, n.d.). As Wikipedia is a written encyclopaedia containing citations to published resources and most world languages do not have a writing system of their own without knowledge published in peer-reviewed journals and other notable publications, Wikipedia also becomes unsuitable in most cases for the under-resourced languages discussed in this article.

1.3 SOCIO-ECONOMIC AND POLITICAL CONTEXTS

Throughout this project, content creators, journalists, scholars, field linguists/documentarians and many other practitioners who create and share linguistic content in their focus languages are broadly referred to as archivists, a self explainatory term "multimedia language archivist" (an archivist making audio-visual and other multimedia documentation of languages) that is shortened in places plainly to "archivist", and the activism relating to the documentation of languages as "language archivism" (portmanteau of "activism" and "archiving"). Similarly, the reference to low and limited resources means the financial, human, socio-political, institutional/infrastructural, technical and all other essential areas of resource that many languages lack, which in turn impact their survival, use or further development. Availability of funding through public policies, people, public and private institutional infrastructures, technical know-how, active participation of the civil society organisations are some of the most critical resources for the survival of languages. A language speaker community's political and economic power limits the community's access to resources

for the active inter-generational use and development of their language. There is often a direct impact of colonial and postcolonial exploitation of many marginalised communities through political, environmental and socioeconomic means that are carried forward by the neighbouring majoritarian communities. Such exploitation not only impact on the cultural disintegration and assimiliation of many marginalised groups in the mainstream societies while still being oppressed by the dominant groups (such as indigenous peoples throughout the world) but also highly restricted access to the resources for the growth of languages.

2. DESIGN OF OPENSPEAKS

Supporting archivists who are documenting endangered, indigenous, minority and other low or limited-resource languages has been the key focus area for the OpenSpeaks project since its inception in 2015 and launch in 2017. Each of these language categories has its own sets of challenges and specific needs. Albeit the definition of each category varies from linguist to linguist, the lack of resources remains a critical barrier for documentation for most/all the languages in these categories. The status quo of availability (or dearth) of resources for each language has to be critically examined to identify which specific resources or what kind of workflow are relevant for a specific documentation project.

2.1 DESIGN PRINCIPLES AND SCOPE

The design principles of the OpenSpeaks project lie broadly in documenting the methodologies, best practices and other learning from the archival works of archivists who are documenting indigenous, endangered and other under-resourced languages, as audio and video, and creating educational materials that beginner and intermediate-level archivists can use.

The modules and tutorials inside OpenSpeaks display know-how of specific areas for audio-visual documentation of languages, but also frameworks to help each archivist to position their language of focus so that they can figure out for themselves which of the resources are relevant in their context. The entire base content of this project is available publicly in simple English to help archivists with basic English fluency access easily and help translators localise the project into other world languages.

2.2 OPENSPEAKS AS AN OPEN EDUCATIONAL RESOURCE

OpenSpeaks as a resource is intended to be used for a beginner-to-intermediate-level multimedia language-archivist, it is assumed in most cases that the archivist might or might not have access to many critical resources. The project uses a self-assessment (a series of questions) that the archivist can take to decide for themselves.

Time: The following questions are to define the time constraint/availability of both the archivist and the persons they are interviewing. The self-assessment questions prompts the archivist asking “how much time I can contribute for the documentation?” and “how much time my interviewees can contribute to the documentation?”. The second question is quite dependent on the “funds” factor to be discussed later in this section.

Hardware: Hardware/equipment, software and the operational knowledge of the same can be clubbed as technical resources. A large part of the operational segment of the language documentation being a technical process, some amount of prior technical planning is needed. Having a smartphone that can record and has a storage for at least one-two hour(s) of audio/video is the bare minimum for documentation. A few key ques-

Resource	Current status	Support needed	How to procure?
1. Time	5-6 hours/week for 2 months	More time during post-production	Need to find a video editor on a *pro-bono* basis "OR" carve out time for video editing
2. Technical resources	Smartphone for both audio and video recording and a small table-top tripod for fixed recording	A lavalier microphone can enhance interviewee voice quality	I do not have budget to purchase a mic but can borrow from a friend if I reach out to her in advance
3. Editing	I currently do not have any editing software or skills	Need to learn Kdenlive for desktop video editing and Audacity for desktop audio editing	Need to bookmark useful tutorials on Youtube and watch them to learn using these editing tools
4. Funds	I can manage to cover the local travel expenses	Need about $300 to pay a video editor for 15 editing hours; $40 for a microphone; and $30 for local travel	a. I will apply for a small grant to my university department so that I can purchase a mic, remunerate Maya (friend who knows editing) b. Alternatively, I will try to borrow a mic from Kai (friend) and discuss with Maya if she is available for helping with video editing c. If I get less than what I am applying, I will borrow the mic from Kai and pay partially/fully to Maya

Table 1: OpenSpeaks project planning framework

tions in this section could be—"what are the different equipment/hardwarel would need to to be able to document and which ones I already have?" or "do I have an audio recorder or camera or any other accessory (e.g. tripod and a mount for fixing the phone/camera to the tripod, or an external microphone that their phone/camera supports)?" and "which of these hardware I still need but I do not have access to, and how can I get them and learn how to use them?".

Editing: After recording audio/video, editing and processing it before publishing is needed. Sometimes time constraints might delimit an archivist's ability for post-production. If a video is live-streamed on a social media or a platform like YouTube, then there is not much scope for editing. The questions any multimedia language archivist can ask themselves are: "do I have an application on my phone/computer to edit the audio/video (unless they are live-streamed)?", "which apps do I have operational experience with and which are the new apps I need to learn to use?". Some live-streaming processes have options and even hardware set-up for real-time audio/video editing but that is beyond the scope of this article.

Funds: Access to funds to pay for the labour of the archivist and their support team, purchasing/renting equipment, or remunerating the interviewees for their time are some of the finance-related factors. It is important to do the entire assessment in the order shared here. To identify their project-level financial status, the archivist can ask themselves questions such as "how much money will I have to spend on the way I plan to document?". Increasing 10-20% of that amount is generally recommended as costs might increase in

> reality and the work should not get halted because of this. What can also be useful is if the archivist asks themselves—"where am I going to get the money that I estimate for this project?". This question can help them find ways to tap into funding streams unless they already have arranged for them.

The example below is based on a framework for an objective self-assessment and is filled up for the persona of aspecific archivist—a university student who is planning to go to a nearby town and support with documentation of a few low-resource languages.

2.3 LICENSING

As the host platform Wikiversity is available under an open licence, the Creative Commons Attribution-ShareAlike (CC BY-SA) License version 3.0, OpenSpeaks is also licensed under the same CC BY-SA 3.0 License. The reference to "open" in this context means that the project is available under Open Access—international frameworks like the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities among others detail how scientific and humanities content can be made available under open licenses (*Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities*, 2003). Educational Resources with Open Access, as opposed to restricted via content paywalls, are generally referred to as Open Educational Resources or OER in short. OpenSpeaks is also open for public contributions just like other Wikiversity and Wikimedia projects. To summarise, these design principles both equip Open Access and encourage open participation/contribution by everyone.

2.4 INTEGRATING LEARNING FROM FIELD DOCUMENTATIONS AND USER TESTING

Under the aegis of the OpenSpeaks project, several documentary films such as "Gyani Maiya" (2019; archival of Kusunda, moribund language from Nepal), "Remosam" (2019; archival of Bonda, endangered language from India) and "Mage Porob" (2020; archival of Ho, endangered language from India) have been created. The overall learning from these films captures the relations between different native speaker communities and film-makers like myself who are non-native speakers and researchers who are either native speakers or self-taught non-native speakers of the respective languages. In the documentation journey of many low-resourced languages, both initial or even long-term and predominant involvement of non-native speakers can potentially lead to a lower native-community ownership of the documented content and the narratives. In theory, the ethical ideals of language documentation of OpenSpeaks gives the utmost priority to the native speaker community and its ownership of the linguistic content. The challenge, however, is to identify language documentation processes that are led by archivists who are also native speakers, and shape OpenSpeaks toolkit around their challenges, learning and recommendations. Such archivists are not only fewer in number but also have the constraint of time, financial resources and linguistic barriers. Additionally, capturing learning and best practices might not be a priority in their work. Capturing the experiences of some such archivists has only been possible in a handful of cases by conducting 10 user interviews and user testing of OpenSpeaks prototypes in a 2021 incubator project called UNLOCK (Wikimedia Deutschland, 2021).

2.5 COGNIZANCE OF SELF PRIVILEGES

Just like the positioning of each language is unique in the language documentation spectrum, the challenges, opportunities and privileges of each archivist also has to be understood with their socio-economic backgrounds. For instance, my personal ability

to volunteer time to document learning on OpenSpeaks is deeply tied to my own privileges, both historical and systemic privileges such as the discriminatory caste system of India. The caste system has been used as a religious shield by a certain section of the society (often referred to as «upper caste» widely and «oppressor caste» by scholar Suraj Yengde) (Yengde, 2018) to oppress the Dalits (formerly and pejoratively known as "untouchables" who have been discriminated and criminalised for the longest time on the basis of ethnicity), Adivasi (indigenous) peoples, and Bahujan (other minority ethnic and otherwise marginalised groups). By and large, the Dalit-Adivasi-Bahujan people have been deprived of the socio-economic-political agency, and contemporary privileges including access to many forms of digital privileges (e.g. access to devices, access to the Internet, etc.) for the longest time in the Indian subcontinent and even beyond where the South Asian diaspora is present. Asking "what" is documented might be important from a linguistic standpoint to cover a wide range of topics, but "who" is documenting and "whose" voices/narratives they are documenting is very crucial from a class and gender perspective. In my opinion, the gender of the archivist is one of the most ignored but a very critical factor for "what" is documented. As a male documentary film-maker, I have often experienced first-hand the self-censoring by the female and non-binary interviewees, that arises from the gender-based discomfort, who might have shared important insights if they were interviewed by a same gender interviewer. Similar instances are also prevalent in the significant social class hierarchy between the interviewees and the archivists. OpenSpeaks details on identifying the gender and social class hierarchies and including ways in the planning process prior to the documentation to help address the aforementioned issues. More than technological innovations, recognition of the deeply embedded social biases and measures to address the systemic issues (and not just the topical ones) are extremely key to a long-term strategy building for language documentations. As access to formal training, funding and institutional support are scarce for most under-resourced languages, the educational resources in OpenSpeaks are being updated by optimising the project for self-paced learning and breaking the barrier of technical know-how.

3. LEARNING FROM OTHER SOCIO-TECHNICAL INNOVATIONS

3.1 NETWORKED COLLECTIVES

Finding existing practices of technical innovations and other forms of digital activisms for linguistic documentation/conservation is relatively easier than building a network of the key active stakeholders of language digital activism.As a large network that is connected to hundreds of individuals and organisations around the world that are involved in language digital activism for low-resourced languages, Rising Voices has been instrumental in facilitating knowledge and resource exchange practices. Such individuals and organisations exchange knowledge, advisory support and even collaborate with each other with and beyond the facilitation of Rising Voices. On the other hand, a collective like Rising Voices also creates a network effect by bringing together civil society actors, activists and even international organisations such as UNESCO. One such example is the Language Digital Activism Project (Rising Voices 2021) by Rising Voices that aims at providing a roadmap for language digital activism by furthering collaboration between current and new actors—the latter was possible through a series of focused workshops for activists. Continuous discussions of both shared and unique issues, and avenues for

sharing learning and resources that are of common interest strengthen the collaborations further.

3.2 INDIGENOUS COMMUNITY MEDIA

There are many forms of indigenous community media platforms, some are led by the communities themselves and some by local NGOs with participation from the community. Community radios by and large include local producers and local listeners. While the broadcasting method for community radios are generally FM-based, a pre-digital-era technology. Communities combine contemporary practices such as promotion and outreach through social media. "Radio Indígena 94.1 FM" in the Venture County of California serves the immigrant Mexican workers who speak mostly Mixtec or Zapotec dialects. Mixteco/Indígena Community Organizing Project (MICOP), the organisation behind this FM channel has managed to make policy-level changes by passing overtime laws for agricultural workers (Jiménez, 2020). Community radio organisations being causal agents for access to human rights are not uncommon. Even though language documentation or protection is not always a focus for many organisations, use of indigenous languages or any other under-represented language/dialect as a medium of outreach and other means of communications eventually lead to language documentation. In the Chhattisgarh state of India, the Central Gondwana Net Swara (CGNet Swara) works as a rural news portal that uses a technology as pervasive as phone calls for collecting recorded audio news from citizen journalists who are mostly members of the Gond indigenous tribe. Interactive voice recording (IVR), the technology behind CGNet Swara is a rather old but effective technology as it only requires one to have a landline or mobile connectivity (Pain, 2017). As different parts of India experience frequent internet shutdowns and slowdowns, India happens to be home to the largest number of internet shutdowns in the world (Keelery, 2021), relying on more basic technology that is more reliable in the given context has been effective for CGNet Swara. Voice-based journalism also cuts the barrier of literacy of dominant languages or even the internet/digital literacy. Many Gondi and Kurukh-language speakers (both endangered (vulnerable) languages) also contribute news in the respective languages which are further translated into Hindi. CGNet Swara articles being available in Hindi, the official language of Chhattisgarh and a dominant language in central and northern India, increases the chances of the news being more widely disseminated. More than 30 percent of Chhattisgarh's population includes different indigenous groups, collectively termed as the Adivasis in the socio-ethnic context and Scheduled Tribes (ST) as per the official classifications (Ministry of Tribal Affairs, Government of India, 2011) who have witnessed historical oppression through mining, India's oppressive caste system and significantly low-income jobs in addition to being the victims of constant political battle between the "Naxalite" movement and the government. The CGNet Swara model is a great example of the contextual framing of technology for a low-resource language to understand what works in a particular demographic and to take practical action for documenting human rights issues (Bali et al., 2019). Use of citizen journalism as a tool has also been effective in this case for documentation and digital use of the Gondi and Kurukh languages.

3.3 MASAKHANE: LOCAL COMMUNITY-LED LANGUAGE TECHNOLOGY INNOVATION

Language communities that might be linguistically divided can still find issues that are relevant to each other's languages. A technological solution for one language that usually

takes longer to create for the first time could become a template for other languages that share the same linguistic and/or other demographic factors. Identifying common grounds between languages with a shared history of colonisation and modern-day technological challenges has been the solidarity link for the Masakhane community (https://www.masakhane.io/). Masakhane is connecting speakers of 38 African languages from 30 different countries, and aims to find technological solutions in the Natural Language Processing (NLP) research. There are more than 2,000 languages that are spoken just within Africa. The long-lasting colonisation of many communities from Africa has caused severe damage by separating communities from their own languages by replacing native languages with dominant languages such as Arabic, English, French and Portugese. Since 2019, Masakhane's grassroots approach is resulting in building a solidarity network of more than 1,000 contributors from the region and innovating for the future of the African languages. Masakhane describes itself as an "open-source, continent-wide, distributed, online research effort for machine translation for African languages›› (Orife et al., 2020). The organisation's work includes building a language-tech community, and supporting them with necessary resources for conducting open, participatory and multidisciplinary research. By removing all kinds of eligibility criteria, Masakhane equips new individual contributors with tools and helps them by connecting with mentors and other contributors so that, together, they can find areas of interest, and collectively strengthen different open-source projects. Because of the large network of contributors and a focus on research, the organisation is able to identify and prioritise the areas for contribution. For instance, Machine translation, a method of automatically translating text in one language to another is not available in a large number of African languages and is creating a huge knowledge gap as the majority of the Internet's content is available only in dominant languages. Masakhane has built tools for community contribution and testing of neural machine translation (NMT) tools which are the first of its kind of an attempt to bridge the gap of accessing information through translation.

Masakhane's work underlines the value of building a contributor community for technological innovations because of its philosophical foundation—"Umuntu Ngumuntu Ngabantu" (isiZulu-language quote translating to "a person is a person through another person" or "I am because you are")—which finds a natural solution in open-source software. Masakhane indicates a shift from leaving the fate of local languages in the hands of large foreign corporations like Google or Microsoft to centering technological innovations around the socio-linguistic issues, and equipping local communities to lead those innovations.

4. CONCLUSION

The historical background for each under-resourced language provides a context on how the right to speak in one's own language is directly tied to their right to access information which, in turn translates to and the right to life, and how these rights are systematically violated by using dominant languages as a tool of oppression. The languages that are on the brink of extinction certainly need an immediate response and practical action in the form of documentation. However, the long-term goal has to be equipping native speakers with funding, educational and other resources so that they are in charge of building linguistic-technological solutions. Decolonisation of technology calls for reversing the historical equation of people with power through their privileges controlling the social,

political and economic power hierarchy to stay in power. In the context of endangered, indigenous and other low/limited-resource languages, the imminent danger to languages exist both from the outside and even within the native speaker community because of the far-reaching oppression. Instead of asking which technological solutions will address the issue of low usage of one such language, the framing of the problems needs an anthropological lens. For instance, asking "why a particular language is dominant over another and who is gaining economic and political power because of that structure?" can help find the actors in play. In most cases, the linguistic issues might warrant a technological solution at their outset. But, not knowing the social, economic and political reasons that lead to those issues could lead to creation of isolated technology, isolated from the people whom the solution could actually benefit. Many social issues are extremely complex and time-consuming to solve. It is highly unlikely that archivists, who are, in most cases, individuals with a limited access to human, financial and other resources, might not be able to address such larger issues. However, addressing the structural issues during creation of their solutions would stop the repetition of the same vicious cycle of oppressive social ways that are otherwise prevalent in technology. OpenSpeaks draws philosophical and practical inspirations from organisations like Masakhane which embodies the bottom-up approach of changing the tech society, global feminist campaigns like WhoseKnowledge? which is working towards "decolonisation of the Internet", or movement networks like Rising Voices which is creating an avenue for exchanging ideas and resources. The underlying theme of most such grounded works is approaching linguistic issues as social issues and using technology as a tool to drive innovations in social and behavioural practices. As OpenSpeaks design constantly absorbs the learning from other movement leaders, the design process and the educational content it has to offer to the language archivists is shifting towards reflecting the same learning. Instead of a prescriptive body of content, the new design is aimed at helping archivists view their language documentation work through the same lenses of social justice, and get trained on available resources that fit their needs.

5. REFERENCES

Bali, K., Choudhury, M., Sitaram, S., Seshadri, V., & Microsoft Research Labs, Bangalore, India. (2019). ELLORA: Enabling Low Resource Languages with Technology. *Proceedings of the Language Technologies for All (LT4All)*, 160–163. https://lt4all.elra.info/proceedings/lt4all2019/pdf/2019.lt4all-1.41.pdf

Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities. (2003). https://openaccess.mpg.de/Berlin-Declaration

Hale, K., Krauss, M., Watahomigie, L. J., Yamamot, A. Y., Craig, C., Jeanne, L. M., & England, N. C. (1992). Endangered Languages. *Language*, *68*(1), 42. https://doi.org/10.2307/416368

International Telecommunication Union. (2021). *ITU-D ICT Statistics: Statistics*. International Telecommunication Union. https://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx

Jiménez, C. (2020). Radio Indígena and Indigenous Mexican Farmworkers in Oxnard, California. In *Digital Activism, Community Media, and Sustainable Communication in Latin America* (p. 366). Springer International Publishing; Palgrave Macmillan. https://doi.org/10.1007/978-3-030-45394-7

Keelery, S. (2021, March 3). *Number of government-imposed internet shutdowns in India from 2012 to 2021*. statista. Recu-

perat el 17 d'octubre de 2021, de https://www.statista.com/statistics/1095035/india-number-of-internet-shutdowns/

Ministry of Tribal Affairs, Government of India. (2011). *State wise Scheduled Tribe population and decadal change by residence: Census 2011*. Tribal Profile. Recuperat el 17 d'octubre de 2021, de https://tribal.nic.in/ST/Tribal%20Profile.pdf

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

Orife, I., Kreutzer, J., Sibanda, B., Whitenack, D., Siminyu, K., Martinus, L., Ali, J. T., Abbott, J., Marivate, V., & Kabongo, S. (2020, March 13). Masakhane -- Machine Translation For Africa. *arXiv: Computation and Language*. https://arxiv.org/abs/2003.11529

Pain, P. (2017, August 4). Framing citizen activism: a comparative study of the CGNET Swara and Mobile Voices projects. *Media Asia*, *44*(2), 107-120. https://doi.org/10.1080/01296612.2016.1277825

Panigrahi, S. (Director). (2019). *Gyani Maiya* [ज्ञानी मैया] [Film]. Entertainment Identifier Registry. https://ui.eidr.org/view/content?id=10.5240/52AE-86BB-F84D-03B2-D938-U (Obra original publicada l'any 2019)

Panigrahi, S. (Director). (2019). *Mage Porob* [ ] [Film]. O Foundation. http://dx.doi.org/10.17613/t2cb-bg17 (Original work published 2019)

Pulgarín, A. M. R., & Woodhouse, T. (2021). *The Costs of Exclusion: Economic Consequences of the Digital Gender Gap*. Alliance for Affordable Internet, The Web Foundation. https://webfoundation.org/docs/2021/10/CoE-Report-English.pdf

Rajpal, S. (2018, August 25). O Foundation and National Geographic set out to document endangered languages before they are lost forever. *edex, The New Indian Express*. https://www.edexlive.com/happening/2018/aug/25/o-foundation-and-national-geographic-set-out-to-document-endangered-languages-before-they-are-lost-f-3728.html/

Rising Voices. (2021). *Language Digital Activism Toolkit*. Rising Voices. Recuperat el 18 d'octubre de 2021, de https://rising.globalvoices.org/languagetoolkit/about/

UNESCO Ad Hoc Expert Group on Endangered Languages. (2003). *Language Vitality and Endangerment*. UNESCO. https://ich.unesco.org/doc/src/00120-EN.pdf

Wikimedia contributors / Wikimedia Foundation. (n.d.). *Data: Wikipedia statistics*. Wikimedia Commons. Recuperat el 17 d'octubre de 2021, de https://commons.wikimedia.org/wiki/Data:Wikipedia_statistics/meta.tab

Wikimedia Deutschland. (2021). *OpenSpeaks Accessibility*. UNLOCK. Recuperat el 17 d'octubre de 2021, de https://www.wikimedia.de/unlock-projects/openspeaks-accessibility

World Bank Group. (2017). *Individuals using the Internet (% of population)*. World Bank Group. https://data.worldbank.org/indicator/IT.NET.USER.ZS

Yengde, S. (2018, June 3). Dalit Cinema. *South Asia: Journal of South Asian Studies*, *41*(3), 503-518. https://doi.org/10.1080/00856401.2018.1471848Hale, K., Krauss, M., Watahomigie, L. J., Yamamot, A. Y., Craig, C., Jeanne, L. M., & England, N. C. (1992). Endangered Languages. *Language*, *68*(1), 42. https://doi.org/10.2307/416368Moseley, C. (2010). *Atlas of the World's Languages in Danger*. UNESCO Atlas of the World›s Languages in Danger. http://www.unesco.org/culture/en/endangeredlanguages/atlasPanigrahi, S. (Director). (2019). *Gyani Maiya* [ज्ञानी मैया] [Film]. Entertainment Identifier Registry. https://ui.eidr.org/view/content?id=10.5240/52AE-86BB-F84D-03B2-D938-U (Original work published

2019)Rajpal, S. (2018, August 25). O Foundation and National Geographic set out to document endangered languages before they are lost forever. *edex, The New Indian Express*. https://www.edexlive.com/happening/2018/aug/25/o-foundation-and-national-geographic-set-out-to-document-endangered-languages-before-they-are-lost-f-3728.html/ Panigrahi, S. (Director). (2019). *Mage Porob* [ ] [Film]. O Foundation. http://dx.doi.org/10.17613/t2cb-bg17 (Original work published 2019)International Telecommunication Union. (2021). *ITU-D ICT Statistics: Statistics*. International Telecommunication Union. https://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspxWorld Bank Group. (2017). *Individuals using the Internet (% of population)*. World Bank Group. https://data.worldbank.org/indicator/IT.NET.USER.ZSPulgarín, A. M. R., & Woodhouse, T. (2021). *The Costs of Exclusion: Economic Consequences of the Digital Gender Gap*. Alliance for Affordable Internet, The Web Foundation. https://webfoundation.org/docs/2021/10/CoE-Report-English.pdf-Wikimedia contributors / Wikimedia Foundation. (n.d.). *Data:Wikipedia statistics*. Wikimedia Commons. Retrieved October 17, 2021, from https://commons.wikimedia.org/wiki/Data:Wikipedia_statistics/meta.tab*Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities*. (2003). https://openaccess.mpg.de/Berlin-DeclarationWikimedia Deutschland. (2021). *OpenSpeaks Accessibility*. UNLOCK. Retrieved October 17, 2021, from https://www.wikimedia.de/unlock-projects/openspeaks-accessibilityYengde, S. (2018, June 3). Dalit Cinema. *South Asia: Journal of South Asian Studies*, *41*(3), 503-518. https://doi.org/10.1080/00856401.2018.1471848Rising Voices. (2021). *Language Digital Activism Toolkit*. Rising Voices. Retrieved October 18, 2021, from https://rising.globalvoices.org/languagetoolkit/about/Jiménez, C. (2020). Radio Indígena and Indigenous Mexican Farmworkers in Oxnard, California. In *Digital Activism, Community Media, and Sustainable Communication in Latin America* (p. 366). Springer International Publishing; Palgrave Macmillan. https://doi.org/10.1007/978-3-030-45394-7Pain, P. (2017, August 4). Framing citizen activism: a comparative study of the CGNET Swara and Mobile Voices projects. *Media Asia*, *44*(2), 107-120. https://doi.org/10.1080/01296612.2016.1277825Keelery, S. (2021, March 3). *Number of government-imposed internet shutdowns in India from 2012 to 2021*. statista. Retrieved October 17, 2021, from https://www.statista.com/statistics/1095035/india-number-of-internet-shutdowns/Bali, K., Choudhury, M., Sitaram, S., Seshadri, V., & Microsoft Research Labs, Bangalore, India. (2019). ELLORA: Enabling Low Resource Languages with Technology. *Proceedings of the Language Technologies for All (LT4All)*, 160–163. https://lt4all.elra.info/proceedings/lt4all2019/pdf/2019.lt4all-1.41.pdfMinistry of Tribal Affairs, Government of India. (2011). *State wise Scheduled Tribe population and decadal change by residence : Census 2011*. Tribal Profile. Retrieved October 17, 2021, from https://tribal.nic.in/ST/Tribal%20Profile.pdfOrife, I., Kreutzer, J., Sibanda, B., Whitenack, D., Siminyu, K., Martinus, L., Ali, J. T., Abbott, J., Marivate, V., & Kabongo, S. (2020, March 13). Masakhane -- Machine Translation For Africa. *arXiv: Computation and Language*. https://arxiv.org/abs/2003.11529

OPENSPEAKS: TRANSFORMAR L'APRENENTATGE DE LES INNOVACIONS TECNOLÒGIQUES EN DOCUMENTACIÓ DE LLENGÜES AMB POCS RECURSOS EN RECURSOS EDUCATIUS OBERTS

Subhashish Panigrahi
subhashish@theofdn.org

Subhashish és un investigador i cineasta que treballa a les interseccions entre la justícia social, la documentació lingüística i la tecnologia. Ha tingut papers catalitzadors per a comunitats a tota l'Àsia-Pacífic a organitzacions sense ànim de lucre, entre elles Wikimedia Foundation, Centre for Internet and Society, Mozilla i Internet Society. És un Explorador de National Geographic, i ha ajudat moltes comunitats de parlants de llengües en perill d'extinció, realitzant documentals sobre aquestes llengües. Subhashish és el creador del projecte OpenSpeaks i co-fundador de la O Foundation.

RESUM

El projecte OpenSpeaks és un joc d'eines per a arxivistes que creen documentació audiovisual de llengües amb recursos escassos o limitats. El projecte està disponible en línia a https://en.wikiversity.org/wiki/OpenSpeaks/. A aquest article, considero algunes de les innovacions tecnològiques que s'utilitzen per protegir llengües en perill d'extinció, indígenes i d'altres amb pocs recursos, i detallo com aquestes innovacions s'avaluen per crear Recursos Educatius Oberts per a arxivistes. Entre altres observacions, emfatitzo que el ressorgiment i l'arxivament lingüístic poden semblar processos tecnològics des d'un punt de vista superficial, però estan directament lligats a la jerarquia social, econòmica i política i, per tant, sempre s'haurien de tractar des de la perspectiva de la justícia social.

1. INTRODUCCIÓ

1.1 LLENGÜES EN PERILL

El progressiu declivi de l'ús d'una llengua per parlants nadius, especialment entre els joves, posa en perill aquesta llengua. El Grup d'Experts de la UNESCO encarregat de les Llengües en Perill (UNESCO Ad Hoc Expert Group on Endangered Languages, 2003) defineix que una llengua està en perill quan els seus parlants deixen d'utilitzar-la, l'utilitzen en un nombre de dominis comunicatius cada cop més reduït, i deixen de transmetre-la d'una generació a la següent. És a dir, no hi ha parlants nous, ni adults ni infants[1]. Els lingüistes han creat diferents marcs i metodologies per identificar les llengües en perill. Com apunta el lingüista Michael Krauss, la taxa d'extinció lingüística és extraordinàriament alta, mentre que la supervivència lingüística és molt baixa. El concepte de "llengua en perill" s'emmarca de forma similar al d'espècies biològiques en perill d'extinció, i sovint s'entén de la mateixa manera que l'extinció d'una espècie. Utilitzant aquesta conceptualització,

1.- "A language is in danger when its speakers cease to use it, use it in an increasingly reduced number of communicative domains, and cease to pass it on from one generation to the next. That is, there are no new speakers, adults or children." (UNESCO Ad Hoc Expert Group on Endangered Languages, 2003).

Krauss també va crear la distinció entre llengües en perill i "llengües moribundes"; les últimes inclouen llengües que ja no parlen els descendents de la comunitat parlant (Hale et al., 1992). L'Atles UNESCO de les llengües del món en perill (en anglès, *UNESCO Atlas of the World's Languages in Danger*) inclou una llista completa de 2.464 llengües en diversos nivells de perill: des de "vulnerable" (llengua parlada per la majoria d'infants en entorns tancats com la llar) a "extinta" (no queda cap parlant viu), passant per "en clar perill" (no la parlen els infants), "en perill greu" (només la parlen les generacions més grans, però potser els pares no la parlen amb els fills), i "en perill crític" (parlada amb poca freqüència per semi-parlants, que potser la parlen parcialment) (Moseley, 2010). Els recursos educatius a OpenSpeaks s'han enriquit a través de múltiples projectes de documentació lingüística: la documentació del dialecte Baleswari de la llengua Odia (un dialecte poc documentat d'una llengua força dominant) l'any 2015, abans de la creació d'OpenSpeaks; i un projecte d'arxivament de l'any 2018 amb el suport de la National Geographic Society per documentar la llengua moribunda Kusunda (Panigrahi, 2019/2019) de l'oest del Nepal i les llengües indígenes Bonda (Rajpal, 2018) i Ho (Panigrahi, 2019/2019) d'Odisha, un estat de l'est de l'Índia.

1.2 EL PAPER DE LA TECNOLOGIA EN LA PROTECCIÓ, DOCUMENTACIÓ I REVITALITZACIÓ DE LLENGÜES

Les llengües en general, i les llengües en perill i d'altres amb pocs recursos en particular, es veuen impactades de diverses maneres a mesura que la tecnologia afecta les societats d'arreu del món. Ha augmentat significativament l'accés a la tecnologia digital, especialment gràcies a l'omnipresència de telèfons intel·ligents i el cost decreixent de les comunicacions mòbils de dades. Tanmateix, encara que la penetració d'Internet està creixent a un ritme excepcionalment alt a tot el món, mentre un 51% de les persones ja tenen accés a Internet l'any 2019 (International Telecommunication Union, 2021), només el 14% de tota la població a països pobres pot permetre's utilitzar Internet l'any 2017 (World Bank Group, 2017). L'assequibilitat d'Internet es veu encara més afectada quan parlem d'etnicitat i gènere. La Fundació World Wide Web estima que en una dècada, els països amb una renda baixa i mitjana-baixa han perdut al voltant d'1 bilió de dòlars en Producte Interior Brut (PIB) ja que les dones no podien accedir-hi ni participar en línia degut a diverses barreres. A més, la "bretxa digital de gènere" només ha disminuït del 30,9% al 30,4% des de l'any 2011 (Pulgarín & Woodhouse, 2021). Des del mes d'octubre de l'any 2021, existeixen Viquipèdies actives en 312 llengües (un 4,8% de les 6.500 llengües reconegudes per la UNESCO) d'arreu del món (Wikimedia contributors / Wikimedia Foundation, n.d.). Donat que la Viquipèdia és una enciclopèdia escrita que conté citacions a recursos publicats i la majoria de llengües mundials no tenen un sistema d'escriptura propi sense coneixements publicats a revistes arbitrades i d'altres publicacions notables, la Viquipèdia també esdevé inadequada en la majoria de casos per a les llengües amb pocs recursos esmentades a aquest article.

1.3 Contextos socioeconòmics i polítics

Al llarg d'aquest projecte, els creadors de continguts, els periodistes, els acadèmics, els lingüistes/documentalistes de camp i molts altres professionals que creen i comparteixen continguts lingüístics en les seves llengües d'interès es denominen en general, fent servir un terme autoexplicatiu, com a "arxivista multimèdia de llengües" (un arxiver que realitza documentació audiovisual i d'altres tipus de les llengües) i que s'escurça en alguns llocs simplement com a "arxiver" o "arxivista", i l'activisme relacionat amb la documentació de les llengües com a "arxivisme

lingüístic" (portmanteau de " activisme" i "arxivar"). De la mateixa manera, la referència als recursos escassos i limitats es refereix als recursos financers, humans, sociopolítics, institucionals/infraestructurals, tècnics i tots els altres àmbits essencials dels quals moltes llengües manquen, cosa que alhora repercuteix en la seva supervivència , ús o desenvolupament posterior. La disponibilitat de finançament a través de polítiques públiques, les persones, les infraestructures institucionals públiques i privades, els coneixements tècnics i la participació activa de les organitzacions de la societat civil són alguns dels recursos més crítics per a la supervivència de les llengües. El poder polític i econòmic d'una comunitat de parlants d'una llengüa limita l'accés de la comunitat als recursos per a l'ús actiu intergeneracional i el desenvolupament de la seva llengua. Sovint hi ha un impacte directe de l'explotació colonial i postcolonial de moltes comunitats marginades a través de mitjans polítics, ambientals i socioeconòmics que a càrrec de comunitats majoritàries veïnes. Aquesta explotació no només repercuteix en la desintegració cultural i l'assimilació de molts grups marginats a les societats majoritàries mentre segueixen sent oprimits pels grups dominants (com els pobles indígenes de tot el món), sinó que també restringeix en gran mesura l'accés als recursos per al creixement de les llengües.

2. EL DISSENY D'OPENSPEAKS

Donar suport a arxivistes que documenten llengües en perill, indígenes, minoritàries i d'altres amb recursos escassos o limitats ha estat l'àrea clau d'actuació del projecte OpenSpeaks des de la seva concepció l'any 2015 i el seu llançament l'any 2017. Cadascuna d'aquestes categories lingüístiques té els seus propis reptes i necessitats específiques. Tot i que la definició de cada categoria varia de lingüista a lingüista, la manca de recursos continua essent una barrera crítica per a la documentació de la gran majoria/totes les llengües d'aquestes categories. S'ha d'examinar de forma crítica l'statu quo de disponibilitat (o escassetat) de recursos per a cada llengua per identificar quins recursos específics o quin tipus de flux de treball són pertinents per un projecte de documentació específic.

2.1 PRINCIPIS DE DISSENY I ABAST

Els principis de disseny del projecte OpenSpeaks rauen en termes generals en documentar les metodologies, bones pràctiques i altres lliçons de les obres d'arxivament d'arxivistes que documenten llengües indígenes, en perill i d'altres amb pocs recursos, en suport d'àudio i vídeo, i creen materials educatius que poden utilitzar arxivistes principiants i de nivell intermedi.

Els mòduls i tutorials dins d'OpenSpeaks mostren coneixements d'àrees concretes per la documentació audiovisual de llengües, però també marcs per ajudar a cada arxivista a posicionar la seva llengua d'estudi perquè puguin desxifrar quins recursos són rellevants al seu context. Tots els continguts de base d'aquest projecte estan disponibles públicament en anglès senzill per ajudar a arxivistes amb un nivell bàsic d'anglès a accedir-hi i ajudar a traductors a localitzar el projecte en altres llengües del món.

2.2 OPENSPEAKS COM A RECURS EDUCATIU OBERT

OpenSpeaks com a recurs està destinat a arxivistes multimèdia de llengües, principiants o de nivell intermedi; es dona per fet que, en la majoria de casos, l'arxivista podria no tenir accés a gaires recursos crítics. El projecte fa servir una autoavaluació (una sèrie de preguntes) que l'arxivista pot respondre per decidir per si mateix.

Temps: Les següents preguntes són per definir la limitació/disponibilitat de temps tant de l'arxivista com de les persones entrevistades. L'arxivista podria fer una autoavaluació preguntant-se "Quant de temps puc contribuir a la documentació?" i "Quant de temps poden contribuir les persones entrevistades a la documentació?". La segona pregunta depèn força del factor "Fons", del qual parlarem una mica més avall.

Maquinari: El maquinari/equip, el programari i els coneixements operatius dels mateixos es poden agrupar com a recursos tècnics. Com gran part del segment operatiu de la documentació lingüística és un procés tècnic, es necessita certa planificació tècnica prèvia. Tenir un telèfon intel·ligent que pugui gravar almenys una o dues hores d'àudio/vídeo es el mínim necessari per a la documentació. Algunes preguntes clau d'aquesta secció podrien ser: "Quins són els diferents maquinaris/equips que necessitaria per poder documentar i quins ja tinc?" o "Tinc una gravadora d'àudio o una càmera o qualsevol altre accessori (p. ex. un trípode i una muntura per subjectar el mòbil/la càmera al trípode, o un micròfon extern compatible amb el mòbil/la càmera)?" i "A quins d'aquests equips que necessito no hi tinc accés, i com puc aconseguir-los i aprendre a utilitzar-los?".

Edició: Després de gravar l'àudio/vídeo, cal editar-lo i processar-lo abans de publicar-lo. A vegades les limitacions de temps poden disminuir les capacitats de postproducció de l'arxivista. Si un vídeo es retransmet en directe a les xarxes socials o a una plataforma com YouTube, no hi ha gaires possibilitats d'editar-lo. L'arxivista multimèdia de llengües pot preguntar-se: "Tinc accés a una aplicació al mòbil/ordinador per editar l'àudio/vídeo (a no ser que sigui una retransmissió en directe)?", "Amb quines aplicacions tinc experiència funcional i quines de noves he d'aprendre a utilitzar?". Alguns processos de retransmissió en directe tenen opcions i fins i tot equips muntats per editar àudio/vídeo en directe, però això va més enllà d'aquest article.

Fons: L'accés a fons per pagar la feina de l'arxivista i del seu equip de suport, comprar/llogar equips, o remunerar als entrevistats pel seu temps són alguns dels factors econòmics. És important dur a terme l'avaluació en l'ordre establert aquí. Per identificar la seva situació econòmica a nivell de projecte, l'arxivista pot preguntar-se "Quants diners tindré per invertir en la manera que penso documentar?". Generalment, es recomana elevar aquesta xifra un 10-20%, ja que els costs reals poden augmentar i la feina no s'hauria d'aturar per això. També pot ser útil preguntar-se "D'on trauré els diners que calculo que necessito per aquest projecte?". Aquesta pregunta els pot ajudar a buscar fonts de finançament, a no ser que ja els hagin acordat.

L'exemple que es presenta a continuació es basa en un marc per a una autoavaluació objectiva i l'omple un arxiver concret: un estudiant universitari que té previst anar a una ciutat propera i donar suport a la documentació d'unes quantes llengües de recursos escassos.

2.3 LLICÈNCIA

Com la plataforma amfitriona Wikiversity és de llicència lliure, la llicència Creative Commons Attribution-ShareAlike (CC BY-SA) versió 3.0, OpenSpeaks també és obert amb la mateixa llicència CC BY-SA 3.0. "Obert" en aquest context significa que el projecte està disponible per Accés Obert[2]; marcs internacionals com la Declaració de Berlín sobre

2.- Open Access

Recurs	Situació actual	Suport necessari	Com obtenir-lo?
1. Temps	5-6 hores/ setmana durant 2 mesos	Més temps durant la postproducció	Necessitat de trobar un editor de vídeo voluntari "O" fer temps per editar el vídeo
2. Recursos tècnics	Telèfon intel·ligent per fer gravacions tant d'àudio com de vídeo i un petit trípode de sobretaula per a gravacions fixes	Un micròfon de solapa pot millorar la qualitat de veu de l'entrevistat	No tinc pressupost per comprar un micròfon però me'l pot prestar una amiga si l'aviso amb temps
3. Edició	Actualment no tinc programari ni habilitats d'edició	He d'aprendre a utilitzar les aplicacions d'escriptori Kdenlive per a l'edició de vídeo i Audacity per a l'edició d'àudio	He de desar tutorials útils de YouTube i mirar-los per aprendre a utilitzar aquestes eines d'edició
4. Fons	Puc cobrir les despeses de transport local	Necessito uns 300 $ per pagar a un editor de vídeo 15 hores d'edició; 40 $ per un micròfon; i 30 $ pel transport local	a. Demanaré una petita subvenció al meu departament universitari per poder comprar un micròfon i remunerar a la Maya (amiga que sap editar) b. Sinó, intentaré que el Kai (amic) em presti un micròfon i parlaré amb la Maya de si està disponible per ajudar-me amb l'edició de vídeo c. Si rebo menys del que demano, li demanaré al Kai que em presti el micròfon i li pagaré a la Maya la totalitat/parcialment

Taula 1: Marc de planificació de projecte OpenSpeaks

l'Accés Obert al Coneixement en Ciències i Humanitats, entre d'altres, especifiquen com es poden oferir continguts científics i d'humanitats mitjançant llicències lliures (*Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities*, 2003). Generalment, ens referim als Recursos Educatius amb Accés Obert, en contraposició als continguts restringits mitjançant murs de pagament, com a Recursos Educatius Oberts o REO (o OER, per les seves sigles en anglès, *Open Educational Resources*). OpenSpeaks també està obert a contribucions públiques com d'altres projectes de Wikiversity i Wikimedia. Resumint, aquests principis de disseny tant permeten l'Accés Obert com animen a la participació/contribució oberta a tothom.

2.4 INTEGRAR LLIÇONS DE LA DOCUMENTACIÓ DE CAMP I PROVES D'USUARI

Sota la tutela del projecte OpenSpeaks, s'han creat diverses pel·lícules documentals com "Gyani Maiya" (2019; arxivament de Kusunda, una llengua moribunda del Nepal), "Remosam" (2019; arxivament de Bonda, una llengua en perill de l'Índia) i "Mage Porob" (2020; arxivament de Ho, una llengua en perill de l'Índia). Les lliçons globals d'aquestes pel·lícules capten les relacions ebtre diferents parlants nadius i cineastes com jo que són parlants no nadius i investigadors que són o bé parlants nadius o parlants no nadius autodidactes de les llengües respectives. Al procés de documentació de moltes llengües amb pocs

recursos, tant la implicació inicial com fins i tot a llarg termini i predominant de parlants no nadius pot donar lloc a una comunitat nativa més petita que acabi essent propietària dels continguts i la narrativa documentada. En teoria, l'ideal ètic de la documentació lingüística d'OpenSpeaks dona màxima prioritat a la comunitat nativa i a que aquesta sigui propietària dels continguts lingüístics. El repte, però, rau en identificar processos de documentació lingüística liderats per arxivistes que també són parlants nadius, i que el joc d'eines OpenSpeaks prengui forma al voltant dels seus reptes, lliçons i recomanacions. No només hi ha pocs arxivistes nadius, sinó que també tenen les limitacions afegides de temps, recursos econòmics i barreres lingüístiques. A més, és possible que captar lliçons i bones pràctiques no sigui una prioritat en la seva feina. Captar les experiències d'alguns d'aquests arxivistes només ha estat possible en comptades ocasions gràcies a 10 entrevistes a usuaris i proves d'usuari de prototips d'OpenSpeaks a un projecte incubadora l'any 2021 anomenat UNLOCK (Wikimedia Deutschland, 2021).

2.5 CONSCIÈNCIA DE PRIVILEGIS PROPIS

Igual que el posicionament de cada llengua és únic a l'espectre de la documentació lingüística, els reptes, les oportunitats i els privilegis de cada arxivista també s'han d'entendre en relación als seus antecedents socioeconòmics. Per exemple, la meva capacitat personal per dedicar temps a documentar l'aprenentatge a OpenSpeaks està profundament lligada als meus propis privilegis, tant històrics com sistèmics, com ara el sistema discriminatori de castes de l'Índia. El sistema de castes ha estat utilitzat com a escut religiós per un determinat sector de la societat (al qual l'acadèmic Suraj Yengde sol anomenar àmpliament "casta superior" i "casta opressora") (Yengde, 2018) per oprimir els dalits (abans coneguts pejorativament com a "intocables", que han estat discriminats i criminalitzats durant molt de temps pel seu origen ètnic), als pobles adivasi (indígenes) i als bahujan (altres grups ètnics minoritaris i marginats per altres motius). En general, els Dalit-Adivasi-Bahujan han estat privats de l'agència socioeconòmica-política i dels privilegis contemporanis, inclòs l'accés a moltes formes de privilegis digitals (per exemple, accés a dispositius, accés a Internet, etc.) durant molt de temps temps al subcontinent indi i fins i tot més enllà, on és present la diàspora del sud d'Àsia. Preguntar "què" es documenta pot ser important des del punt de vista lingüístic per cobrir una àmplia gamma de temes, però "qui" documenta i "de qui" són les veus/narracions que documenten és crucial des d'una perspectiva de classe i de gènere. Al meu entendre, el gènere de l'arxivista és un dels factors més ignorats però alhora més crítics per al "què" es documenta. Com a documentalista masculí, sovint he experimentat de primera mà l'autocensura per part de les entrevistades femenines i no binàries, que sorgeix de la incomoditat basada en el gènere, i que podrien haver compartit importants punts de vista si haguessin estat entrevistades per un entrevistador del mateix gènere. També es donen casos similars respecte al valor significatiu de la jerarquia de classe social entre els entrevistats i els arxivers. OpenSpeaks detalla la identificació de les jerarquies de gènere i de classe social i la inclusió de formes en el procés de planificació previ a la documentació per ajudar a abordar els problemes esmentats. Més que les innovacions tecnològiques, el reconeixement dels prejudicis socials profundament arrelats i les mesures per abordar els problemes sistèmics (i no sols els tòpics) són extremadament clau per crear una estratègia a llarg termini per a la documentació lingüística. Atès que l'accés a la formació formal, el finançament i el suport institucional són escassos per a la majoria de llengües amb pocs recursos, els recursos educatius d'OpenSpeaks s'estan actualitzant en op-

timitzar el projecte per a l'aprenentatge al ritme de cadascú i trencant la barrera dels coneixements tècnics previs.

3. APRENDRE D'ALTRES INNOVACIONS SOCIOTÈCNIQUES

3.1 COL·LECTIUS CONNECTATS

Trobar pràctiques existents d'innovacions tècniques i d'altres formes d'activisme digital per a la documentació/conservació lingüística és relativament més fàcil que construir una xarxa amb els principals actors actius de l'activisme digital lingüístic. Com a gran xarxa connectada a centenars d'individus i organitzacions de tot el món que participen en l'activisme digital lingüístic per a les llengües pobres en recursos, Rising Voices ha estat fonamental per facilitar les pràctiques d'intercanvi de coneixements i recursos. Aquests individus i organitzacions intercanvien coneixements, assessoren i fins i tot col·laboren entre si i més enllà de la facilitació de Rising Voices. D'altra banda, un col·lectiu com Rising Voices també crea un efecte de xarxa en reunir actors de la societat civil, activistes i fins i tot organitzacions internacionals com la UNESCO. Un exemple n'és el Projecte d'Activisme Digital Lingüístic (Rising Voices 2021) de Rising Voices, l'objectiu del qual és proporcionar un full de ruta per a l'activisme digital lingüístic fomentant la col·laboració entre els actors actuals i els nous; això darrer va ser possible gràcies a una sèrie de tallers centrats en els activistes. Els debats continus sobre qüestions compartides i úniques, i les vies per compartir l'aprenentatge i els recursos d'interès comú reforcen encara més les col·laboracions.

3.2 MITJANS DE COMUNICACIÓ COMUNITARIS INDÍGENES

Existeixen molts tipus de plataformes de mitjans de comunicació comunitaris indígenes, alguns liderats per les mateixes comunitats i d'altres per les ONG locals amb la participació de la comunitat. Les ràdios comunitàries generalment inclouen productors i oients locals. Encara que el mètode de comunicació de les ràdios comunitàries sol ser FM, una tecnologia anterior a l'era digital, les comunitats ho combinen amb pràctiques contemporànies com la promoció i participació a través de les xarxes socials. "Radio Indígena 94.1 FM" al Comptat de Ventura de Califòrnia atén als treballadors mexicans immigrants que parlen principalment els dialectes mixteca o zapoteca. Mixteco/Indígena Community Organizing Project (MICOP). L'organització darrere d'aquest canal FM ha aconseguit fer canvis a nivell de política amb l'aprovació de lleis d'hores extres per als treballadors agrícoles (Jiménez, 2020). És comú que organitzacions de ràdios comunitàries esdevinguin agents causals per a l'accés a drets humans. Encara que per moltes organitzacions la documentació o protecció lingüística no és sempre el centre d'atenció, l'ús de llengües indígenes o qualsevol altra llengua (o dialecte) infrarepresentada com a mitjà de participació i altres formes de comunicació porten finalment a la documentació lingüística. A l'estat de Chhattisgarh de l'Índia, el Central Gondwana Net Swara (CGNet Swara) funciona com un portal de notícies rurals que utilitza trucades telefòniques per recollir notícies gravades en àudio per periodistes ciutadans que són principalment membres de la tribu indígena Gond. La gravació de veu interactiva (IVR, per les seves sigles en anglès, *interactive voice recording*), la tecnologia darrere de CGNet Swara, és una tecnologia força vella però efectiva, ja que només s'ha de disposar d'una línia de telèfon fix o connexió mòbil (Pain, 2017). Donat que a diferents parts de l'Índia l'Internet sovint s'apaga o s'alenteix, l'Índia és el país on l'Internet s'interromp amb més freqüència al món (Keelery, 2021), comptar amb una tecnologia més bàsica més fiable dins el context

donat ha esdevingut una solució efectiva per a CGNet Swara. El periodisme basat en la veu també trenca la barrera d'alfabetització de llengües dominants o fins i tot l'alfabetització digital/d'Internet. Molts parlants de gondi i kurukh (ambdues llengües vulnerables o en perill) també aporten notícies a les seves respectives llengües que llavors es tradueixen a l'hindi. El fet que l'hindi sigui una llengua oficial de Chhattisgarh, una llengua dominant al centre i nord de l'Índia, i la llengua en la qual es troben la majoria dels articles a CGNet Swara, augmenta les possibilitats de que les notícies es difonguin més àmpliament. Més del 30% de la població de Chhattisgarh inclou diferents grups indígenes, anomenats col·lectivament els adivasi dins del context socioètnic i Tribus Reconegudes (TR) segons les classificacions oficials (Ministeri d'Afers Tribals, Govern de l'Índia, 2011)[3], que han estat oprimits històricament a través de la mineria, el sistema de castes divisiu de l'Índia i feines mal remunerades, a més de ser víctimes de la batalla política constant entre el moviment naxalita i el govern. El model de CGNet Swara és un bon exemple de la contextualització de la tecnologia per a una llengua amb pocs recursos per entendre el que funciona a un grup demogràfic concret i prendre accions pràctiques per documentar qüestions de drets humans (Bali et al., 2019). L'ús del periodisme ciutadà com a eina també ha estat efectiva en aquest cas per a la documentació i l'ús digital de les llengües gondi i hurukh.

3.3 MASAKHANE: INNOVACIÓ TECNOLÒGICA LINGÜÍSTICA LIDERADA PER LA COMUNITAT LOCAL

Les comunitats lingüístiques que puguin estar dividides lingüísticament poden trobar-se problemes que siguin rellevants a les llengües de l'altre. Una solució tecnològica per una llengua que normalment prengui més temps en crear-se en un primer moment pot esdevenir una plantilla per d'altres llengües que comparteixin els mateixos factors lingüístics i/o demogràfics. Identificar els punts en comú entre llengües que comparteixen una història de colonització i reptes tecnològics actuals ha estat l'enllaç solidari per la comunitat Masakhane (https://www.masakhane.io/) que connecta parlants de 38 llengües africanes de 30 països diferents amb l'objectiu de trobar solucions tecnològiques en la recerca del processament del llenguatge natural (PLN). A l'Àfrica es parlen més de 2.000 llengües. La llarga colonització de moltes comunitats de l'Àfrica ha provocat danys significatius al separar les comunitats de les seves llengües pròpies i substituir les llengües natives per llengües dominants com l'àrab, l'anglès, el francès i el portuguès. Des de l'any 2019, l'enfocament polític comunitari de Masakhane ha contribuït a elaborar una xarxa solidària de més de 1.000 col·laboradors de la regió i a innovar pel futur de les llengües africanes. Masakhane es defineix com un esforç de recerca de traducció automàtica en línia, distribuïda, continental, de codi obert, per llengües africanes[4] (Orife et al., 2020). La feina de l'organització inclou construir una comunitat lingüística i tecnològica, i recolzar-la amb els recursos necessaris per fer recerca oberta, participativa i multidisciplinària. Al desfer-se de qualsevol criteri d'elegibilitat, Masakhane proveeix d'eines als nous col·laboradors individuals i els ajuda posant-los en contacte amb mentors i d'altres col·laboradors perquè, junts, puguin trobar àrees d'interès, i enfortir col·lectivament diferents projectes de codi obert. Gràcies a l'àmplia xarxa de col·laboradors i el seu enfocament en la

3.- Scheduled Tribes (ST), Ministry of Tribal Affairs, Government of India.

4.- : "Masakhane constitutes an open-source, continent-wide, distributed, online research effort for machine translation for African languages."

recerca, l'organització pot identificar i prioritzar les àrees de col·laboració. Per exemple, la traducció automàtica, un mètode per traduir automàticament un text d'una llengua a una altra, encara no està disponible per un gran nombre de llengües africanes, fet que crea una diferència de coneixement enorme, ja que la majoria dels continguts a Internet estan disponibles en llengües dominants. Masakhane ha elaborat eines per a la col·laboració comunitària i per posar a prova eines de traducció automàtica neuronal (NMT, per les seves sigles en anglès, *neural machine translation*) que són les primeres en intentar eliminar les distàncies en l'accés a la informació a través de la traducció.

La feina de Masakhane emfatitza el valor de construir una comunitat de col·laboradors per a innovacions tecnològiques degut als seus fonaments filosòfics —"Umuntu Ngumuntu Ngabantu" (cita en la llengua isiZulu que vol dir "una persona és persona a través d'una altra persona" o "Jo soc perquè tu ets")— que troben una solució natural al programari de codi obert. Masakhane indica un canvi de deixar el destí de les llengües locals en mans de grans societats estrangeres com Google o Microsoft a centrar les innovacions tecnològiques al voltant de qüestions sociolingüístiques, i proveir les comunitats locals perquè puguin liderar aquestes innovacions.

4. CONCLUSIONS

El fons històric de cada llengua amb pocs recursos proporciona un context sobre com el dret a parlar la llengua pròpia està directament relacionat al dret d'accés a la informació i, per tant, al dret a la vida, i com aquests drets es violen sistemàticament mitjançant l'ús de llengües dominants com a eina d'opressió. Les llengües que estan a punt d'extingir-se certament necessiten una resposta immediata i acció pràctica en forma de documentació. Tanmateix, cal que l'objectiu a llarg termini sigui proveir els parlants nadius amb els recursos econòmics i educatius entre d'altres perquè ells mateixos siguin els encarregats de construir solucions tecnològiques lingüístiques. La descolonització de la tecnologia requereix invertir l'equació històrica de persones amb poder que, a través dels seus privilegis, controlen les jerarquies de poder social, polític i econòmic per mantenir-se al poder. Dins el context de llengües en perill, indígenes i d'altres amb pocs o escassos recursos, el perill imminent prové tant de fora com de dins de la comunitat nativa a causa de la opressió de gran abast. Enlloc de preguntar-se quines solucions tecnològiques tractaran el problema del poc ús d'una d'aquestes llengües, cal contextualitzar els problemes des d'una perspectiva antropològica. Per exemple, preguntar-se "Perquè una llengua concreta domina una altra i qui guanya poder econòmic i polític gràcies a aquesta estructura?" pot ajudar a trobar els actors que hi participen. En la majoria de casos, els problemes lingüístics poden necessitar una solució tecnològica des dels seus inicis. Però desconèixer els motius socials, econòmics i polítics que causen aquests problemes podria donar lloc a la creació d'una tecnologia aïllada de les persones que realment podrien beneficiar-se'n. Molts problemes socials són extremadament complexos i solucionar-los requereix molt de temps. És molt poc probable que els arxivistes, que són, en la majoria de casos, individus amb accés limitat a recursos humans, econòmics i d'altres, puguin arribar a resoldre problemes tan grans. No obstant això, tractar els problemes estructurals durant la creació de les solucions aturaria la repetició del mateix cercle viciós d'opressió social que segueix essent habitual a la tecnologia. OpenSpeaks pren inspiració filosòfica i pràctica d'organitzacions com Masakhane, que plasma l'enfocament ascendent de canviar la societat tecnològica, campanyes feministes globals

com WhoseKnowledge?, que treballen per la "descolonització de l'Internet", o xarxes de moviment com Rising Voices, que crea una via per compartir idees i recursos. El tema subjacent a la majoria d'aquestes obres és l'enfocament dels problemes lingüístics com a problemes socials i la utilització de la tecnologia com a eina per impulsar innovacions en les pràctiques socials i conductuals. Com el disseny d'OpenSpeaks absorbeix constantment els coneixements d'altres líders del moviment, el procés de disseny i els continguts educatius que ofereix als arxivistes lingüístics avancen cap a reflectir les mateixes lliçons. Enlloc de ser continguts prescriptius, el nou disseny vol ajudar als arxivistes a veure la seva feina de documentació lingüística a través de la mateixa perspectiva de justícia social, i formar-se en els recursos disponibles que encaixin amb les seves necessitats.

5. REFERÈNCIES

Bali, K., Choudhury, M., Sitaram, S., Seshadri, V., & Microsoft Research Labs, Bangalore, India. (2019). ELLORA: Enabling Low Resource Languages with Technology. *Proceedings of the Language Technologies for All (LT4All)*, 160–163. https://lt4all.elra.info/proceedings/lt4all2019/pdf/2019.lt4all-1.41.pdf

Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities. (2003). https://openaccess.mpg.de/Berlin-Declaration

Hale, K., Krauss, M., Watahomigie, L. J., Yamamot, A. Y., Craig, C., Jeanne, L. M., & England, N. C. (1992). Endangered Languages. *Language*, *68*(1), 42. https://doi.org/10.2307/416368

International Telecommunication Union. (2021). *ITU-D ICT Statistics: Statistics*. International Telecommunication Union. https://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx

Jiménez, C. (2020). Radio Indígena and Indigenous Mexican Farmworkers in Oxnard, California. In *Digital Activism, Community Media, and Sustainable Communication in Latin America* (p. 366). Springer International Publishing; Palgrave Macmillan. https://doi.org/10.1007/978-3-030-45394-7

Keelery, S. (2021, March 3). *Number of government-imposed internet shutdowns in India from 2012 to 2021*. statista. Recuperat el 17 d'octubre de 2021, de https://www.statista.com/statistics/1095035/india-number-of-internet-shutdowns/

Ministry of Tribal Affairs, Government of India. (2011). *State wise Scheduled Tribe population and decadal change by residence: Census 2011*. Tribal Profile. Recuperat el 17 d'octubre de 2021, de https://tribal.nic.in/ST/Tribal%20Profile.pdf

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

Orife, I., Kreutzer, J., Sibanda, B., Whitenack, D., Siminyu, K., Martinus, L., Ali, J. T., Abbott, J., Marivate, V., & Kabongo, S. (2020, March 13). Masakhane -- Machine Translation For Africa. *arXiv: Computation and Language*. https://arxiv.org/abs/2003.11529

Pain, P. (2017, August 4). Framing citizen activism: a comparative study of the CGNET Swara and Mobile Voices projects. *Media Asia*, *44*(2), 107-120. https://doi.org/10.1080/01296612.2016.1277825

Panigrahi, S. (Director). (2019). *Gyani Maiya* [ज्ञानी मैया] [Film]. Entertainment Identifier Registry. https://ui.eidr.org/view/content?id=10.5240/52AE-86BB-F84D-03B2-D938-U (Obra original publicada l'any 2019)

Panigrahi, S. (Director). (2019). *Mage Porob* [ ] [Film]. O Foundation. http://dx.doi.org/10.17613/t2cb-bg17 (Original work published 2019)

Pulgarín, A. M. R., & Woodhouse, T. (2021). *The Costs of Exclusion: Economic Consequences of the Digital Gender Gap*. Alliance for Affordable Internet, The Web Foundation. https://webfoundation.org/docs/2021/10/CoE-Report-English.pdf

Rajpal, S. (2018, August 25). O Foundation and National Geographic set out to document endangered languages before they are lost forever. *edex, The New Indian Express*. https://www.edexlive.com/happening/2018/aug/25/o-foundation-and-national-geographic-set-out-to-document-endangered-languages-before-they-are-lost-f-3728.html/

Rising Voices. (2021). *Language Digital Activism Toolkit*. Rising Voices. Recuperat el 18 d'octubre de 2021, de https://rising.globalvoices.org/languagetoolkit/about/

UNESCO Ad Hoc Expert Group on Endangered Languages. (2003). *Language Vitality and Endangerment*. UNESCO. https://ich.unesco.org/doc/src/00120-EN.pdf

Wikimedia contributors / Wikimedia Foundation. (n.d.). *Data: Wikipedia statistics*. Wikimedia Commons. Recuperat el 17 d'octubre de 2021, de https://commons.wikimedia.org/wiki/Data:Wikipedia_statistics/meta.tab

Wikimedia Deutschland. (2021). *OpenSpeaks Accessibility*. UNLOCK. Recuperat el 17 d'octubre de 2021, de https://www.wikimedia.de/unlock-projects/openspeaks-accessibility

World Bank Group. (2017). *Individuals using the Internet (% of population)*. World Bank Group. https://data.worldbank.org/indicator/IT.NET.USER.ZS

Yengde, S. (2018, June 3). Dalit Cinema. *South Asia: Journal of South Asian Studies*, *41*(3), 503-518. https://doi.org/10.1080/00856401.2018.1471848Hale, K., Krauss, M., Watahomigie, L. J., Yamamot, A. Y., Craig, C., Jeanne, L. M., & England, N. C. (1992). Endangered Languages. *Language*, *68*(1), 42. https://doi.org/10.2307/416368Moseley, C. (2010). *Atlas of the World's Languages in Danger*. UNESCO Atlas of the World's Languages in Danger. http://www.unesco.org/culture/en/endangeredlanguages/atlasPanigrahi, S. (Director). (2019). *Gyani Maiya* [ज्ञानी मैया] [Film]. Entertainment Identifier Registry. https://ui.eidr.org/view/content?id=10.5240/52AE-86BB-F84D-03B2-D938-U (Original work published 2019)Rajpal, S. (2018, August 25). O Foundation and National Geographic set out to document endangered languages before they are lost forever. *edex, The New Indian Express*. https://www.edexlive.com/happening/2018/aug/25/o-foundation-and-national-geographic-set-out-to-document-endangered-languages-before-they-are-lost-f-3728.html/Panigrahi, S. (Director). (2019). *Mage Porob* [ ] [Film]. O Foundation. http://dx.doi.org/10.17613/t2cb-bg17 (Original work published 2019)International Telecommunication Union. (2021). *ITU-D ICT Statistics: Statistics*. International Telecommunication Union. https://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspxWorld Bank Group. (2017). *Individuals using the Internet (% of population)*. World Bank Group. https://data.worldbank.org/indicator/IT.NET.USER.ZSPulgarín, A. M. R., & Woodhouse, T. (2021). *The Costs of Exclusion: Economic Consequences of the Digital Gender Gap*. Alliance for Affordable Internet, The Web Foundation. https://webfoundation.org/docs/2021/10/CoE-Report-English.pdfWikimedia contributors / Wikimedia Foundation. (n.d.). *Data:Wikipedia statistics*. Wikimedia Commons. Retrieved October 17, 2021, from https://commons.wikimedia.org/wiki/Data:Wikipedia_statistics/meta.tab*Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities*. (2003). https://openaccess.

mpg.de/Berlin-DeclarationWikimedia Deutschland. (2021). *OpenSpeaks Accessibility*. UNLOCK. Retrieved October 17, 2021, from https://www.wikimedia.de/unlock-projects/openspeaks-accessibilityYengde, S. (2018, June 3). Dalit Cinema. *South Asia: Journal of South Asian Studies*, *41*(3), 503-518. https://doi.org/10.1080/00856401.2018.1471848Rising Voices. (2021). *Language Digital Activism Toolkit*. Rising Voices. Retrieved October 18, 2021, from https://rising.globalvoices.org/languagetoolkit/about/Jiménez, C. (2020). Radio Indígena and Indigenous Mexican Farmworkers in Oxnard, California. In *Digital Activism, Community Media, and Sustainable Communication in Latin America* (p. 366). Springer International Publishing; Palgrave Macmillan. https://doi.org/10.1007/978-3-030-45394-7Pain, P. (2017, August 4). Framing citizen activism: a comparative study of the CGNET Swara and Mobile Voices projects. *Media Asia*, *44*(2), 107-120. https://doi.org/10.1080/01296612.2016.1277825Kelery, S. (2021, March 3). *Number of government-imposed internet shutdowns in India from 2012 to 2021*. statista. Retrieved October 17, 2021, from https://www.statista.com/statistics/1095035/india-number-of-internet-shutdowns/Bali, K., Choudhury, M., Sitaram, S., Seshadri, V., & Microsoft Research Labs, Bangalore, India. (2019). ELLORA: Enabling Low Resource Languages with Technology. *Proceedings of the Language Technologies for All (LT4All)*, 160–163. https://lt4all.elra.info/proceedings/lt4all2019/pdf/2019.lt4all-1.41.pdfMinistry of Tribal Affairs, Government of India. (2011). *State wise Scheduled Tribe population and decadal change by residence : Census 2011*. Tribal Profile. Retrieved October 17, 2021, from https://tribal.nic.in/ST/Tribal%20Profile.pdfOrife, I., Kreutzer, J., Sibanda, B., Whitenack, D., Siminyu, K., Martinus, L., Ali, J. T., Abbott, J., Marivate, V., & Kabongo, S. (2020, March 13). Masakhane -- Machine Translation For Africa. *arXiv: Computation and Language*. https://arxiv.org/abs/2003.11529

RISING VOICES: INDIGENOUS LANGUAGE DIGITAL ACTIVISM

Eddie Avila

During three days in October 2014, 30 speakers of Indigenous languages of Mexico gathered at the Juan de Córdova library in the heart of downtown Oaxaca, the capital city of a state known for its cultural and linguistic diversity. This particular occasion marked the first ever meeting of its kind, convening language champions actively leveraging the power and the potential of the internet, digital media, and other technologies to promote and share their native languages in digital spaces.

For many, this type of "un-conference" was a revelation. The participants themselves offered to lead workshops or facilitate discussions. Many remarked that they never imagined that there were others like themselves who were adopting these digital strategies to make sure their language and culture were reflected online.

Of course, the use of digital tools had been playing an important role in language documentation and promotion for years before this first meeting. However, it is most likely the Oaxaca 2014 meeting was the first time that these types of activities had been described as "Indigenous language digital activism."

This label was used on purpose, not to be provocative, but to recognize that, to advance change, it would take intentional steps towards this end. Of course, in some circles in many parts of the world, the label "activist" could land one in hot water for those resisting being seen as agitators. However, as many participants noted, in order to see change happen on a local, national, or even international level, some type of challenge to the status quo needs to happen. In this case, the status quo usually meant Indigenous languages being relegated to a lower status by governments or technology companies that fail to provide conditions so that Indigenous languages can flourish.

The changes that some digital activists want to help shape include wanting to see more digital content available in their language to pique the interest of younger generations encouraging them to keep their mother languages.

Since that first gathering, more and more people have identified with the label because in it they found a certain kinship with others all working towards similar changes. The concept also picked up steam on international platforms, becoming part of the recommendations for the upcoming International Decade of Indigenous Languages 2022-2032. The Chapoltepek [Los Pinos] Declaration stated the need for the:

> *Promotion of networks of digital activists and champions for the teaching and learning of Indigenous languages, as well as the exchange of best practices related to the use of technology.*

While it may be satisfying to see the work of Indigenous language digital activists named in such spaces, and the attention that the work has received in local, national, and international press, this work has grown to a point where it is not enough to merely be recognized. Instead, this movement seeks to have a greater influence to affect positive change.

However, what does that change look like? How can the impact on languages affected by digital activism go beyond simple online metrics? And who should be the ones who decide what it means to be successful?

THE ROLE OF RISING VOICES IN THIS MOVEMENT

Rising Voices (RV), the digital inclusion and equity arm of the organization Global Voices, was one of three partners that were the catalysts for the First Gathering of Indigenous Language Digital Activism held in Oaxaca in 2014. While it may have been the first activity held under the umbrella of "Indigenous language digital activism," it certainly was not the first experience with the adoption of digital tools for language promotion.

In 2007, one of the first communities supported by RV was a project called Voces Bolivianas (Bolivian Voices) that trained Indigenous Aymara university students in El Alto, Bolivia to create and maintain their own blog. With the realization that this personal space online could be whatever its creator wanted to see reflected, some of the workshop participants decided that they wanted their blog to be in the Aymara language. It was here that the first Aymara language blog was likely created, which in turn opened up the possibilities for the use of this language in even more spaces online.

Since that early experience with the Aymara language, participants in that project would be seen as pioneers. Many of those participants would go on to form their own collectives, driven by the desire to create more digital content in the Aymara language for use by future generations. With this inspiring others from other language communities, many more projects and campaigns would soon appear. However, it was quite evident that this type of language promotion was already happening across the Americas and beyond — it just needed to be further amplified.

After seeing the positive consequences of the Oaxaca gathering, Rising Voices partnered with allies to convene other language digital activists in Bolivia, Colombia, Ecuador, Guatemala, Chile, and Peru. They followed a similar format, where the participants themselves helped set the agenda, but more importantly they set the stage for peer learning and experience exchange. One meeting in the city of Otavalo, Ecuador, particularly stood out because the gathering focused on a single language, Kichwa, unlike other national gatherings where dozens of languages were represented. The entirety of the Kichwa Digital Activism gathering was conducted entirely in this language, spoken by approximately half a million Ecuadorians. The message was unmistakable: the workshop organizers wanted to demonstrate that the language was fully functional to be used in all contexts, from traditional ceremonies to modern life, but especially in digital contexts. There was no need to borrow words from the Spanish languages; Kichwa was perfectly capable of conveying concepts.

One outcome resulting from these national gatherings has been the formation of networks of mutual support. Since they were able to relate to one another, the participants remained in touch long after the meetings ended. Moving entirely to a virtual setting, these networks would go on to celebrate each other's success and help further amplify these movements.

It became clear that the demand for these types of connections would outnumber the spaces available. For each of the national gatherings, the applications to attend would often be five times the number of spaces available. In order to acknowledge this growing

need, other types of activities were planned to diversify the opportunities available.

With more and more digital projects appearing online, RV created a beta version of an online directory or database to highlight efforts across the region. These examples serve as inspiration for other communities wanting to do something for their language through the use of apps, podcasts, online radio, and localization projects, among other tools. While the directory is far from complete since new projects or campaigns appear all the time, users can search the directory according to country, language, and type of digital platform utilized.

Another of Rising Voices's major activities has been the use of participatory social media accounts, such as @ActLenguas (Language Activism). Started in 2019 to coincide with the International Year of Indigenous Languages, this rotating Twitter account has featured several different Indigenous language digital activists who serve as guest hosts. Each week, they shared their personal stories, as well as information about their motivations, challenges, successes, and other work they recommend. This platform provides a direct line to an audience of nearly 15,000 interested in this topic. Activists from Chile to Mexico have all taken a turn and have had their stories amplified on a global scale. The account was featured by Twitter Mexico as a recommended account, as well as highlighted during the National Fair of Indigenous Languages of Mexico, organized by the Institute of Indigenous Languages of the Mexican government.

Rising Voices's Activismo Lenguas initiative, which coordinates all of these projects and campaigns, was awarded the International Mother Language Award from the International Mother Language Institute of the government of Bangladesh in 2020.

Even though Rising Voices's name was on the Award, the recognition was a tribute to the amazing efforts of all of the digital activists who have taken part in the events, activities, and campaigns organized by Rising Voices. It cements RV's role in the movement as a connector and amplifier in order to facilitate spaces that make peer learning a possibility.

BEYOND LIKES AND FOLLOWS

As the field of digital activism continues to expand, not only do the number of online initiatives designed to promote native languages in digital spaces increase, but so does the growing importance on an international stage. There is a general sense that there is something larger just around the bend.

Yes, anecdotes abound about how inspiring it has been to see one's own language and culture on the internet, and how it has been a constant motivator. For many, it has been enough to validate the effort. Knowing that there are a handful of people that may have been inspired by this work can keep one going for a long duration of time. However, there is also a sense that this must be more than just a curiosity or much more than just headlines in the mainstream media.

How can digital activists know that their work is having the intended impact?

Since much of this emerging digital content is concentrated on private technology platforms, it is difficult to know the full story. These platforms have a unique and enviable vantage point with access to all of the analytics related to language use online. This valuable data would be a treasure trove not only for language communities to gain a better snapshot of how often digital content in their language is being used, by whom, and what for, but for

researchers, policymakers and other allied partners who could use this information to complement traditional data collected by the censuses and other initiatives. Such information could help with strategic planning on how to apply limited time and resources to gain the highest return on investment.

So much could be gained by having access to the data. Distinguishing between a true rise in the use of the language on these platforms and a perceived sense due to the algorithms showing more content in this language could help shape one's approach to digital activism.

Without access to this usage data, language communities must rely on other ways to understand what to make of this effort. Visible stats, such as views, follows, retweets, and downloads could be one way to get a general sense of the reach. When prominent Indigenous TikTokers receive hundreds of thousands, if not millions of views, is this enough a sign that the reach is having an impact? Who is on the other side of those views? Is a child or teenager seeing themselves reflected in these spaces be enough to spur them to take on the language.

Why would demonstrating impact be so important?

The vast majority of Indigenous language digital activists immersed in this field are doing so out of a love and commitment for their language and culture, and they do not need to justify their work to anyone.

However, there is also a stark urgency. Estimates indicate that nearly 40% of the world's languages are considered to be endangered, and if the current rate continues, then 90% of the world's 7000 languages will be extinct over the next century. The resources we have to counter this are limited, making it paramount that we understand the best way to deploy the resources we have in the time we have to push back against this trajectory. Understanding and measuring the impact will give us vital information about what works where and how. Demonstrated impact can also serve to attract more resources to this work.

MAKING AN IMPACT DURING URGENT TIMES FOR INDIGENOUS LANGUAGES

So what can digital activism accomplish when language loss is such a deep-rooted challenge?

The causes of language loss are quite complex, involving historical, political, and cultural factors that span centuries. Certainly, in many parts of the world, well-documented examples of forced displacement or elimination of Indigenous peoples and their languages have had devastating effects felt centuries later. And in modern times, globalization has only created further challenges for Indigenous languages to overcome the influence of dominant languages.

Against these odds, it may seem naive to think that digital activism will drastically alter the fate of these languages. Thinking about it in those terms can be counterproductive, though, since digital activism alone is not the magical solution. However, it can be an important part of an integrated approach to attempt to halt or even reverse language loss.

When one learns about the fundamental motivations of digital activists, there is a desire for change. Whether it is about changing attitudes, motivating young people, or documenting the language for multiple purposes, many digital activists have a clear change they would like to see.

Examining Indigenous language digital activism through the lens of the broader field of digital activism, including, for example, environmental digital activism or digital activism for women's rights, can provide some insights on the difficulties of assessing change. The challenges to measuring impact in digital activism include figuring out what other factors or conditions are affecting the outcome. In a 2010 interview with ReadWrite, one of the early observers of digital activists, Mary Joyce, indicated that "The measuring of impact thus becomes extremely subjective."

That commentary resonates with Indigenous language digital activism because of the wide range of changes or impact that is sought. As mentioned earlier, changes can include raising awareness, changing attitudes, or even just creating the conditions for the use of the language online. Subjectivity in measuring impact allows for greater flexibility, taking into consideration that many digital activists may be engaged in this work without necessarily labeling it as such.

For all those that are satisfied with creating content and highlighting their language in digital spaces, there are many more that want to go further. In a recent Rising Voices survey of 46 experienced Indigenous language digital activists in Latin America, one question asked about their aspirations for future involvement with their language. Of the small sample size, 83% indicated that they want to advance through new roles, including working in the public sector to affect policies, join the academic sector to contribute knowledge, seek a job at a non-governmental organization to work in favor of their language. Only 17% expressed a desire to continue practicing language digital activism on a voluntary basis.

OPPORTUNITIES TO PUSH THE ENVELOPE

With the International Decade of Indigenous Languages set to begin in 2022, we have another golden opportunity to support Indigenous language digital activists and their work. Among the ways to provide support is amplifying their work to raise awareness of their presence on a local, national, and international stage, and convening spaces where they can reflect on the challenges at hand, but also to seek ways to maximize the available resources, including human resources.

Rising Voices has partnered with UNESCO to develop a toolkit resource for language communities interested in adopting digital technologies and adapting them for their self-determined language needs. Instead of providing a recipe, the resource seeks to provide a road map sharing possible paths and the resources necessary to achieve their intended change.

The resource is based on nearly a decade of best practices and lessons learned from the hundreds of digital activists who have shared their work with different networks, events, campaigns, and other research activities. Naturally, new approaches and experiences constantly emerge, and, to help address this new information, we also convened a number of allied partners from around the world and created an advisory group to provide additional perspectives on what is needed for their own language community.

In addition, a global survey was made available in eight languages and provided feedback about aspirations, challenges, and needs for Indigenous language communities. In total, nearly 900 responses were collected, providing insights on questions that speakers of Indigenous languages already have and how such a resource might be useful.

The result of this curation of experiences and best practices, coupled with the sharing of needs from the advisory group, allied partners, and respondents of the survey was to organize the information into tactics. Eight strategies were identified that most digital activism can be organized.

Tactic 1: Facilitating digital communication in Indigenous languages
Tactic 2: Multiplying Indigenous language content online
Tactic 3: Normalizing the use of Indigenous languages online
Tactic 4: Educating in and teaching Indigenous languages online
Tactic 5: Reclaiming Indigenous languages and knowledges
Tactic 6: Imagining and creating new media in Indigenous languages
Tactic 7: Defending spaces for Indigenous languages and linguistic rights
Tactic 8: Protecting linguistic heritage and communities

Each tactic provides an overview of its overall objective and existing examples of digital projects or campaigns employing the strategy. In addition, there are links to existing guides and tutorials, and other resources that can help to employ the strategy. Tips for measuring impact are also included in each module.

The toolkit resource is designed to be a reference document with nearly 150 pages of information. As part of the project, Rising Voices partnered with UNESCO and an organization called P2PU to adapt this resource document into participatory workshops sessions. With the COVID-19 pandemic still preventing many in-person gatherings, RV organized a series of workshops consisting of four sessions with additional readings in between sessions. Within each workshop cohort, there may have been 10-15 different language communities represented. Even though the presentation of tactics and challenges were some of the learning materials shared, it was the small breakout groups that seemed to be uplifting.

Similar to the experience from the first gathering of Indigenous Language Digital Activists in Mexico in 2014, encountering like-minded people to share aspirations and current needs turned out to be an inspiration for many of the participants. For emerging digital activists in Nigeria, Ghana, India, Central America, South America, and Southeast Asia, the sessions were a chance to reflect on the challenges at hand, exchanging stories of personal motivations, and exploring what it means to have an impact or seek change, all in a safe space among peers.

It is by facilitating this kind of knowledge sharing between and within Indigenous language communities that we can strengthen their resilience and their ability to occupy and retain digital space. That can become the foundation of further efforts to halt and reverse language loss, and provide a base from which to demand/lobby for space in policy conversations to make further concrete plans.

NEXT STEPS

Going forward, as had been recommended by the Chapoltepek [Los Pinos] Declaration of 2020, there is a need to support peer learning networks of digital activists. There is no better time than the present to demonstrate to Indigenous language digital activists that their work matters, sending the message that their activities are having an impact as they themselves define. Some of the ways to do that include continuing to amplify their

work to inspire others and convening spaces for peer learning, but also providing learning resources that can help them reflect on their own work. With at least ten years to further collaborate with the start of the International Decade of Indigenous Languages 2022–2032, the current groups of digital activists can be instrumental in encouraging and inspiring a new generation of digital activists to keep these efforts going.

TECNOLOGIES DE LA LLENGUA I REVITALITZACIÓ LINGÜÍSTICA (RISING VOICES)

Eddie Avila

Durant tres dies a l'octubre de 2014, es van reunir 30 parlants de llengües indígenes de Mèxic a la biblioteca Juan de Córdova al centre d'Oaxaca, la capital d'un estat conegut per la seva diversitat cultural i lingüística. Aquesta ocasió en concret va ser la primera trobada d'aquest tipus; i reunia defensors lingüístics que utilitzen el poder i potencial de l'Internet, els mitjans digitals, i d'altres tecnologies de forma activa per promoure i compartir les seves llengües natives a espais digitals.

Per a molts, aquest tipus de "no-conferència" va ser una revelació. Els mateixos participants es van oferir a liderar tallers i organitzar debats. Molts van comentar que mai haurien imaginat que poguessin haver-hi altres persones com ells que adoptessin aquestes estratègies digitals per assegurar-se que la seva llengua i cultura estiguessin presents a Internet.

És clar, feia anys que les eines digitals ja jugaven un paper significatiu a la documentació i promoció lingüística molt abans d'aquesta primera reunió. Tanmateix, és probable que la reunió a Oaxaca l'any 2014 fos el primer cop que s'hagin descrit aquests tipus d'activitats com a "activisme digital per les llengües indígenes".

Es va utilitzar aquest terme expressament, no com una provocació, sinó per reconèixer que, per avançar cap al canvi, cal prendre passos deliberats en aquesta direcció. Sens dubte, a alguns cercles d'arreu del món, el terme "activista" podria deixar amb l'aigua al coll als que no volen ésser considerats agitadors. No obstant, com molts participants han apuntat, perquè s'esdevinguin canvis a nivell local, nacional o fins i tot internacional, cal desafiar l'statu quo d'alguna manera. En aquest cas, l'statu quo normalment volia dir que les llengües indígenes perdien prestigi davant de governs o empreses tecnològiques que no proporcionen les condicions adequades perquè les llengües indígenes puguin desenvolupar-se.

Alguns activistes digitals volen ajudar a originar canvis com veure més continguts digitals disponibles en la seva llengua per poder despertar l'interès de les generacions més joves i animar-les a mantenir les seves llengües maternes.

Des d'aquella primera trobada, cada cop més persones se senten identificades amb aquest terme perquè hi troben certa afinitat amb d'altres que treballen per canvis similars. El concepte també va agafar impuls a plataformes internacionals, i forma part de les recomanacions per a la propera Dècada Internacional de Llengües Indígenes 2022-2032. La Declaración de Los Pinos (Chapoltepek) exposa la necessitat de:

> *Promoción de redes de activistas y defensores digitales de la enseñanza y el aprendizaje de lenguas indígenas, así como del intercambio de mejores prácticas relacionadas con el uso de la tecnología.*

Encara que pugui resultar satisfactori veure com se cita la feina d'activistes digitals per les llengües indígenes a aquests espais, i l'atenció que aquesta feina ha rebut a la premsa local, nacional i internacional, aquesta obra ha crescut fins a tal punt que només aquest reconeixement ja no és suficient. En canvi, aquest moviment cerca tenir una major influència per repercutir canvis positius.

Tot i això, com és aquest canvi? Com l'impacte en les llengües afectades per l'activisme digital pot anar més enllà de senzilles mètriques a Internet? I qui hauria de decidir què significa tenir èxit?

EL PAPER DE RISING VOICES DINS D'AQUEST MOVIMENT

Rising Voices (RV), la secció d'igualtat i inclusió digital de l'organització Global Voices, va ser un dels tres catalitzadors de la Primera Trobada d'Activisme Digital per les Llengües Indígenes celebrada a Oaxaca l'any 2014. Tot i que potser va ser la primera activitat realitzada sota el paraigües de "l'activisme digital per les llengües indígenes", certament no va ser la primera experiència adoptant eines digitals per a la promoció lingüística.

L'any 2007, una de les primeres comunitats recolzades per RV va ser un projecte anomenat Voces Bolivianas que formava a estudiants universitaris indígenes aimares a El Alto, Bolívia, sobre com crear i mantenir el seu propi blog. A l'adonar-se que aquest espai personal en línia podia ser el que ells volguessin, alguns participants van decidir que volien que el seu blog estigués en la llengua aimara. Segurament va ser llavors quan va crear-se el primer blog en aimara, fet que va obrir la porta a que s'utilitzés aquesta llengua a encara més espais en línia.

Gràcies a aquella primera experiència amb la llengua aimara, els participants d'aquell projecte van ser considerats uns pioners. Molts d'aquells participants van passar a crear els seus propis col·lectius, impulsats pel desig de crear més continguts digitals en la llengua aimara per les generacions futures. Això va inspirar a més persones d'altres comunitats lingüístiques, i aviat van sorgir molts més projectes i campanyes. Tanmateix, era força evident que aquest tipus de promoció lingüística ja estava passant al continent americà i més enllà, simplement calia difondre-la més.

Després de veure les experiències positives de la trobada a Oaxaca, Rising Voices va convocar d'altres activistes digitals lingüístics a Bolívia, Colòmbia, l'Equador, Guatemala, Xile i el Perú. Seguien un format similar, on els mateixos participants ajudaven a establir l'ordre del dia, però per sobre d'això, deixaven espai per l'aprenentatge entre iguals i l'intercanvi d'experiències. Una reunió a la ciutat d'Otavalo, Equador, va destacar especialment perquè la trobada se centrava en una única llengua, el kichwa, a diferència d'altres trobades nacionals on s'utilitzaven dotzenes de llengües. Tota la trobada d'Activisme Digital en Kichwa es va dur a terme en aquesta llengua, parlada per aproximadament mig milió d'equatorians. El missatge era clar: els organitzadors volien demostrar que la llengua podia utilitzar-se en tots els contextos, des de cerimònies tradicionals fins a activitats a la vida moderna, però especialment en contextos digitals. No calia manllevar paraules de la llengua espanyola; el kichwa podia transmetre tots els conceptes perfectament.

Un resultat d'aquestes trobades nacionals ha estat la formació de xarxes de suport mutu. Com podien empatitzar els uns amb els altres, els participants van continuar en contacte més enllà de les reunions. En traslladar-se a un escenari completament virtual, aquestes xarxes van utilitzar-se per celebrar els èxits dels companys i ajudar a difondre encara més aquests moviments.

Es va fer evident que la demanda d'aquests tipus de connexions superaria l'oferta de places. A cadascuna de les trobades nacionals, les sol·licituds de participació sovint eren cinc cops el nombre de places. Per reconèixer aquesta necessitat creixent, es van planejar altres tipus d'activitats per diversificar les oportunitats disponibles.

Amb l'aparició de cada cop més projectes digitals a l'Internet, RV va crear una versió beta d'un directori o base de dades en línia per destacar les iniciatives a tota la regió. Aquests exemples serveixen d'inspiració per a d'altres comunitats que volen ajudar a la seva llengua a través d'aplicacions, podcasts, ràdio en línia, i projectes de localització, entre d'altres eines. Tot i que el directori no és exhaustiu, ja que sovint apareixen nous projectes o campanyes, els usuaris poden filtrar-los per país, llengua i tipus de plataforma digital utilitzada.

Una altra de les activitats principals de Rising Voices ha estat l'ús de comptes participatius a xarxes socials, com @ActLenguas (Activismo digital de lenguas indígenas). Iniciat l'any 2019 per coincidir amb l'Any Internacional de les Llengües Indígenes, diferents activistes digitals per les llengües indígenes han gestionat aquest compte de Twitter per torns. Cada setmana, compartien les seves històries personals, a més d'informació sobre les seves motivacions, reptes, èxits i altres obres que recomanen. Aquesta plataforma dona accés directe a un públic de gairebé 15.000 persones interessades en aquest tema. Activistes des de Xile fins a Mèxic han tingut el seu torn i han pogut difondre les seves històries a escala global. El compte va ser recomanat per Twitter México, a més d'aparèixer a la Feria de las Lenguas Indígenas Nacionales de Mèxic, organitzada per l'Institut Nacional de Llengües Indígenes del Govern de Mèxic.

La iniciativa Activismo Lenguas de Rising Voices, que coordina tots aquests projectes i campanyes, va guanyar l'*International Mother Language Award* [premi de llengua materna internacional] de l'*International Mother Language Institute* [Institut de llengua materna internacional] del Govern de Bangladesh l'any 2020.

Tot i que Rising Voices va ser l'entitat guardonada, el premi reconeixia els esforços de tots els activistes digitals que han participat als esdeveniments, les activitats i les campanyes organitzades per Rising Voices. Aquest reconeixement consolida el paper de RV dins el moviment com a institució connectora i difusora d'espais que possibiliten l'aprenentatge entre iguals.

MÉS ENLLÀ DE M'AGRADES I SEGUIDORS

A mesura que se segueix expandint l'activisme digital, no només augmenten el nombre d'iniciatives dissenyades per promoure les llengües natives a espais digitals, també ho fa la seva importància creixent a l'escenari internacional. S'intueix que alguna cosa més gran es troba molt a la vora.

És cert que hi ha moltes anècdotes sobre com d'inspirador ha sigut veure la llengua i cultura pròpies a l'Internet, i com això ha estat una motivació constant. Per a molts, ha sigut suficient per a que l'esforç valgués la pena. Saber que aquesta feina pot haver inspirat a unes poques persones pot mantenir la motivació per continuar durant molt de temps. Tanmateix, també sembla que això ha d'anar més enllà d'ésser una curiositat o només titulars als mitjans convencionals.

Com poden els activistes digitals saber que la seva feina té l'impacte pretès?

Com que molts d'aquests continguts digitals emergents es concentren a plataformes tecnològiques privades, es fa difícil conèixer-ne l'abast. Aquestes plataformes tenen una perspectiva privilegiada única i envejable de totes les analítiques relacionades amb l'ús de les llengües a Internet. Aquestes valuoses dades serien un tresor no només per a les comunitats lingüístiques que podrien tenir una millor idea de la freqüència amb la qual s'utilitzen els continguts digitals en la seva

llengua, qui els utilitza i amb quina finalitat, sinó també per a investigadors, legisladors i d'altres col·laboradors que podrien utilitzar aquesta informació per complementar les dades tradicionals recollides als censos i a d'altres iniciatives. Aquesta informació podria ajudar a planificar una estratègia amb temps i recursos limitats per obtenir la màxima rendibilitat.

Tenir accés a aquestes dades seria de valor incalculable. L'enfocament de l'activisme digital es beneficiaria de la distinció entre un creixement real en l'ús de la llengua a aquestes plataformes i una falsa percepció causada pels algoritmes que mostren més continguts en aquesta llengua.

Sense accés a aquestes dades d'ús, les comunitats lingüístiques han de dependre d'altres maneres d'entendre aquest esforç. Les estadístiques visibles com les vistes, els seguidors, els retuits, i les descàrregues podrien ser una manera de tenir una idea general del seu abast. Quan *TikTokers* indígenes reben centenars de milers, o fins i tot milions de vistes, és senyal suficient de que estan tenint impacte? Qui es troba a l'altra banda d'aquestes vistes? És suficient que un infant o adolescent es vegi reflectit a aquests espais per animar-los a fer servir la llengua?

Perquè és tan important demostrar impacte?

La gran majoria dels activistes digitals per les llengües indígenes immersos en aquest camp ho fan des de l'amor per i el compromís amb la seva llengua i cultura, i no necessiten justificar la seva feina davant de ningú.

Tanmateix, també existeix una clara urgència. Les estimacions indiquen que gairebé el 40% de les llengües del món es considera en perill, i si la tendència actual continua, el 90% de les 7.000 llengües del món s'extingirà durant el proper segle. Els recursos disponibles per contrarestar això són limitats, per tant és d'importància cabdal comprendre la millor manera d'implementar els recursos dels que disposem dins el període de temps que tenim per fer retrocedir aquesta trajectòria. Comprendre i mesurar l'impacte ens donarà informació vital sobre què funciona a on i com ho fa. Demostrar-ne l'impacte també pot servir per atraure més recursos a aquesta tasca.

TENIR IMPACTE EN MOMENTS D'URGÈNCIA PER A LES LLENGÜES INDÍGENES

Llavors, què pot aconseguir l'activisme digital quan la desaparició de les llengües és un repte tan profundament arrelat?

Les causes de la desaparició de les llengües són força complexes, i inclouen factors històrics, polítics i culturals presents durant segles. Certament, a moltes regions del món, existeixen casos ben documentats del desplaçament forçós o l'eliminació de pobles indígenes i les seves llengües, dels quals se senten els efectes devastadors segles més tard. En l'actualitat, la globalització ha dificultat encara més que les llengües indígenes puguin superar la influència de les llengües dominants.

Amb aquest pronòstic, pot semblar ingenu pensar que l'activisme digital té la capacitat de canviar decisivament el destí d'aquestes llengües. Pensar així pot resultar contraproduent però, perquè l'activisme digital per si sol no és la solució màgica. Tanmateix, pot formar part d'un conjunt de mesures per intentar frenar o fins i tot invertir la desaparició de la llengua.

Entre les motivacions fonamentals dels activistes digitals, hi ha la d'un desig de canvi.

Tant si es tracta de canviar actituds, motivar la gent jove, o documentar la llengua per qualsevol motiu, molts activistes digitals tenen en ment que volen veure un canvi clar.

Examinar l'activisme digital per les llengües indígenes des de la perspectiva més àmplia de l'activisme digital, incloent, per exemple, l'activisme digital pel medi ambient o l'activisme digital pels drets de les dones, pot donar informació sobre les dificultats d'avaluar els canvis. Els reptes de mesurar l'impacte en l'activisme digital inclouen investigar quins altres factors o condicions afecten el resultat. A una entrevista l'any 2010 amb ReadWrite, una de les primeres observadores dels activistes digitals, Mary Joyce, va indicar que mesurar l'impacte, doncs, esdevé extremadament subjectiu.

Aquest comentari està en consonància amb l'activisme digital per les llengües indígenes donada l'àmplia varietat de canvis o l'impacte cercats. Com ja s'ha comentat, els canvis poden incloure crear consciència, canviar actituds, o simplement crear les condicions per l'ús de la llengua a Internet. La subjectivitat en la mesura de l'impacte permet major flexibilitat, tenint en compte que molts activistes digitals poden estar involucrats en aquesta tasca sense necessàriament anomenar-la així.

Encara que hi ha moltes persones que es conformen amb crear continguts i fomentar la seva llengua a espais digitals, existeixen moltes d'altres que volen donar un pas més. A una enquesta recent de Rising Voices a 46 activistes digitals per les llengües indígenes a Llatinoamèrica, en un apartat se'ls preguntava sobre les seves intencions d'implicar-se amb la seva llengua en el futur. D'aquesta petita mostra, el 83% indicava que volia avançar cap a nous rols, incloent treballar al sector públic per influir en política, unir-se al sector acadèmic per contribuir amb coneixements, i cercar feina a una organització no governamental per treballar a favor de la seva llengua. Només el 17% expressava un desig de continuar practicant l'activisme digital lingüístic de forma voluntària.

OPORTUNITATS PER ARRIBAR MÉS ENLLÀ

Amb la Dècada Internacional de Llengües Indígenes que començarà l'any 2022, tenim una altra excel·lent oportunitat per donar suport als activistes digitals per les llengües indígenes i la seva tasca. Per donar-hi suport, podem difondre la seva tasca per crear consciència sobre la seva presència a l'escenari local, nacional i internacional, i crear espais on poder reflexionar sobre els reptes que els ocupen, però també cercar maneres de maximitzar els recursos disponibles, entre ells els recursos humans.

Rising Voices s'ha associat amb la UNESCO per desenvolupar un joc d'eines per a comunitats lingüístiques interessades en adoptar tecnologies digitals i adaptar-les a les seves pròpies necessitats lingüístiques. Enlloc de donar instruccions concretes, aquest joc d'eines vol ser un full de ruta i compartir els recursos necessaris i les possibles vies per aconseguir el canvi desitjat.

Aquest instrument es basa en gairebé una dècada de bones pràctiques i lliçons apreses dels centenars d'activistes digitals que han compartit la seva tasca en diferents xarxes, esdeveniments, campanyes i d'altres activitats de recerca. Per suposat, sorgeixen noves visions i experiències constantment, i, per ajudar a tractar aquesta nova informació, es van reunir alguns col·laboradors d'arreu del món que van crear un grup assessor per donar diferents perspectives sobre les necessitats a les seves pròpies comunitats lingüístiques.

A més, es va oferir una enquesta global disponible en vuit idiomes que recollia les aspiracions, els reptes i les necessitats de les comunitats de llengües indígenes. En total es van recollir prop de 900 respostes que proporcionen informació sobre les preocupacions dels parlants de llengües indígenes i com d'útil podria ser aquest tipus d'instrument.

La finalitat d'aquesta col·lecció d'experiències i bones pràctiques, conjuntament amb les necessitats compartides pel grup assessor, els col·laboradors i els enquestats, era organitzar la informació en tàctiques. Es van identificar vuit estratègies segons les quals es pot organitzar gran part de l'activisme digital.

Tàctica 1: Facilitar la comunicació digital en llengües indígenes
Tàctica 2: Multiplicar els continguts en llengües indígenes a Internet
Tàctica 3: Normalitzar l'ús de llengües indígenes a Internet
Tàctica 4: Educar en i ensenyar llengües indígenes en línia
Tàctica 5: Recuperar coneixements i llengües indígenes
Tàctica 6: Idear i crear nous mitjans de comunicació en llengües indígenes
Tàctica 7: Defensar els espais per les llengües indígenes i els drets lingüístics
Tàctica 8: Protegir el patrimoni lingüístic i les comunitats

Cada tàctica dona una visió general del seu objectiu principal i exemples de projectes o campanyes digitals que implementen l'estratègia. A més, inclou enllaços a guies i tutorials, i d'altres recursos que poden ajudar a implementar l'estratègia. Cada mòdul també inclou consells sobre com mesurar l'impacte.

El joc d'eines està pensat per ser un document de referència i consta de gairebé 150 pàgines d'informació. Dins del projecte, Rising Voices es va associar amb la UNESCO i una organització anomenada P2PU per adaptar aquest document convertint-lo en tallers participatius. Com la pandèmia del COVID-19 encara impedeix moltes trobades presencials, RV va organitzar una sèrie de quatre tallers amb lectures addicionals entre sessions. Cada grup dels tallers incloïa entre 10 i 15 comunitats lingüístiques diferents. Tot i que es van compartir aquestes tàctiques i reptes com a material educatiu, sembla que van ser els grups reduïts el que va resultar més inspirador.

De forma similar a la primera trobada d'Activistes Digitals per les Llengües Indígenes a Mèxic l'any 2014, trobar-se amb persones del mateix parer amb qui compartir aspiracions i necessitats va inspirar a molts participants. Per a activistes digitals emergents a Nigèria, Ghana, l'Índia, l'Amèrica Central, l'Amèrica del Sud i el Sud-est Asiàtic, les sessions eren una oportunitat per reflexionar sobre els seus reptes, intercanviar històries i motivacions personals, i esbrinar què significa tenir impacte o cercar el canvi, tot això en un espai segur entre iguals.

Facilitar aquest tipus d'intercanvi de coneixements entre i dins de les comunitats de llengües indígenes és la manera d'enfortir la seva resiliència i la capacitat d'ocupar i mantenir la seva posició a espais digitals. Això pot convertir-se en els fonaments per seguir frenant i invertint la desaparició de les llengües, i així poder reclamar el seu dret a participar en converses polítiques per poder concretar estratègies futures.

ELS SEGÜENTS PASSOS

En el futur, com recomana la Declaración de Los Pinos (Chapoltepek) de 2020, caldrà donar suport a les xarxes d'aprenentat-

ge entre activistes digitals. Ara és el millor moment per demostrar als activistes digitals per les llengües indígenes que la seva feina importa i que les seves activitats tenen l'impacte que ells mateixos defineixen. Algunes maneres de fer-ho són continuar difonent la seva tasca per inspirar a d'altres i crear espais per a l'aprenentatge entre iguals, però també proporcionar recursos educatius que els poden ajudar a reflexionar sobre la seva pròpia feina. Com queden almenys deu anys per seguir col·laborant gràcies a l'inici de la Dècada Internacional de Llengües Indígenes 2022-2032, els grups actuals d'activistes digitals poden ser fonamentals per animar i inspirar a les noves generacions d'activistes digitals a continuar amb aquestes iniciatives.

BEYOND TECHNOLOGICAL SOLUTIONS: HOW WE CREATE A WORLD THAT SUSTAINS ITS LANGUAGES

Steven Bird
Charles Darwin University, Australia

Have you ever run across text in the "wrong" language then with a copy-paste and a click you could have it in the "right" language and get on with your work? If so, you just embraced a theory of Language, a theory that human language is a tool for communication. And what about your brief exposure to another language? Each time you clicked that translate button you reinforced the tacit assumption that linguistic diversity is an obstacle to be solved by technology.

Pause for a moment to imagine how speakers of endangered languages theorise about language? A Usarufa man once said to me: "if we stop speaking our language the Kamano people will chase us off our land." A Kunwinjku woman told me: "I can't tell that story, it's from Kudjekbin country." The theory? Language is identity, country, title to land.

We always knew language was more than a tool. For example, poetry is sometimes defined as language that cannot be translated. We see it with individual words: you can't just translate Portuguese *saudade* as "longing" or Dutch *gezellig* as "cozy." The nuances are lost.

This failure of translation is often a source of pride. Speakers may explain that their language offers its own way of seeing the world. This is especially striking in the case of endangered languages.

Carrier, a language of British Columbia (pop. 600) has a word *k'onih'azi* which translates as "newly-wed beaver couple." Dalabon, from Arnhem Land in Northern Australia (pop. <10), *dalabborrord* means "the place on a tree where two branches rub together." In Nootka, a language of Vancouver Island (pop. 130), *ši·ša·wi·taqyo* is translated as "powered by a monstrous supernatural porcupine-like creature." Thus, a small local language is coupled with a way of being in the world. Its stock of words hints at people's preoccupations and worldviews.

So you see, it's not just a few nuances. Cosmology is lost in translation. We cannot "save languages" merely by recording and translating them.

There are other ways to sustain the 4,500+ languages that are still vigorous. We can seek to address the threats they face. We can nurture the cultural ecosystems that enable languages to thrive. We can create entirely new spaces for minority languages in our towns and cities. We can create a world that sustains its languages.

This chapter suggests concrete actions that you can take to help create this future world. You don't need to know another language. You don't need to join a campaign. But you do need to have the nerve to do something personal and risky, and enter the vulnerable place of being in the minority. Are you game?

GREET PEOPLE IN THEIR LANGUAGE

Do you cross paths with someone whose first language is not the same as yours? Or perhaps, not your dialect? You might see her at work, school, the park, a supermarket. Your challenge is to elicit a simple greeting like "hello," preferably one that works at any time of day.

Each time you see her, use the greeting. Notice any effects this has on your connection. Soon you might be finding out more about the language, the culture, and the local community. You could use a flashcard app to help you to memorise greetings in several languages.

If you're a school teacher, why not learn greetings in the home languages of your students? You could try them in class, school assembly, or with parents.

Note that some people don't like to be publicly recognised as speaking a foreign language. Take care not to make anyone feel uncomfortable, exoticised, or a subject of your showing off. Using a greeting as you pass someone in the hallway is different to calling out across the room at a public meeting. Also note that, in some languages, greetings vary depending on personal status or direction of movement. Speakers might have to compromise to work out a greeting that you can use.

Hello!
Anyi paranga ra (Ma'di)
Te aso tokereka (Takū)
Ngudda kamak (Kunwinjku)
Abilaki (Eskayan) Dahooja (Carrier)
Wú cjêew (Shilluk)
Aŋpétu wašté (Lakota) Masikati (Shona)
Palya (Pitjantjatjara) Gude (Tok Pisin)

LEARN TO PRONOUNCE PEOPLE'S NAMES

> "Jo ... Ja ... Joh-von. Ja-Va. Ah, f*ck it, we'll call you Joe."
>
> –Alec Patric, *Black Rock White City*

Your friend has a foreign name and you suspect you don't say it correctly. Is it a problem for her when people mispronounce her name? Would she like you to learn to say it properly? Prepare for this activity by downloading a voice recorder app.

1. Ask your friend, "can I try something?" Then open your voice recorder app and say your friend's name as best you can. (This makes it obvious that you're recording, and will help you notice pronunciation differences later.)
2. Look to your friend for correction. Put the phone nearer to her and ask her to say it again, slowly.
3. Play back the recording. This shows her what you captured. Try saying her name correctly and look to your friend for correction. Notice how she moves her lips. Ask her to correct you in future if you don't get it right.
4. Later on your own, listen and practice. The goal is improvement, not perfection! Rename the recording so you can find it easily.

Once you've had a bit of practice, try this with colleagues or acquaintances. If you're a teacher, ask children or their parents to pronounce their names while you record. You could make a class activity out of pronouncing names. You could ask someone with a hard-to-pronounce name to say all the different versions he has heard people use. Verbalising the incorrect version helps children hear the differences. Note that some people prefer others to use a local version of their name; my friend Ruprecht asks English-speaking friends to call him Rupert.

PARTICIPATE IN A LOCAL COMMUNITY FESTIVAL

Many cities have cultural minorities that hold annual festivals. These are typically single-culture single-language events, open to the public but usually only attended by members of the community.

Take yourself along! Ask if you can watch. There'll be cuisine and costume in abundance. Your task is to pick up some of the language. Find someone to teach you a greeting then use it with other people. There might be a word on a sign and you can ask what it means and how to say it. Buy something from a food stall then ask how to say "I would like..." Use this expression at another stall, and point at the same time. See if you can pick up the culturally appropriate way of pointing; it might use the eyes, the lower lip, or a nod of the head.

You may be invited to further events. As you connect with community members, try to receive generosity and openness without feeling the need to make it into a transaction. Resist the temptation to solve problems, to intervene, to campaign, or to throw money at a situation. Avoid being cast in the mould of a teacher of the local majority language.

Focus your limited time and effort on openness and connection. Learn more names and greetings. Memorise a popular song in the language. Find out what people think about your interest. What's their theory of language? Are they concerned about keeping their language strong? Remember that the simple fact of your presence and interest is a powerful act of recognition. You don't need to do much more.

RECONNECT WITH YOUR HOME LANGUAGE

Did you have another language at home when you were growing up? You might have relatives who still speak the language, like an old aunt who lives nearby, or a cousin back in the home country. Could you try using your original home language again?

What would it be like to speak it well? You might have a different personality in this language. You could develop a new connection with the wider family. You might hear stories about what your parents were like when they were young.

Look into the possibility of attending night or weekend classes. Arrange conversation practice, even online. Seek out opportunities to hear the language. It won't matter that you don't understand everything at first. Immerse yourself – it's a great way to learn!

Listening to an older person share stories in her language (Samarkand)

RAISE BILINGUAL CHILDREN

The benefits of bilingualism are well-known. Many bilingual children develop better social cognition, a deeper understanding that others see the world differently. As bilingual adults they may be better world citizens

and better able to learn a third or fourth language. In old age, bilinguals have been found to experience slower cognitive decline and delayed onset of dementia. There are no known disadvantages to a bilingual upbringing.

In spite of this, parents who have grown up with another language often think they should speak the dominant language at home. How else will their children learn it? However, children become fluent in the dominant language regardless of what language is spoken at home. Schools are starting to recognise this, and some run language immersion programs. Is there one near you?

Imagine what it would be like for a child to stay strongly connected with her family's origins while becoming a full participant in our society and economy? What conversations and connections are easier when parent and child are fluent in the same language?

If you decide to raise bilingual or multilingual children, you may face resistance. It helps if you have two languages as a normal part of your family life and if your children are motivated to speak them. It is also good if your children hear the language from other people and other sources such as books, videos, songs, and the Internet.

SPEAK THE ORIGINAL LANGUAGE OF YOUR PLACE

The original language of a place has status, simply because it was there first. The land you stand on has been hearing this language for centuries! This is why Sarah Palin said: "If you're in America, speak American!"

Palin was speaking in defence of English as though its future was threatened. It is worth trying to imagine what it would be like if your mother tongue was endangered. We are going to adopt Palin's language for our life-affirming purposes. *Palin's Principle: Learn to speak the original language where you live.*

What is the original language of your area? Is it still spoken? Can you learn some words? Can you enrol in a course?

Your goal is not fluency. Mastering a language is like mastering a profession or sport or musical instrument. It takes about 10,000 hours! Still, you can be on a journey of discovery, learning useful words and phrases, finding out what placenames mean, and the associated stories.

You could adopt this language as a mascot language at your work or school, with posters, hosted visits, naming of year levels, and so on.

Remember that speakers of small languages generally do not think of their language as a tool that others can just pick up and use for their own ends. Their language is an intimate part of their identity. You will need to build trust and seek permission from the right people.

PLAY LANGUAGE GAMES

There are many language games you can try. Here are just a few of the games described in more detail at languageparty.org.

Foursquare Hello: This is a version of the children's game where we use greetings in each others languages at the same time as passing a ball. If you make a mistake you go back to bottom position. The goal is to reach King Polyglot position and fend off all challengers.

Garden of Words

Cacophony Line

Hip Hello: We learn the hippest slang greetings, with facial expressions and hand gestures... be ultra-cool in another language with a single phrase!

Garden of Words: Divide into pairs. The "sculptors" think of an untranslatable word from their first language such as an idea or emotion and express it by moving their partners, the "clay," into position. We display the foreign language words at their feet and then wander around the sculpture garden and guess the meanings.

MorphoLogical: A game where we apply some of the world's strangest word-formation rules to invent new words, then introduce them into casual conversation.

Cacophony Line: Four volunteers stand at the front with their backs to the audience, and turn around at random to tell a story in their language, stopping when the next person turns. This is a hilarious fast-paced language game.

440 – Four People Four Languages Zero Barriers: Four people have an improvised conversation in four languages, reading each other's facial expressions and body language to decide how to respond.

THROW A LANGUAGE PARTY!

It is nothing short of miraculous that our world is home to over 4,500 vigorous languages. How can this be after centuries of colonialism, nationalism, globalism, and worst of all, mockery and put-downs by people who speak the dominant language? How better to respond to this news than by throwing a party? A language party!

It is time to get together with your new friends to celebrate the world's linguistic diversity. Gather people together and experience stories in the way they have travelled down the generations: in spoken language.

The format is simple: invite people to share a 3-5 minute story in their first language then explain in the majority language. Encourage folklore in preference to narratives of trauma and displacement. You'll be surprised how readily speakers of small languages are able to tell good stories! You could ask for songs as well.

The group needs to be prepped. As host you need to encourage people not to *listen-to-understand* but rather to *listen-to-appreciate*. This is language as art, music, spoken soul. No-one will understand everything

Four People Four Languages Zero Barriers

that is said, but everyone can listen to how each language sounds, paying attention to its rhythms and melodies, to gestures and facial expressions, and guess what each story is about.

Before each story, ask storytellers to teach everyone a greeting. Practice it until people say it correctly. Then ask the storyteller to open the story with this greeting. You can find out more about this format of storytelling at languageparty.org.

WHEN YOU LEAVE YOUR COMFORT ZONE...

When you try these activities you will feel vulnerable. You are reaching out to people in ways they do not expect. They may be suspicious of your motives. They may read your discomfort, provoking their own. Remember that you're trying something new. Like learning to ride a bicycle, there are skills to develop. Don't give up the first time you fall!

Something else is also going on: you're bucking a trend. You're trying to connect with people who may have experienced a lifetime of alienation by your culture. Perhaps they are not instantly grateful that you decided to notice them. You might have caught them at a bad moment, or needed a culturally-appropriate introduction.

When things don't work out, the alienation you feel is a reminder of the alienation that anyone in a minority feels when they try to fit into the dominant culture and are rebuffed, laughed at, or ignored. You are doing this voluntarily and can retreat to your comfort zone at any moment. What would it be like to have no escape?

Connecting across entrenched and invisible barriers is difficult. But it gets easier as you discover friendly people, build trust, find your

groove, and stop worrying about embarrassing yourself. Remember, you are helping to create new ways and new places for people to belong. A special reward awaits. Someone else will come to belong in your place in a new way. *You*.

FURTHER READING

Austin, Peter K (2008). *One Thousand Languages: Living, Endangered, and Lost.* University of California Press.

Evans, Nicholas (2009). *Dying Words: Endangered Languages and What They Have to Tell Us.* Blackwell.

Grosjean, François (2009). What parents want to know about bilingualism. *The Bilingual Family Newsletter*, 26(4), 1-6. francoisgrosjean.ch/for_parents_en.html

Hinton, Leanne (2001). How to Keep Your Language Alive: A Commonsense Approach to One-On-One Language Learning. Heyday.

OTHER RESOURCES

languageparty.org, untranslatable.org, wikitongues.org, elalliance.org, livinglanguages.org.au, ilinative.org, languageconservancy.org, ethnologue.com, psychologytoday.com/blog/life-bilingual, multilingualliving.com, bilingualism-matters.org

ACKNOWLEDGEMENTS

I am indebted to Manuel Maqueda, Robyn Perry, Nadia Chaney, and Michael Margolis for helping to shape the ideas presented here. Thanks to Lauren Gawne, Antonella Sorace, and Hakan Seyalioglu for feedback on earlier drafts. Special thanks to the speakers in our language parties, for your courage to step out on stage and share a story in your heart language. We were with you all the way.

This article originally appeared in Seyalioglu and Hymes (eds). Dialect, A Game About Language and How it Dies, Thorny Games, 2018.

MÉS ENLLÀ DE LES SOLUCIONS TECNOLÒGIQUES: COM CREAR UN MÓN QUE SOSTINGUI LES SEVES LLENGÜES

Steven Bird
Universitat Charles Darwin, Austràlia

T'has trobat mai amb un text en la llengua "equivocada", i llavors copiant i enganxant amb un clic el podies tenir en la llengua "correcta" i seguir amb la teva feina? Si és així, acabes d'acceptar una teoria de la Llengua, la teoria de que la llengua humana és una eina comunicativa. I què passa amb la teva breu exposició a una altra llengua? Cada cop que cliques el botó de "traduir", reforces la suposició tàcita que la diversitat lingüística és un obstacle que la tecnologia ha de resoldre.

Atura't a pensar un moment sobre com els parlants de llengües en perill teoritzen sobre la llengua. Un home usarufa em va explicar una vegada: "si deixem de parlar la nostra llengua, el poble kamano ens farà fora de la nostra terra". Una dona kunwinjku em va dir: "No puc explicar aquella història, és de la nació kudjekbin". Quina és la teoria? Que la llengua és identitat, nació, títol de propietat.

Sempre hem sabut que la llengua és més que una eina. Per exemple, a vegades es defineix la poesia com un llenguatge intraduïble. Ho veiem amb paraules individuals: no es pot simplement traduir el *saudade* portuguès per "nostàlgia" o el *gezellig* neerlandès per "acollidor". Es perden els matisos.

Aquesta intraduïbilitat sovint és motiu d'orgull. Els parlants expliquen que la seva llengua ofereix una visió pròpia del món. Això és especialment sorprenent en el cas de les llengües en perill.

El carrier, una llengua de la Columbia Britànica (600 parlants) té una paraula, *k'onih'azi*, que traduïda significa "parella de castors noucasats". En dalabon, d'Arnhem Land al nord d'Austràlia (<10 parlants), *dalabborrord* significa "el lloc a un arbre on es freguen dues branques". En nootka, una llengua de la illa de Vancouver (130 parlants), *šiˑšaˑwiɬ-taqyo* es tradueix com "impulsat per una criatura monstruosa sobrenatural semblant a un porc espí". Per tant, una llengua local minoritària està connectada a una manera de ser al món. El seu vocabulari dona indicis de les preocupacions dels seus parlants i la seva visió del món.

Així doncs, no es tracta simplement de matisos. Durant el procés de traducció, es perd la cosmologia. Per "salvar les llengües", no n'hi ha prou amb simplement gravar-les i traduir-les.

Hi ha d'altres maneres de sostenir les més de 4.500 llengües que segueixen actives. Podem intentar ocupar-nos de les amenaces a les quals s'enfronten. Podem nodrir els ecosistemes culturals que permeten que les llengües prosperin. Podem crear espais completament nous per a les llengües minoritàries als nostres pobles i ciutats. Podem crear un món que sostingui les seves llengües.

Aquest article proposa accions concretes per ajudar a crear aquest món del futur. No cal conèixer una altra llengua. No cal unir-se a una campanya. Però sí cal que et mullis, que t'arrisquis, i que t'atreveixis a baixar la guàrdia i formar part de la minoria. Te la jugues?

SALUDA A LA GENT EN LA SEVA LLENGUA

Et creues amb algú que no té la mateixa llengua materna que tu? O potser no el mateix dialecte? Potser veus a aquesta persona a la feina, a l'escola, al parc o al supermercat. Es tracta d'utilitzar una salutació senzilla com "hola", preferiblement una que funcioni a qualsevol moment del dia.

Cada vegada que us creueu, saluda-la amb aquesta expressió. Fixa't si això afecta d'alguna manera la vostra relació. És possible que aviat descobreixis alguna cosa més sobre la llengua, la cultura i la comunitat local d'aquesta persona. Podries fer servir una aplicació mòbil per ajudar-te a memoritzar salutacions en diverses llengües.

Si ets professor, perquè no aprendre salutacions en les llengües dels teus alumnes? Podries practicar-les a classe, a les reunions escolars, o amb els pares.

Recorda que hi ha persones que no els agrada reconèixer en públic que parlen una llengua estrangera. Tingues cura de que ningú se senti incòmode, ni que t'estàs lluint a costa seva. Saludar a algú quan te'l creues al passadís és diferent de fer-ho aixecant la veu a una reunió pública. Tingues en compte també que, en algunes llengües, les salutacions varien segons l'estatus personal o la direcció si s'està en moviment. Pot ser que els parlants hagin d'adaptar la salutació lleugerament per trobar-ne una que puguis fer servir.

Hola!
Anyi paranga ra (Ma'di)
Te aso tokereka (Takū)
Ngudda kamak (Kunwinjku)
Abilaki (Eskayan) Dahooja (Carrier)
Wú cjêew (Shilluk)
Aŋpétu wašté (Lakota) Masikati (Shona)
Palya (Pitjantjatjara) Gude (Tok Pisin)

APRÈN A PRONUNCIAR ELS NOMS DE LA GENT

> "Jo... Ja... Joh-von. Ja-Va. Ai, a la m*rda, et direm Joe."
>
> –Alec Patric, *Black Rock White City*

Tens una amiga estrangera i sospites que no pronuncies bé el seu nom. Es molesta quan la gent pronuncia malament el seu nom? Li agradaria que aprenguessis a dir-lo bé? Prepara't per aquesta activitat descarregant-te una aplicació de gravació de veu.

1. Demana-li a la teva amiga si pots provar una cosa, comença a gravar i pronuncia el seu nom de la millor manera possible. (Això posa en evidència que estàs gravant, i t'ajudarà a adonar-te de les diferències de pronunciació després.)
2. Demana-li que et corregeixi. Apropa-li el telèfon i demana-li que el repeteixi una altra vegada a poc a poc.
3. Reprodueix la gravació per ensenyar-li el que has gravat. Intenta pronunciar bé el seu nom i torna a demanar-li que et corregeixi. Fixa't en com mou els llavis. Demana-li que et corregeixi en el futur si t'equivoques.
4. Més tard, escolta i practica en privat. L'objectiu és millorar, no la perfecció! Edita el nom de la gravació perquè la puguis trobar amb facilitat.

Quan tinguis una mica de pràctica, prova-ho amb amics o coneguts. Si ets professor, demana als infants o els seus pares que pronunciïn els seus noms mentre els graves. Podries fer-ne una activitat de classe. Podries demanar-li a algú amb un nom difícil de pronunciar que enumeri totes les maneres diferents que ha sentit dir el seu nom. Verbalitzar les versions incorrectes ajuda als infants a adonar-se de les diferències. Tingues en compte que algunes persones prefereixen que els anomenin amb una versió local del seu nom; el meu amic Ruprecht demana

als seus amics angloparlants que li diguin Rupert.

PARTICIPA A LES FESTES LOCALS D'UNA COMUNITAT

Moltes ciutats compten amb minories culturals que celebren festes anuals. Acostumen a ser esdeveniments d'una única cultura en una sola llengua, obertes al públic però on normalment només assisteixen membres d'aquella comunitat.

Ves-hi! Demana permís per observar-los. Hi haurà molta gastronomia i vestits tradicionals. L'exercici consisteix en aprendre algunes paraules. Troba algú que t'ensenyi una salutació i practica-la amb d'altres persones. Potser hi ha una paraula a un cartell i pots preguntar-ne el significat i com es pronuncia. Compra alguna cosa de menjar i demana com dir "M'agradaria...". Fes servir aquesta expressió a una altra parada, assenyalant alhora. Intenta imitar la manera culturalment apropiada d'assenyalar; podria ser utilitzant els ulls, el llavi inferior, o assentint amb el cap.

Potser et conviden a d'altres esdeveniments. A mesura que connectes amb membres de la comunitat, intenta rebre la seva generositat i acollida sense que es converteixi en una transacció. No caiguis en la temptació de resoldre problemes, intervenir, fer campanya o posar-hi diners. Evita convertir-te en el professor de la llengua majoritària local.

Centra el teu temps i esforços limitats en ser obert i cercar una connexió. Aprèn més noms i salutacions. Aprèn de memòria alguna cançó popular en la llengua. Investiga què pensa la gent del teu interès. Quina és la seva teoria de la llengua? Els preocupa mantenir activa la seva llengua? Recorda que simplement amb la teva presència i interès els estàs reconeixent de forma significativa. No cal fer gaire més.

TORNA A CONNECTAR AMB LA TEVA LLENGUA MATERNA

A la teva infantesa, parlàveu alguna altra llengua a casa? Potser tens familiars que encara la parlen, alguna tieta que visqui a prop, o algun cosí al país d'origen. Podries intentar tornar a utilitzar la teva llengua materna original?

Com seria parlar-la amb fluïdesa? Potser tens una altra personalitat en aquesta llengua. Podries establir un nou vincle amb la família més llunyana. Potser escoltes històries sobre com eren els teus pares quan eren joves.

Investiga si podries fer classes els vespres o caps de setmana. Busca algú amb qui practicar, encara que sigui en línia. Cerca oportunitats per escoltar la llengua. No importarà si al principi no ho entens tot. Submergeix-t'hi, és la millor manera d'aprendre!

CRIA INFANTS BILINGÜES

Els beneficis del bilingüisme són àmpliament coneguts. Molts infants bilingües desenvolupen una millor cognició social, comprenen de forma més profunda que els altres veuen el món des d'una perspectiva diferent. Com a adults bilingües, potser són millors ciutadans del món i tenen facilitats per aprendre una tercera o quarta llengua. En la vellesa, s'ha des-

Una dona escolta a una persona gran que comparteix històries en la seva llengua (Samarkand).

cobert que el seu declivi cognitiu és més lent i la demència apareix més tard. No es coneix cap inconvenient de la criança bilingüe.

Tot i això, els pares que s'han criat amb una altra llengua sovint pensen que a casa s'hauria de parlar la llengua dominant. Com l'aprendran sinó els seus fills? No obstant, els infants acaben parlant la llengua dominant amb fluïdesa sense importar la llengua parlada a casa. Les escoles estan començant a reconèixer això, i algunes tenen programes d'immersió lingüística. Tens algun d'aquests programes a la vora?

Imagina com seria per un infant mantenir un vincle estret amb els orígens de la seva família i alhora participar activament a la nostra societat i economia? Quines converses i relacions són més fàcils quan pares i fills parlen la mateixa llengua amb fluïdesa?

Si decideixes criar infants bilingües o multilingües, potser et trobes amb certa oposició. Resulta pràctic si teniu dues llengües que formen part de la vostra vida familiar i si els teus fills estan motivats per parlar-les. També és recomanable que escoltin la llengua per part d'altres persones i fonts com llibres, vídeos, cançons i l'Internet.

PARLA LA LLENGUA ORIGINAL DE LA TEVA TERRA

La llengua d'una regió és important, senzillament perquè hi era abans que tu. Fa segles que la terra que trepitges escolta aquesta llengua! Per aquest motiu, Sarah Palin va dir: "Si estàs a Estats Units, parla americà!".

Palin defenia l'anglès com si el seu futur estigués en perill. Val la pena intentar imaginar com seria si la teva llengua materna estigués en perill. Adoptarem el llenguatge de Palin pels nostres propis propòsits vitals. *El principi de Palin: Aprèn a parlar la llengua original del lloc on vius.*

Quina és la llengua original de la teva zona? Encara es parla? Pots aprendre algunes paraules? Pots apuntar-te a un curs?

L'objectiu no és parlar-la amb fluïdesa. Dominar una llengua és com dominar una professió, un esport, o un instrument. Implica vora les 10.000 hores! Tot i així, pots prendre-t'ho amb curiositat, aprendre paraules i expressions útils, descobrir el que signifiquen els topònims, i les seves històries.

Podries adoptar-la i utilitzar-la en determinades situacions a la feina o a l'escola, a cartells, visites, per anomenar els cursos, etc.

Recorda que els parlants de llengües minoritàries generalment no pensen en la seva llengua com un instrument que d'altres poden utilitzar pels seus propis objectius. La seva llengua forma part de la seva identitat més íntima. Hauràs de guanyar-te la seva confiança i demanar permís a les persones adequades.

JUGA A JOCS LINGÜÍSTICS

Hi ha molts jocs lingüístics per provar. Aquests són alguns dels jocs descrits amb detall a languageparty.org.

Foursquare Hello: És una versió d'un joc infantil on utilitzem salutacions en les llengües dels altres membres del grup mentre ens passem la pilota. Si t'equivoques tornes a la darrera posició. L'objectiu és convertir-se en el Rei Poliglot i superar a tots els teus adversaris.

Hip Hello: Aprenem les salutacions més de moda en argot, també amb expressions facials i gestos amb les mans... Sigues el més enrotllat en una altra llengua amb una única frase!

Garden of Words

Cacophony Line

Garden of Words: Poseu-vos per parelles. Els "escultors" pensen una paraula intraduïble en la seva llengua materna, com un concepte o una emoció, i l'expressen col·locant la seva parella, el "fang", en la posició adequada. Col·loquem les paraules en les llengües estrangeres als seus peus i ens passegem pel jardí d'escultures i n'endevinem els significats.

MorphoLogical: Un joc on apliquem algunes de les normes de formació de paraules més estranyes del món per inventar-ne de noves, i les fem servir a converses informals.

Cacophony Line: Quatre voluntaris es col·loquen davant, d'esquena al públic, i es giren de forma aleatòria per explicar-nos una història en la seva llengua, aturant-se quan es gira la següent persona. És un joc lingüístic molt ràpid i graciós.

440 – Four People Four Languages Zero Barriers: Quatre persones tenen una conversa improvisada en quatre llengües, i s'han de basar en les expressions facials i el llenguatge corporal dels companys per decidir com respondre.

MUNTA UNA FESTA LINGÜÍSTICA!

És un miracle que existeixin més de 4.500 llengües actives al món. Com pot ser possible això després de segles de colonialisme, nacionalisme, globalització i, el que és pitjor, les burles i el menyspreu dels parlants de les llengües dominants? La millor manera de donar-hi resposta és muntar una festa. Una festa lingüística!

És el moment de trobar-te amb els teus nous amics i celebrar la diversitat lingüística del món. Reuneix a la gent i viu les històries de la mateixa forma que s'han transmès de generació en generació: mitjançant la llengua oral.

El format és ben senzill: convida a la gent a compartir una història de 3 a 5 minuts en la seva llengua materna i llavors explicar-la en la llengua dominant. Anima'ls a compartir històries folklòriques en comptes d'històries traumàtiques. Et sorprendrà la facilitat que tenen els parlants de llengües minoritàries per explicar grans històries! També pots demanar cançons.

Cal que el grup vingui preparat. Com a amfitrió, cal animar a la gent a "escoltar per apreciar" més que "escoltar per comprendre". Es tracta de llengua en forma d'art, de música,

Four People Four Languages Zero Barriers

d'ànima parlada. Ningú entendrà tot el que es digui, però tothom pot escoltar com sona cada llengua, prestar atenció al seu ritme i la seva melodia, als gestos i les expressions facials, i endevinar de què tracta cada història.

Abans de cada història, demana als narradors que ensenyin una salutació. Practiqueu fins que tothom la digui bé. Llavors demana al narrador que comenci la història amb aquesta salutació. Pots trobar més informació sobre aquest format de narració a languageparty.org.

QUAN DEIXES LA TEVA ZONA DE CONFORT...

Quan provis aquestes activitats, et sentiràs vulnerable. T'estàs apropant a la gent de maneres que no s'esperen. Poden sospitar de les teves intencions. Poden notar la teva incomoditat, i que això també els faci sentir incòmodes. Recorda que estàs provant quelcom nou. És com aprendre a anar amb bicicleta, són habilitats que s'han de treballar. No et rendeixis el primer cop que caiguis!

Alhora, estàs anant contracorrent. Estàs intentant establir un vincle amb gent que potser s'ha sentit sempre rebutjada per la teva cultura. Potser no se senten agraïts en el moment que has decidit fixar-te en ells. Potser tenen un mal dia, o necessitaven una presentació culturalment apropiada.

Quan les coses no van rodades, recorda que el rebuig que sents és el mateix que sent qualsevol persona d'una minoria quan intenta formar part de la cultura dominant i aquesta l'exclou, l'ignora o se'n burla d'ella. Això ho fas voluntàriament i pots tornar a la teva zona de confort en qualsevol moment. Com seria no tenir escapatòria?

És complicat establir un vincle superant barreres arrelades i invisibles. Però es fa

més fàcil a mesura que et trobes amb persones amables, et guanyes la seva confiança, et sents a gust i deixes de patir per fer el ridícul. Recorda, estàs ajudant a crear noves maneres i nous espais perquè la gent senti acollida. I això té una recompensa especial. Hi ha una altra persona que sentirà que pertany al teu espai d'una altra manera. *Tu*.

ALTRES LECTURES

Austin, Peter K (2008). *One Thousand Languages: Living, Endangered, and Lost.* University of California Press.

Evans, Nicholas (2009). *Dying Words: Endangered Languages and What They Have to Tell Us.* Blackwell.

Grosjean, François (2009). What parents want to know about bilingualism. *The Bilingual Family Newsletter*, 26(4), 1-6. francoisgrosjean.ch/for_parents_en.html

Hinton, Leanne (2001). How to Keep Your Language Alive: A Commonsense Approach to One-On-One Language Learning. Heyday.

ALTRES RECURSOS

languageparty.org, untranslatable.org, wikitongues.org, elalliance.org, livinglanguages.org.au, ilinative.org, languageconservancy.org, ethnologue.com, psychologytoday.com/blog/life-bilingual, multilingualliving.com, bilingualism-matters.org

AGRAÏMENTS

Els agraeixo a Manuel Maqueda, Robyn Perry, Nadia Chaney i Michael Margolis per ajudar-me a donar forma a les idees presentades més amunt. Gràcies a Lauren Gawne, Antonella Sorace i Hakan Seyalioglu pels seus comentaris als esborranys inicials. Gràcies especialment als narradors de les nostres festes lingüístiques, pel vostre coratge al sortir a l'escenari i compartir una història en la llengua del vostre cor. Us acompanyarem sempre.

Aquest article va aparèixer originalment a Seyalioglu and Hymes (eds). Dialect, A Game About Language and How it Dies, Thorny Games, 2018.

Linguapax International
In official partnership with UNESCO (consultative status)

Carrer de Maria Aurèlia Capmany, 14-16
08001 Barcelona
Tel. 932 701 620
www.linguapax.org