On the Difficulties of Using NLP for Language Revitalization

Anonymous ACL submission

Abstract

This paper is dedicated to discussing the general difficulty of using emerging Natural Language Processing (NLP) technologies to the revitalization of languages. Previous literature had described the social causes of language shift; legal prohibitions, social and economic marginalization as well as a lack of inclusion in public life have been identified as the main factors for the non-viability of minority languages. As such, as innovative as they may be, these emerging tools are not enough to rescue languages and the core issues must be addressed if meaningful results are expected.

1 Introduction

Language extinction is a phenomenon that has been attested since the beginning of human history, but there is growing evidence that language extinction is happening at a never-before-seen rate: one language is estimated to go extinct every two weeks (Evans & Levinson 2009). Through analogies with conservation biology, it has been concluded that languages are even more endangered than wildlife (Sutherland 2003; Skutnabb-Kangas 2000: 83), with Skutnabb-Kangas stating that between 50 and 90% of the world's language could go extinct within the next century, whereas the corresponding figure for animal species is between 2 and 20%. Krauss (1992) had famously estimated that 90% of the world's languages would go extinct within the 21st century.

Given the urgency of the crisis, it is not surprising to see many looking to new technologies to revitalize and save many of the world's currently shifting languages.

However, together with the great excitement and "hype" that comes with these shiny new tools, it is important to temper our expectations and remember the first principles; in this paper, we will thus be reviewing the causes of language shift, as described in the literature, and what this means for the impact that NLP can have for the revitalization of minority languages.

2 Background: The Causes of Language Shift

Language shift is a social evolution in which the entire community progressively abandons their language over the course of the passing generations. Languages may be receding in some regions while thriving in others. The causes of LS must therefore be due to the social environment that the speakers find themselves.

Fundamentally, languages which provide greater social mobility and economic opportunities are chosen as a lingua franca, paving the way for language shift (Kandler & Unger 2010). The social environment thus determine which languages are more socially and economically viable.

Some have argued that power relations and structural forces are to blame for language shift and that members of minority ethnic groups are usually disadvantaged, socially, politically or economically, relative to the speakers of the dominant linguistic group. (Skutnabb-Kangas 2000: 29; Fishman 1991: 59). Inherently, Skutnabb-Kangas believes that language shift must be motivated by a combination of punitive measures to discourage the use of one language, and social and economic rewards for shifting to the dominant one. Fishman likewise (1991: 56) considered the prohibition of language use and advocacy as an "obvious" cause of language shift.

As such, an imbalance in the social power between two linguistic groups as being the root cause for shift. It is thus motivated by an attempt by speakers to escape linguistic injustice and

discrimination, which Skutnabb-Kangas refers to as linguicism, while Fishman called it a "cruel dilemma": to transmit the language, together with its social disadvantages, or stop its transmission to ensure a better future for the children (Fishman 1991 : 60).

Hale (1998) likewise considers linguistic justice to be essential for languages to be maintained. In doing so, he considered both elements of public policy and wider economic forces:

"*The condition which must prevail in order to halt language loss is a form of sociopolitical and economic justice in which this choice is not limited."*

The rewards for assimilating linguistically may include "books, radio licenses, food, clothes, additions to teacher salaries" (Skutnabb-Kangas 2000: 412). Punishments include corporal punishment for using the language, particularly in schools, or any other direct sanctions and prohibitions (Skutnabb-Kangas 2000: 347).

Formal education in particular, may be the most significant factor in language shift (Skutnabb-Kangas 2000: 29) and cross-cultural studies of loss among the Australian Aboriginals, the Sami and the Inuits, has led some to brand schooling as one of the *causes* of shift (Skutnabb-Kangas 2000: 97; Austin & Sallabank 2011: 6), because education is both a reward, as well as a setting where punishments for speaking the minority language may be regularly handed out.

Thus, socially dominant languages provide distinct advantages that minority ones do not; this typically comes in the way of widespread use in institutions and administration (Austin & Sallabank 2011: 96), which can quickly become "weapons of assimilation" if certain languages are excluded (Houston 2003). This is in fact so significant that language vitality scales often use the degree of institutional support as an indicator of language endangerment, such as in the case of the EGIDS (Lewis & Simons 2010) or by UNESCO (Austin & Sallabank 2011: 38).

The role of institutions is thus considered crucial in language survival, as in the words of Fishman (Fishman et al. 2013):

"*No speech community can maintain two languages on a stable basis (past three generations) if they are both used in the same social functions and, therefore, stable societal bilingualism (diglossia) depends on institutionally protected functional sociolinguistic compartmentalization, so no ethnocultural collectivity can maintain two cultures on a stable basis past three generations if they are both implemented in the same social functions (family, friendship, work, education, religion, etc), and therefore, stable societal multiculturism (di-ethnia) depends on institutionally protected ethnocultural compartmentalization.*"

2.1 Economic Dimension of Language Shift

The dominance of one linguistic group over the other does not only entail institutions and public policy. More generally, speakers of the receding language have to come to rely upon the dominant group; this often means that their livelihood is no longer entirely in their power and that they depend on the dominant group economically (Fishman 1991: 60).

As a result, language shift occurs when there is unfair competition between the languages on the marketplace (Austin & Sallabank 2011: 405). For this to be possible, the traditional method of livelihood must first be rendered obsolete, either through physical disruptions, e.g. deforestation, confiscation of material resources, or through more subtle psychological means, e.g. by enticing speakers with greater employment opportunities outside of the linguistic community (Austin & Sallabank 2011 : 405), in a form of cultural propaganda, which may be intentional, or a side-effect of a large difference in economic power and development between the two groups.

Capitalist market forces also inherently seek the highest possible profit for the lowest possible cost; this tends to result in centralization and cultural homogenization (Skutnabb-Kangas 2000: 656), by making products and services available only in (economically meaningful) national lingua francas. For example, video game translations are typically only available in Spanish, but rarely, if ever, provided in non-state language such as Galician or Basque (Fernández-Costales 2017).

3 Technical Difficulties: Data-Poor, Cash-Poor Minority Languages

Machine translation is powered by machine learning, which relies on algorithms and previous training to predict an "outcome" based on a given input (Yang 2019: 161). In the case of machine translation, the input would be a sample of text or speech in a given language, and the corresponding prediction would be the translation in a

corresponding language. In order to train the algorithm, large amounts of data are required, to allow it to go through many attempts and compare its own prediction with the actual translation, before attempting to improve it (Yang 2019: 161).

However, due to their much lesser institutional support, minority languages tend to be a lot more data-poor than widely spoken socially dominant languages, a challenge oft-mentioned in articles on the matter (Arkhangelskiy & Medvedva 2016; Ambati & Carbonell 2009). Almost every aspect of minority languages is comparatively under-utilized; they are typically excluded from certain spheres of use, such as governance and education, thus decreasing the funds available for them, as well as decreasing their overall usage – this results in less data being generated in them.

Their lesser use also means that some of them even lack a standard writing system, or they may lack one altogether[1]. This might result in minority language speakers resorting to the dominant language for writing, in both formal and informal contexts, further limiting the possibility of storing linguistic data in its written form.

In media, minority languages also tend to be under-represented, both due to the economic efficiency of targeting a wide consumer base through lingua francas – which are often the dominant languages replacing them – as well as the lack of funding that comes with exclusion and institutional marginalization.

In fact, corporations such as Google have already cited all of these issues explicitly as a justification for the lack of support for many Native American languages (Hilleary 2021).

In addition, creating NLP technology is a highly expensive endeavor; salaries within the industry are high; in the US, a machine learning engineer position yields an average of US$113,000 per year (Payscale 2021). On the topic of financing, when it comes to language revitalization, Fishman had already suggested that the minority language speakers should fun the initiative themselves, at least in the initial stages, noting that "It may seem unfair that the poor should have to tax themselves for their own betterment, but that is the way of the world and if Xmen do not labor on behalf of Xish before the world as a whole is changed, no one will do it (or pay someone else to do it)" (Fishman 1991 : 98).

Thus, the issue of funding is one that must also be addressed; either the private sector or government agencies will thus have to take on the costs associated with development. In many ways, this is already apparent, for example, a quick glance at the languages available on Google Translate shows three types of languages:

National, dominant languages of independent states, e.g. English, German, Chinese

Regional and minority languages, with a national or autonomous regional government sympathetic to revitalization or preservation of the language, e.g. Scottish Gaelic or Basque

Non-state languages with a large number of speakers, e.g. Hausa, Igbo

This would suggest a pattern where private companies invest in NLP technology if it allows them access to a large, otherwise untapped market. When that is not the case, non-profit organizations, e.g. governments, must step in to provide funding. Scottish Gaelic for example, a language of a mere 50,000 speakers (National Records of Scotland 2011), who are all fully bilingual in English, was added to Google Translate in 2016. As this is unlikely to bring additional income to the company, it is likely that this was due to partial funding and support from the Scottish government. It was in fact reported in the news that the Scottish government had "backed" the plan to develop Gaelic support, although it is not entirely clear to what extent actual funding was involved, but it seems that the "tax-payer funded Gaelic Board" could have been involved (Pauling 2015).

4 Outlook: Can NLP Revitalize Languages?

As previously discussed, languages do not shift due to a lack of opportunities to learn them, rather they shift due to a lack of opportunities to benefit from their use. Many of the currently shifting and endangered languages in fact have a wealth of resources to learn them. Breton for example, has had dictionaries since the XVth century, with the release of the Catholicon (Trepos 1964) and grammars have likewise been available since the XVIIth (Hewitt), but language shift from Breton to French is a long-standing phenomenon, with the Breton domain shrinking gradually over time in favour of French (Even 1987 : 157). Irish is also

[1] Although many languages may simply lack a writing system for cultural reasons.

universally taught in schools within the republic of Ireland but only about 2% make it their daily language outside the education (Petit 2016).

It is thus a lack of economic opportunities in which the minority language is viable, let alone useful, that is often the cause of language shift in the first place. One could therefore wonder whether NLP acts to replace those few employment opportunities available for speakers of minority languages, by for example, offering automatic translation, when this could be a source of employment.

Taking the television industry as an analogy, Ó Ceallaigh, in his thesis about the economic crisis and its impact on the state of the Irish language (2020: 123), noted that "that those involved in the technical aspects of TG4's productions simply do not know Irish" and that "it is for this reason that Irish is used only about 50% of the time on the set of one of the station's flagship shows". As university courses of highly technical fields, such as computer science, are a lot more likely to be available in dominant languages such as English, it is likely that the same applies in the IT industry, even when developing tools for minority languages.

5 Conclusion

Given that language shift is a social issue, rather than a technological limitation, it likely that NLP technologies' contribution to revitalization can only be modest if used on their own. Prohibitions, discrimination, marginalization and economic reliance on national languages must also be solved in order to revitalize shifting languages.

In addition, given that certain tools such as automatic translation see limited use even for more widely spoken languages, it is difficult to imagine how much practical use such technology could have for minority languages.

In fact, Fishman had even noted (1991: 67; 107) that developing minority language media such as radio and television, is often not worth the amount of resources that it required, but could nevertheless increase the prestige and boost the speakers' self-image. It is likely that the same applies to NLP for minority languages, which may not represent the most efficient way of revitalizing endangered languages, assuming that funds are limited, as they very often tend to be. Perhaps dedicating more resources to making public services, such as education, available in minority languages could be a more worthwhile endeavor. Reserving some employment opportunities for speakers of minority languages, or requiring employees to undergo language training could also make the minority language more viable, especially given that language is a tool of social communication.

Nevertheless, just as was the case for minority language media, the development of minority language NLP would provide speakers of minority languages some employment opportunities where their mother tongue is an asset and can be used professionally, rather than exclusively informally – as is usually the case in diglossic societies. In addition, as part of larger, more ambitious preservation efforts, developing NLP tools for minority languages could be a way of normalizing, and ultimately, modernizing languages all too often left behind, providing a boost to the self-esteem of marginalized individuals, who may start to feel like the inclusion of their language normalizes their existence, in a world where their very presence may be treated as an oddity. In a more practical sense, it is also important to allow minority languages to be usable directly in the digital world, without having to constantly rely on national languages. Perhaps the development of NLP for minority languages is also worth pursuing, if only for the sake of the benefits that they could one day bring to society, even if they are not immediately obvious in their current sake.

References

Ambati, Vamshi, & Jaime G. Carbonell. 2009. *Proactive learning for building machine translation systems for minority languages*. Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing (pp. 58-61).

Arkhangelskiy, Timofey, & Maria Medvedeva. 2016. *Developing Morphologically Annotated Corpora for Minority Languages of Russia*. CLiF (pp. 1-6).

Austin, Peter K. & Sallabank, Julia (Eds.). 2011. *The Cambridge handbook of endangered languages*. Cambridge University Press.

Coldewey, Devin. 2019. *Microsoft adds Māori to translator as New Zealand pushes to revitalize the language.* (Available online at, https://techcrunch.com/2019/11/22/microsoft-adds-maori-to-translator-as-new-zealand-pushes-to-revitalize-the-language/?guccounter=1, Accessed 14 October 2021.)

CRYSTAL, David. 2000. *Language death*. Ernst Klett Sprachen.

EVEN, Arzel. 1987. *Istor ar yezhoù keltiek: Ar yezhoù indezeuropek, ar c'helt-edaleg, an hengeltieg, ar galianeg, ar predeneg, ar brezhoneg, ar c'herneveg*. Hor Yezh.

Fernández-Costales, Alberto. 2017. *On the sociolinguistics of video games localisation: Localising games into minority languages in Spain*. The Journal of Internationalization and Localization 4.2, 120-140.

Fishman, Joshua A. 1991. *Reversing language shift: Theoretical and empirical foundations of assistance to threatened languages* (Vol. 76). Multilingual matters.

Fishman, Joshua A., Michael H. Gertner, Esther G. Lowy, & William G. Milán. (2013). *The rise and fall of the ethnic revival*. De Gruyter Mouton.

Hale, Ken. 1998. *On endangered languages and the importance of linguistic diversity*. Endangered languages: Language loss and community response, ed. by Lenore Grenoble and Lindsay Whaley, p. 192-216.

HEWITT, Steve. *Background information on Breton*. (Available online at: https://d1wqtxts1xzle7.cloudfront.net/30212312/Hewitt_-_Background_information_on_Breton.pdf?1353316789=&response-content-disposition=inline%3B+filename%3DBackground_information_on_Breton.pdf&Expires=1636406381&Signature=eLAmTGwfAvSX3Jh3VfICE~9x6p8VbLFVsjH~NJPav-ON3SqXPg2t8zU7X3gmRVdqP6VrCPM~OjktICjP32vASMePovXfQVAddnLpBhjSI57WJhDD2a8LpBuksJ7kdgcUVWcez8ELJ5MO3xt4Si7bmm1UIY2N5q~iUEJuTFzqcOvUbRrTUowb3YgurQua2TQ3b-uxRiAWWovm5ssSSmIKaMqyNO5rTp9a3bI97n9n9Wf4H20T~yMcdTiavRkhLzZQxKY82td2Qd1tyO~ZyGpjyT6bLSZuTG81CgRWi9A1BNbNx8vg~iTklsob~m5Ia9lU2zouTzFgoTinDioYEds4Bg__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA, Accessed 27 October 2021)

Hilleary, Cecily. 2021. *Google explains why App Can't translate most Native American languages.* (Available online at: https://www.voanews.com/a/usa_google-explains-why-app-cant-translate-most-native-american-languages/6204275.html, Accessed 4 October 2021.)

Houston, Robert A. 2003. *'Lesser-used' languages in historic Europe: models of change from the 16th to the 19th centuries*. European Review, 11(3), 299-324.

Kandler, Anne, Roman Unger, and James Steele. 2010. *Language shift, bilingualism and the future of Britain's Celtic languages*. Philosophical Transactions of the Royal Society B: Biological Sciences 365.1559 (2010): 3855-3864.

Krauss, Michael. 1992. *The world's languages in crisis*. Language (Baltimore), *68*(1), 4-10. Evans and Levinson

Lewis, M. Paul & Simons, Gary F. (2010). Assessing endangerment: expanding Fishman's GIDS. (Available online at: https://www.lingv.ro/RRL%202%202010%20art01Lewis.pdf, Accessed 4 October 2021.)

National Records of Scotland. 2011. *Scotland's Census 2011.* (Available online at, https://www.nrscotland.gov.uk/files/statistics/annual-review-2013/html/rgar-2013-scotlands-census-2011.html, Accessed 27 October 2021.)

Ó Ceallaigh, Ben. 2020. *Neoliberalism and language shift: the Great Recession and the sociolinguistic vitality of Ireland's Gaeltacht, 2008-18*.

Pauling, Tim. 2015. *Support for putting Gaelic on Google translation service.* (Available online at: https://www.pressandjournal.co.uk/fp/politics/scottish-politics/457655/support-for-putting-gaelic-on-google-translation-service/, Accessed 5 October 2021.)

PayScale. 2021. *Average Machine Learning Engineer Salary.* (Available online at, https://www.payscale.com/research/US/Job=Machine_Learning_Engineer/Salary, Accessed 26 October 2021.)

PETIT, Kevin. *Successful Learners of Irish as an L2: Motivation, Identity and Linguistic*. Studia Celtica Posnaniensia, 2016, vol. 1, no 1, p. 39-56.

Skutnabb-Kangas, Tove. 2000. *Linguistic genocide in education--or worldwide diversity and human rights?*. Routledge.

Sutherland, William J. 2003. *Parallel extinction risk and global distribution of languages and species*. Nature, 423(6937), 276-279.

TRÉPOS, Pierre. 1964. *Le Catholicon de Jehan Lagadeuc. Pour son cinquième centenaire*. Annales de Bretagne et des pays de l'Ouest. Persée-Portail des revues scientifiques en SHS, p. 501-552.

YANG, Xin-She. 2019. *Introduction to algorithms for data mining and machine learning*. Academic press.