IRIE
International Review of Information Ethics Vol. 32 (11/2022)

Author: Richard Harvey

Multilingualism in cyberspace: a practical reality?

Abstract:

The loss of a language is often phrased in catastrophic terms despite the evolution of languages being a quite natural process. Here we discuss language loss and language use in the digital domain. We note that while English still dominates cyberspace, other languages are growing rapidly, so it seems likely that the future will not be monolingual but multilingual. We show that efforts to mandate the use of minority languages are unlikely to be successful and can backfire. The question therefore arises of how best to handle minority languages in the digital domain. This article argues that the best, and maybe the only, solution is high-quality machine translation.

Keywords: Communication, Cyberspace, Languages, Linguicide, Multilingualism, Translation

Agenda:

Authors:

Richard Harvey:

- School of Computing Sciences, University of East Anglia, Norwich, NR4 7TJ, UK, Email: r.w.harvey@uea.ac.uk, Website: www.uea.ac.uk

ISSN 1614-1687

Introduction

It is becoming more commonplace to frame linguistic loss as equivalent to the death of biological organisms. Thus, language death can be called *linguicide.* The term is not sufficiently well described to have yet reached the Oxford or Cambridge English Dictionaries but, in many writings, linguicide is used as an equivalent to genocide. There is also *linguicism* which is the analog to racism. Since genocide and racism are bad, the linguistic metaphoric alternatives are also bad. Not only is this stretching metaphor too far (Frank 2008), it is dangerous as it can lead to a form of "culture wars," the "language wars" perhaps, in which one side attributes all linguistic losses to malicious actions, and the other side positions itself as a defender against "woke" show-boating. Such debates can be futile, since language loss is an inevitable part of linguistic development and there is a tendency to ignore the critical issue which is how people or entities who speak different languages might best communicate with each other.

In this paper, we briefly cover some of the misconceptions which appear to have arisen from the biological metaphor. I will argue that language adaptation is completely natural, and efforts to preserve languages beyond their natural life can be as damaging as some efforts to kill certain languages. I then examine the issue of access to digital services and show that efforts in Machine Translation (MT) allow us a possible peaceful resolution between each of the "language wars".

As a brief aside I would ask the reader to note that I often use the term "speech" to mean the use of language expression (as in phrase "free speech" which is not just restricted to talking). This paper is not about the difference between spoken, written and signed languages, it is about whether people can be free to communicate in their chosen form or whether in the interests of practicality these freedoms must be curtailed.

Language myths and realities

Examples of the catastrophisation of language loss are now quite widespread. For example, a note by the Executive Secretary submitted to one of the groups of United Nationals Environment Programme (UNEP) contains the startling statement that "If during the next century we lose more than half of our languages, we also seriously undermine our chances for life on Earth." (UNEP 2004) This statement turns out to be a direct quote from a report produced under the imprimatur of UNESCO and the WWF (Skutnabb-Kangas, Maffi and Harmon 2003). The argument, in a nutshell, in both (UNEP 2004) and (Skutnabb-Kangas, Maffi and Harmon 2003) is that languages are storehouses of historically developed knowledge and when a language is lost, that knowledge is lost. At first sight, this might seem a reasonable assertion. There are, for example, numerous stories of scientific "discoveries" which have turned-out to be rediscoveries of indigenous knowledge – in (Skutnabb-Kangas, Maffi and Harmon 2003) there is an account of how Nordic scientists "discovered" that salmon can spawn in small rivulets leading to the river Teno. Yet it turned out to be very ancient knowledge to some Sámi speakers. Had the researchers spoken Sámi, they would have known that some of these rivulets had names that included the Sámi word for "salmon spawning-bed."

While this story is entertaining, it raises two questions: firstly, had the researchers been more fluent in Sámi, would their research have differed? And secondly, if those rivers had been renamed in, say Finnish or Norwegian (the river Teno flows between Finland and Norway) would the knowledge have been lost? It is difficult to give a definitive answer to the first question since we are not fisheries researchers but, I suspect the answer is "No" -the researchers would have continued since there is widespread scientific mistrust of folklore. Readers may feel strongly that scientists should give more credence to local knowledge or even folklore, and they may be right, but debating whether scientists are arrogant or not is not pertinent since the second question is easy to answer – had the rivers been renamed with appropriate translations then clearly the knowledge would have been retained. Therefore, as far as this well-known example goes, knowledge is not bound to language. That said, there is an elision in the previous argument – what constitutes an "appropriate translation." Even at this stage in the argument, we might suspect that not all translations have had the purest

of motives, nor have they given equal credence to the views of minority speakers and those of the dominant language. We will discuss that later.

The second theme that emerges around language loss is how the dominance of one language, writers usually choose English, is harming minority languages. A particular theme is how digital language death is even more pronounced than the physical. Kornai in 2013 estimated that less than 5% of all languages can still ascend to the digital realm (Kornai 2013). It is very computationally challenging to make such estimates and we ought to note that Kornai uses the computational linguists favourite digital proxy which is Wikipedia. Nevertheless, there is certainly evidence that minority languages are not being used in the digital domain. However, English no longer dominates, and Wikipedia's statistics on page edits show English now represents only 40% of edits. An analysis of users implies that English speaking users might be down to as low as 26% with Chinese, Spanish and Arabic now all having greater than 5% share (Wikipedia: Languages used on the internet 2021). Thus authors argue that "the dominance of English is asserted and maintained by the establishment" (Phillipson 1992) or that "widespread use (of English) threatens other languages" (Pennycock 2017) or that "English hegemony threatens other languages and discriminates against non-English-speaking people" (Tsuda 2008) may soon need to be re-issuing their works replacing English with Chinese, Spanish, Arabic or Russian depending on how the trends develop.

In short, arguments that equate loss of linguistic diversity to loss of biological diversity are tenuous and not evidenced. Furthermore, arguments that state that English is crowding-out other languages are not fully appraised of the current trends in digital language usage. The internet is becoming more, not less, multilingual. The question, therefore, is whether multilinguality in cyberspace can become a practical reality?

Approaches to communication across languages

Tsuda (Tsuda 2008), when discussing what he perceives to be a rise in the use of English, characterises the responses into one of three categories which he calls:

1. The monolingual approach;
2. The global scheme approach;
3. Multilingualism.

The monolingual approach argues that for successful international communication then we need a language that we can all use. For the time being it is English, but English has arisen without much discussion of what the most appropriate language should be. Certainly, there are a number of reasons why English is a less than practical choice. Maybe we should heed the frequent pleas of the French government and return French to its preeminent position as the former language of diplomacy? Or maybe we should return to New Latin as the lingua franca of academic writing? Of course, it is all very well for native English speakers to be enthusiastic about English as an international language, although we should note that there are now considerable dialectical variations among World Englishes (Schneider 2007), since their fluency puts them in a position of some comfort. However, it also puts them at an unfair advantage in accessing information, negotiating and general business dealings.

One well-known alternative is to devise a universal language spoken natively by no-one: an international auxiliary language or IAL. Esperanto is the most widely spoken example with a very regular grammar which its proponents claim, make it easy to learn. It is a fundamentally euro-centric language with most of the vocabulary coming from Romance and Germanic languages so the arguments about unfair advantage apply also to Esperanto to some degree. That said, it is difficult to be optimistic about Esperanto or any other IAL. Like software adoption, there are very considerable hurdles to switching language, not least of which is the huge cognitive effort associated with language learning, so there needs to be very great incentives for learning a language. Hence, English has risen because it is associated with the dominant economies – command of

English is perceived as given access to substantial economic advantages. Even so, despite these advantages, it is often the case that parents will often defer learning English until after their children have achieved some fluency – children can acquire languages with far less cognitive effort than parents.

The global scheme approach is legislation: let us force, or strongly encourage people, by which it is usually meant governments, to recognise linguistic diversity. The most notable example is the Universal Declaration on Human Rights (UHDR), which, in Article 2 states that everyone is entitled to their rights and freedoms (which are listed in the document) irrespective of a number of characteristics which include race, colour, sex, religion, political opinion, origin, birth, status or language. It does not take much reading to realise that the language characteristic is either being ignored or, alternatively, is hugely impractical. For example, Article 10 of the UHDR, relates to being given a fair and independent legal hearing in relation to any criminal charge. If we were to pick one of the more liberal and large US states, the state of California, then there is a publicly accessible webpage[1] that lists the sixteen languages for which certified translators are available. It also provides a longer list where translators may be available. Unsurprisingly, a Sámi speaker, to use our earlier example, would not find a translator. Nor is a Finnish translator available. On the face of it, this is a clear breach of the Universal Declaration of Human Rights but presumably, the citizens of California are satisfied that it is an acceptable compromise. The practical point is that, although the State of California is enormously well resourced and politically minded adhering to the UDHR, it cannot possibly cover all the possible languages. Unlike sex or religion, the language characteristic presents a combinatorial explosion in a person's intersectionality which means that compromises have to be made.

An attempt to tighten-up some of the definitions around language and people was provided by PEN with their International Universal Declaration on Linguistic Rights (PEN International 1998). The document required considerable debate and redrafting due to difficulties of defining terms such as "language community" which, incidentally, is defined as existing in a "territorial space" thus eliminating virtual communities of, say, Esperanto speakers, in the first line of the declaration. The declaration tries to steer a difficult line recommending legal protections, which might be very expensive for the relevant taxpayer, and avoiding situations in which over-bearing governments force their citizens to speak a language. The Universal Declaration on Linguistic Rights thus illuminates an important principle to which we will return: all compelled speech is wrong. I will take this as a basic principle from now on. Said quickly, it is easy to agree that compelled speech is wrong, but it has some important and uncomfortable ramifications.

To see these, consider the situation of the Welsh language in the United Kingdom. In the early part of the 20th century, Welsh looked as though it would die out. However, around 15% of the population of Wales are said to speak Welsh daily and around 22% of the population have the ability to speak Welsh. There is widespread agreement that a critical factor in the rejuvenation of the Welsh language were a series of acts of parliament which, firstly, gave Welsh "equal validity" and secondly forced all organisations in the public sector, and some in the private sector, to speak Welsh. The punishments for not complying are a little arcane, but there is certainly some investigation and public shaming for transgressors. Needless to say, the act imposes considerable costs on society, most of which are borne by non-Welsh taxpayers who relatively frequently enquire, via Freedom of Information Act requests, how much the Welsh language policy costs. From these we know that the Welsh TV stations, S4C spends around £150M, per annum, the Welsh language Service costs around £40M and it seems Welsh Councils speed between £100k and £500k per annum. A further feature of the policy is that Welsh schools must teach Welsh which appears to be unpopular (a poll for ITC news reported that the most popular option was for the Welsh language to be optional rather than mandatory in schools) (Sheldrick 2015). In short, although the Welsh language policy has grown the use of Welsh it has increased cost and has produced some anti-Welsh language sentiment in the process. The nub of the issue seems to be that without compulsion few people would use or learn Welsh; but the compulsion of speech is not popular.

The final option is multilingualism. In this approach, there is no dominant or preferred language. Large bodies such as governments and corporations are expected to transact in multiple languages, and it is often the case

[1] https://www.courts.ca.gov/35273.htm

that citizens will also be multilingual. There are many examples of countries that have two official languages but examples of more than two are considerably rarer due to reasons of cost and practicality. Often governments will adopt a hybrid approach of saying there are "recognised" languages in addition to, or in place of, "official" ones. Slovakia, for example, recognises twelve languages and guarantees their use in municipalities where there is evidence they are needed.

On a transnational level, the two organisations which are notably very committed to multilingualism are the United Nations and the European Union. In the case of the UN, there are six official languages, Arabic, Chinese, English, French, Russian and Spanish and there are frequent General Assembly Resolutions, for example, Resolution A/RES/73/356, which mandate multilingualism and demand that the UN measure it's compliance. The most recent compliance document (United Nations 2021) reports that around 57% of the entities of the secretariat are compliant with the policy. However, the report also measures the percentage of external UN website content that is available in the official languages. There are some very noticeable differences: 99.3% of website content is available in English but only 26% in Chinese and 27% in Mandarin. A principal reason is stated as the lack of timely, affordable high-quality translation. However, it is noticeable that even more skewed are the language requirements by job openings at the UN: 98.7% mention English but Arabic, Chinese, Russian and Spanish are required by fewer than 5% (United Nations 2021). It would seem that the UN is committed to multilingualism up to a point. But when it comes to hiring people who might profess fluency in any other language than English their commitment dies.

The EU supports written multilingualism via the Translation Service (DG Translation or DGT) which handles 24 official languages plus a few others when necessary (European Commission 2021). The activity is impressive, as roughly half of DGT's output concerns law-making which is a particularly technical and demanding domain with short deadlines. That said, the total cost is very significant at around €0.5Bn per annum and, like all translation services, there are persistent pressure points around languages with small numbers of speakers such as Maltese and Irish.

In conclusion, of the three responses outlined in (Tsuda 2008), the monolingual approach may well be the *de facto* one with English dominant but it comes with some disadvantages of which unfairness is the most palpable. The global scheme approach is beset with difficulties and impracticalities. Furthermore, language regulation is a step towards compelled speech which is very unpopular and itself a violation of the Universal Declaration of Human Rights. It is multilingualism that presents the most opportunities but it creates the knotty problem of translation.

Translation

There are a great number of countries where even the most polyglot of people could not possibly speak all the languages in use in that country. One approach, is the one we have discussed with regard to Welsh: compelled speech. We might argue that, while compelled speech is undesirable, it is a rule worth breaking if it is essentially imposing a small burden on the powerful, to allow minority language speakers access to their democratically accountable institutions. But this is a deceptive argument – do speakers of minority languages not also have rights to access the full economic and social benefits of their majority-language-speaking fellow citizens? Does that imply that all organisations should be forced to speak the language of the minority? Or indeed, the majority be forced to speak the language of a minority? That sounds very dangerous indeed. More tellingly, but less obviously, this strategy actually reinforces unequal power relationships.

As we have seen with the EU's DGT, high-quality accurate translation is too expensive to be afforded by most private citizens. Thus, it is inevitable that it will be the powerful who pay, and hence control the translation. This leads to several problems. The first is that without any native speakers in the organisation commissioning the translation, there is a danger of unchecked work. Sometimes this can have comic results. The BBC reported in 2006 that Vale of Glamorgan Council installed temporary road signs instructing "Cyclists dismount." The translated Welsh instruction was "Llid Y Bldren Dymchwelyd" which is gibberish Welsh and could be translated

to mean "bladder inflammation upset" (BBC News 2006). A simple error but one that arose because Glamorgan Council either did not care enough about Welsh to check its signs, or were not competent to do so (Wikipedia asserts that 84% of Glamorgan's population have no knowledge of Welsh). What if a government or region has a noticeable antipathy to a language? Is it really believable, for example, that the Franco government of Spain would be prepared to honestly translate it's decrees and judgements into Basque? There are numerous historical precedents of antipathetic and irrational hatreds of certain languages by the powerful. To insist on translation that can only be paid for by the powerful is a licence for discrimination.

What is needed is a method of communication that allows both parties to speak the language they choose without the possibility of malicious or biased translation. Given that this has to function for all citizens it also has to be provided at minimal cost. The answer is machine translation.

Machine translation (MT) has a long history in computer science but until recently it has produced poor quality results. There have been several recent innovations which collectively are known as deep learning or deep neural networks. Deep because the neural networks have very many layers. Although such systems have been postulated since the 1960s the breakthrough was an algorithm to train them, and a surfeit of supercomputers with which to do the training. An example of a recent state-of-the-art system (Popel, et al. 2020) significantly outperformed professional agency English -to-Czech translation. Furthermore, most users of the system were unable to distinguish the machine from human translation – it passed the Turing test. Although there are plenty of caveats around such systems, (Popel, et al. 2020) is not an isolated result and a summary of a recent "bake-off" among systems concludes that "MT systems seem to reach the quality of human translation in the news domain for some language pairs" (Barrault, et al. 2019). In the interests of fair reporting, we should also note that we do not really yet understand why some language pairs are easier for some machines than others, nor is there a universal solution to translation across multiple domains. However, even though such systems are quite experimental, the technology cycle is now very short and commercial systems such as Google Translate now incorporate new developments very rapidly. Moreover, the modern web-user has numerous choices of systems, so it is possible to run each system and select from among the alternatives. Most importantly from our perspective, the business of translation is now out of the hands of state actors – there are multiple commercial vendors, so the likelihood of systematic bias or malice is reduced.

Even if one is sceptical about machine translation, the activities involved in building accurate machine translation systems which involve obtaining high-quality sources (either audio, sign or text) are incredibly helpful if one wishes to record and preserve a language. Further activity, which might include construction of the standard instruments of computational linguistics such as corpora, grammatical rules, lexicons, dictionaries, pronunciation dictionaries and so on; form the backbone of a systematic and detailed study of a language. Thus, investment in machine translation for minority languages is an enabler for the serious scholastic study of a language as well as providing a vital link between the minority and majority languages.

Conclusions

To reach the conclusions in this paper I have chosen to designate some principles as inviolable. They are:

1. All compelled speech is wrong.
2. Both parties in a conversation should have access to high-quality translation.

Principle 1 is usually uncontroversial when it applies to individual people, but the principle should also apply to legal persons such as governments, companies, and other corporate bodies. This extension may well give rise to some queasiness in those who seek social justice. They may well point out that is unfair to treat the impoverished and multi-billion companies to the same standards. But, as I attempt to show earlier, forcing a very powerful party to communicate in a language that they do not wish to, has many negative consequences, none of which are to the benefit of the language. Furthermore, it is a basic principle of modern legal systems that what is sauce for the goose is sauce for the gander. Principle 2 has turned out to be a practical stumbling point – who can afford high-quality translations? Only large entities – hence it has been argued that we should

compromise Principle 1 and force those entities to speak a language that is not native to them. In the case of a government, it may be perfectly reasonable for them to be required to communicate with its citizens in their native language but, even then, if the government has impure motives, will the translation be high quality and unbiassed? To assume it will be, seems pollyannaish. Machine translation changes the balance of power. Now we all have access to multiple machine translation systems. In narrow domains, MT is better than humans and it is improving all the time, so there is every reason to be optimistic. However, MT is usually improved by the systematic collection of linguistic data such as corpora, language pairs, grammatical data and so on. And, if we are to preserve languages from dying without any accurate record of their construction, then we need large scale datasets which can also be used to build translation systems that allow that language to flourish without the usual economic and social costs associated with speaking only a minority language.

References

Barrault, Loïc, Ondřej Bojar, Marta R Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, et al. 2019. "Findings of the 2019 Conference on Machine Translation (WMT19)." Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1). Florence, Italy: Association for Computational Linguistics. 1--61.

BBC News. 2006. "Bladder alert lost in translation." August 15. Accessed October 30, 2021. http://news.bbc.co.uk/1/hi/wales/south_east/4794753.stm.

European Commission. 2021. Annual Actvitiy Report 2020 DG Translation. Brussells: European Commission.

Frank, Roslyn M. 2008. The language-orgnaism-species analogy: A complex adaptice systems approach to shifting perspectives on "language". Vol. 2, in Body Language and Mind, by Roslyn M Frank, René Dirven, Tom Ziemke and Enrique Bernaárdez. Berlin: Mouton de Gruyter.

Kornai, András. 2013. "Digital Language Death." PLOS One 8 (10): 1--11.

PEN International. 1998. Universal Declaration of Linguistic Rights. Barcelona: PEN International.

Pennycock, Alastair. 2017. The Cultural Poltics of English as International Language. Routledge.

Phillipson, Robert. 1992. Linguistic Imperialism. Oxford: Oxford University Press.

Popel, Martin, Marketa Tomkova, Jakub Tomek, Ondřej Bojar, Jakob Uszkoreit, Zdeněk Žabokrtský, and Łukasz Kaiser. 2020. "Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals." Nature Communications.

Schneider, Edgar W. 2007. Postcolonial English. Cambridge: Cambridge University Press.

Sheldrick, Tom. 2015. "Exclusive poll: 64% oppose compulsory Welsh to age 16." July 10. Accessed October 30, 2021. https://www.itv.com/news/wales/2015-07-10/exclusive-poll-64-oppose-compulsory-welsh-to-age-16.

Skutnabb-Kangas, Tove, Luisa Maffi, and Dave Harmon. 2003. Sharing a world of difference: the earth's linguistic, cultural and biological diversity. Salt Spring Island, British Columbia, Canada: UNESCO, WWF and Terralingua.

Tsuda, Yukio. 2008. "English Hegemony and English Divide." China Media Research 4 (1): 47--55.

UNEP. 2004. INDICATORS FOR ASSESSING PROGRESS TOWARDS THE 2010 TARGET: STATUS AND TRENDS OF LINGUISTIC DIVERSITY AND NUMBERS OF SPEAKERS OF INDIGENOUS LANGUAGES. United Nations Environment Programme, UN.

United Nations. 2021. UN General Assembly Seventy Fifth Session Agenda Item 129: Multilingualism Report of the Secretary-General. UN, Geneva: UN.

2021. Wikipedia: Languages used on the internet. Wikipedia. October 14th. Accessed October 30th, 2021. https://en.wikipedia.org/wiki/Languages_used_on_the_Internet.