"The Jungle Is Neutral" – Newcomer Languages Face New Media

Nicholas Ostler
Foundation for Endangered Languages
172 Bailbrook Lane
Bath BA1 7AA
England
nostler@chibcha.demon.co.uk

Abstract

The origin and early history of research and development in machine translation might suggest that it is only interesting, or applicable, to the greatest of major languages. But founders' effects do not persist in eras of furious technical change, unless they concern essentially arbitrary aspects, such as technical standards. Machine translation, and language technologies more generally, may yet be very useful to minority languages, promoting and extending both their use and their status, in a world where there may be more than one dominant language.

The title quote is borrowed from F. Spencer Chapman's 1949 book on his experience of jungle warfare.

1 Introduction

This is a time of great political and economic turmoil, when the future power-structures of the world are unknown, clouded by changes that have not as yet run their course. One of these unknown future structures is the space that will be available to the world's languages. Will the world's rising powers such China, India, Russia and Brazil simply accept the current dispensation, and communicate in the relatively neutral medium of English, alien and costly though it may be to many of them? But then, have they any power to set up an alternative? Will the world's minority languages continue to yield ground to the dominant languages in their various countries, losing functions, and more and more failing to be picked up at all by rising generations? Or will the emerging situation, technological as well as social and political, offer them new options for survival and utility? In fact, study of the past history of synergies between languages and language technologies suggests some surprising possibilities for the future.

The modern position of the English language clearly owes much to a generalized technological revolution, the massive increase in immediate wealth and power that came from the use of mass production, fossil-fuel-burning industry, and world-shrinking transport and information-exchange; this revolution came about just as the United Kingdom, and then the United States of America, were spreading their political power, and their agents of enterprise, to every corner of the world. These, the switch to the technologies and the spread of the language, happened principally from the 18th to the 20th centuries.

Much more specifically, the application of computer technology to machine translation (MT), as it happens, owes its first surge of development to competition between the USA and the Soviet Union of the 1950s-60s, taking in the early Cold War and the Space Race. It was then adopted and extended by other significant scientific and technical powers of those days, seen for a time (the 1980s and early 1990s) as an important enabling technology by governments such as those of Japan and the European Union. As the principles of the technology became better understood, there even began a drive to create a meta-system which might generate new MT systems on the fly, as and when English-speakers needed access to any other "low-density language ".

This history might in itself suggest that MT is only interesting for, or in practice applicable to, the greatest of major languages – Russian, Japanese, the state languages of western Europe, Chi-

Proceedings of the 13th Annual Conference of the EAMT, pages 1–7,
Barcelona, May 2009

nese and above all to English. And these languages have certainly carried off the early laurels for research and exploration, and limited success, in the field.

But founder effects – defined as the continuing dominance of those who have pioneered a move, even as others join in, and make their own contributions to it – do not necessarily persist in eras of furious technical or political change. For example, English has indeed continued as the language of foreign colonists who came to dominate North America, even though the native tongues of the vast majority of later immigrants were different, mostly Slavic or Germanic. But Portuguese has not sustained its early (16th to 18th century) role as the lingua franca of trade and diplomacy round the Indian Ocean. The collapse of Portuguese mercantile control in the 17th century did not benefit linguistically the Dutch who subverted it, but it did leave the field clear for the later growth of English.

There is no simple rule, then, that decrees the long-term triumph of those who first take up a technical option, or a dominant position in a new world-order. This possibility, of latecomer dominance, is as applicable to the mastery and use of machine-translation technology as it is to the general survival and role of individual languages within the world system. If the two applications are put together, it may even be that the late-coming use of MT technology to give access to other languages, for large or small communities, will give new life-chances to such languages.

To explore these prospects, it is reasonable to ask, and try to answer, three key questions. Firstly, what does it take to unseat an established lingua franca, inhibiting its continuing transmission down the generations? Secondly, can the wider application of MT provide what is required to do this? Thirdly, is the future course of globalization likely to call for some lingua franca in the foreseeable future, whether it will be English, or some successor to it?

2 To Unseat a Lingua franca

The most interesting case of an established lingua franca which came unstuck is that of Latin in Europe after the Renaissance. It is interesting in that there was no evident competitor which supplanted it, and many aspects of the situation, as well as contemporary events, might have been expected to support its role, and indeed to enhance it.

In the late 15th century, the printing press became available in Europe, first in Germany and then in Italy. The first book to be printed was, unsurprisingly, the Vulgate Bible in Latin, and to start with, the vast majority of books that came off the presses, wherever they might be in Europe, were in Latin. The whole of Latin literature that had survived from the classical era was soon in print, and now that texts could be duplicated without error through mechanical production, textual criticism could become systematic, best editions could be reconstructed through comparing the various available manuscripts, and the results agreed as standard. In addition, with books becoming cheap, scholars might now be expected to purchase their own copies, and classes could learn from textbooks. Many of the early best sellers were indeed Latin textbooks. Furthermore, since users of Latin made up the only language-community that could be assumed to be 100% literate (always having learnt the language at school), and since they were distributed throughout Europe except in the orthodox zone (of east and south-east), there would have seemed no doubt at all that Latin would offer the best language in which to print books, and to produce them in the longest print-runs.

At just this time, European mariners – first Portuguese and Spanish, then French, Dutch, English and Danish – discovered the vast potential of the world beyond Europe's coastal waters. European settlements sprang up in the illiterate and sparsely populated lands of the Americas, as well as in the high developed markets of the Indian Ocean, and China beyond. All the captains of these expeditions, as educated men, knew Latin, and the first accounts of their discoveries, by such writers and Peter Martyr, were circulated round Europe in that language. Although the different European nations were in competition to set up these settlements and trading posts, the Catholics among them were explicitly charged by their Pope to win souls for Christ on these expeditions, something that would have been unthinkable for them without use of the Latin language, especially if a native priesthood was to be established and educated in the new lands.

Nevertheless, just when Latin – the textually-based language par excellence – had finally got its classical texts firmly defined and economi-

cally distributed, and seemed poised to travel with European venturers as they spread their interests, their faith and soon their control, round the wider world, it began to lose its pre-emptive dominance. Educated discourse began to be acceptable in vernacular languages, first in the leading powers of the west, France and then England, then in the powers of central Europe, such as the Netherlands, Germany and Italy, and finally in the peripheral powers of the East and North, such as Austria, Hungary, Poland and Sweden. The book markets that had sprung up all over the continent switched during 16th and 17th centuries from Latin to the various vernaculars: even in the New World, the printing-presses were producing texts in the indigenous languages, which the Spanish missionaries had just succeeded in analysing and reducing to (roman) script.

Latin failed for a variety of reasons. One reason may have been "insufficient globalization" of the market: excessive costs of book transportation round Europe, as against the costs of book production, which meant that publishers and booksellers stood to gain more from selling their print-runs close to the home, with local audiences who read in the vernacular, than to an elite, pan-European, market, who could read in Latin. Close at hand, there were always more vernacular readers than Latinists, and now that books were produced in quantity rather than one at a time, those numbers began to count. But in addition to economics, there was a power-shift going on between the classes. The elite were less and less traditionally educated clerics, and more and more urban bourgeoisie who had had a more practical, and vernacular education. National governments too, led off by France and England, as they distanced themselves from Church power, wished to discriminate in favour of their own vernaculars. Already in 1539, King François I had required by the Ordinance of Villers-Cotterêts that official documents, whether from courts or parish registers, should all be produced *en langage maternel françois et non autrement* – "in French mother tongue, and not otherwise", implicitly not in Latin.

What, then, had inhibited the transmission of Latin as a lingua franca? No single language had stepped in to take its place, but the power-structure of society had changed. As people educated without Latin came to assume greater influence (and the influence of the greatest Latin-using power, the Roman Catholic Church, declined) there was simply less call for skill in Latin. The international contacts which were facilitated by use of Latin as a common language were naturally diminished. But this was less crucial, in a new society where separate nation-states dominated in their own interests. The "founder effect", that had transmitted Latin through a good millennium, or 40 generations, when it was not close to the vernacular for much of Europe's population, had been undone.

Founder effects, a.k.a. the force of tradition, are stronger where either there is little cost in sustaining the past pattern (contrast the continuing expense in time and effort to induct new generations in Latin, effectively an artificial language), or the tradition is not at variance with some other new pressure (as Latin was in effect a barrier to entry for less educated bourgeois people). Hence notoriously, wheel gauges have been sustained at 4 ft 8½ in from the Roman empire and its road ruts to the US standard railroad gauge. What motive was there to change as one style of wheeled transport succeeded another? One could also note that Renaissance typographers of the 15th and 16th centuries, choosing the character styles for printing fonts simply took over the styles (**Gothic**, Roman and *Italic*) which were then in vogue in manuscript hands. They have been sustained ever since – though Gothic, the least readable, has lost much ground – since there is no more of the particular dynamic in manual pen movement which had previously driven the changes since the CAPITALS of the Classical age. Even more notoriously, the perverse QWERTY pattern of the English keyboard, invented in 1872, has survived a century of mechanical typing and the first 30 years or so of digital text entry. It is likely to continue to survive unless and until it comes to represent a barrier to entry to some sector of the population of would-be typists and writers, which – hitherto disenfranchised – is yet rising in influence.

This kind of situation is precisely what can be expected to provide at least opposition, and perhaps effective revolution, to the retention of English as a global lingua franca. What about the vast section of the world's population for whom the need to learn English is still a burdensome chore which they would prefer to avoid?

3 When Founder Effects Live On

As a digression, or an examination of the clinging power of a dead or dying lingua franca, we may note that features of an inherited system are not always rationalized away. If they are harmless, or in some way emphasize the (conservative) power of a favoured group, they may be preserved. Hence in the cuneiform ideographic writing which was invented to write Sumerian, but subsequently adopted to write Akkadian, the rebus principle is operated to give punning meanings to characters, using both these languages. But when the script was later adopted to write Elamite, Hittite and Ugaritic, the alien puns were retained (as 'Sumerograms' and 'Akkadograms') in the writing system, although the pronunciation in the new languages would have followed the meaning rather than any attempt to borrow the Sumerian or Akkadian words literally.

This principle, understandable in ideographic scripts like cuneiform or Chinese characters, actually continued to be followed after Aramaic and then Persian came to be written with purely alphabetic scripts. Since, pragmatically, texts written in Aramaic to Persian addressees were usually read out only in Persian translation, the Aramaic came to be seen as an indirect way of writing the Persian. When later, Persian itself came to be written in the Aramaic alphabet, it would be interspersed with large numbers of words written alphabetically in Aramaic, but each to be pronounced as the synonymous word in Persian. These so-called 'Aramaeograms' persisted in use long after Aramaic itself had been forgotten by most scribes.

In certain cases, the pragmatics of these alien carry-overs may be not burdensome or neutral, but even beneficial to receiving users. This has been the case with the alphabetical input of Chinese characters in Chinese and Japanese word-processing.

It had long been an unsolved problem in attempting directly to input Chinese characters from a keyboard that there were just too many keys; when an operator had to chose from an array of several thousand characters, location just took too long, so that direct input by pen was usually preferred. However, in an electronic context, language technologists at Toshiba discovered that candidate characters could be located much more efficiently – and at a speed acceptably close to real-time – if their phonetics were typed in alphabetically, and the reduced set of candidate characters was then used to select the actual character desired. Furthermore, the phonetics are more efficiently typed in using the Roman alphabet than the traditional Japanese phonetic syllabary, the *kana*: such Roman phonetics never appear in the resulting text, but they do facilitate the entry of characters, whether in Japanese or in Chinese.

So in fact, an arbitrary carry-over of apparently irrelevant technical details from one language's technology to another system may, by good luck or good judgment, in fact solve that other's persistent problem. This happens because deep learning, as well as loss of traditional skill, can result from the introduction and acceptance of new technologies from alien sources.

4 Is MT Equal to the Task?

This all provides some kind of answer to the first question: "what does it take to unseat an established lingua franca?" In essence, the answer is the context needs to change so that what was an advantage comes to be seen as a net liability. We proceed to the second question: "can the wider application of Machine Translation provide what is required to do this?" Can the availability of MT cause such a change to the surrounding context for international communication that continuing use of English will be undercut?

Prima facie, the answer to this is unpromising. It has been an unchanging truism of MT, almost from the beginning of its fifty-year history, that its results have been disappointing. The hope that inspired, and for a long time funded, MT was that it could provide a cheap, fast and high-quality substitute for human translators or interpreters, so that in effect the language barrier would go away. This has not happened, though the reasons for this disappointment are not clear and distinct.

Ironically, this ambiguity was dramatized most memorably for me by the Danzin report, which in 1990 evaluated the success of the European Union's 12-year-long EUROTRA project to produce a multilingual MT system among the (then nine) official languages of the Union.[1] Attending

[1] Danzin, A., Allén, S., Coltof, H., Recoque, A., Steusloff, H. and O'Leary, M. (1990) 'Eurotra Programme Assess-

a session of the management committee which oversaw EUROTRA, I was perplexed to note that there seemed to be a radical misunderstanding between two sets of delegates: it was accepted that the project had not delivered the functioning system which had been the goal of the project; but was the report as a whole supportive or dismissive of EUROTRA's work? Did it suggest that more work should be undertaken, or the whole project abandoned as a failure? Broadly, the delegates split along language lines, the Romance-language speakers taking the report as more positive.

As it happened, the report had been written in French, but many of the committee had only read the English translation. On a crucial summary page, I discovered that the report had characterized the project's work as 'insuffisant', whereas the English version had translated this as 'inadequate'.

Arguably, no mistake had been made by the translator, in truth-conditional meaning, or even in style: when quality rather than quantity is being judged, it is much more natural in English to say 'inadequate' than 'insufficient'. But what a difference in connotation! What is called *insufficient* naturally needs to be supplemented, but what is termed *inadequate* is usually being roundly condemned. There could hardly be a clearer example of the treacherous nature of translation, even by the wise for the wise.

But what of MT itself? Have its results been insufficient or inadequate? It is very hard to give a final decision, although it is fairly clear that one concept which underlay most of the early work was basically inadequate.

The original rule-based models of MT which dominated research until the 1990s were essentially attempts to automate the "grammar-translation" approach to language learning. The syntactic rules of the various languages could be represented and programmed, and translation equivalents could be stipulated for lexical items, and for the semantic content of the various constructions. Proper names required access to vast encyclopaedias and gazetteers, seemingly never complete. The systems got larger and larger, and more cumbersome, harder to direct effectively.

Another response which became popular in the 1990s was to increase the role of machine intelligence, allowing inference engines to derive their own rules from exposure to vast amounts of translation equivalence data. This was computer equivalent of the "natural" method of language learning, essentially waiting for competence to arise unconsciously from massive exposure to language data. Perhaps the problems of performance here would ultimately yield as computers got exponentially faster and cheaper.

Yet systems remained lacking any general models which could represent the meaning of texts in the writer's or the reader's understanding, as they flitted from text to text or context to context. Nor was there any general means of selecting appropriate equivalents when language was used metaphorically. It seemed to prove that in practice, it was impossible to divorce the syntactic part of language processing from modelling the meaning of particular texts.

While this technical struggle continued unabated, the actual users of machine translation were devising their own *pis aller*, their own make-do approaches to handling what was available. The technology has begun to come into its own as a support system for human translators, allowing them to evade drudgery of repetitive translation and dictionary look-up. And the field of application has also been transformed by the vast quantities of foreign language text that are now available across the Internet. Automatic systems are proving useful aids to web-surfers, looking for relevant content in foreign disguise, rather than for clean translations of specific documents.

The fact that the technology is already being used serendipitously (rather than developed) by informal and linguistically-informed users is a first sign, I believe, of the actual future that awaits MT, and it is not an inglorious one.

The reason for the chronic dissatisfaction with MT's performance (especially among monolingual Anglophones, one may say) is that it has always been approached from a monolingual point of view, as a tool that is supposed to eliminate language barriers – i.e. as a means of converting all the alien codes into some readily understandable home language. This is the true in-

ment Report', Commission of the European Communities, DG-XIII, March 1990. French original as: Rapport Danzin : Document COM (90) 289 final.

adequacy in our traditional approach to MT. It is comparable to the lingua franca solution to multilingualism: let us find a common means – be it Latin, be it English, be it Esperanto – in which all the languages' texts' meanings can be represented. But a lingua franca is a practical solution in terms of a single language. MT has failed to do anything comparable, at least consistently, or reliably, or at a standard where the user familiar with English (or whatever target language is being attempted) is well satisfied.

But even in the forms currently available over the Internet, MT (and many other ad-hoc devices) already provide a vast number of tools to access and penetrate texts in unknown languages. It is debatable whether this is truly translation, and in many cases, the help is only accessible to those with a partial knowledge of the source language. But it does mean that, increasingly, partial understanding is becoming available, of texts that would in the past have been totally closed books.

Another personal anecdote may illuminate the situation that is emerging. According to Wikipedia, I "can functionally speak 26 different languages", this improbable claim having somehow emerged from the publicity department at Bloomsbury USA when they were designing the paperback jacket for my book *Empires of the Word*. This is harder to disprove than you might think, since paired with 'functionally' the verb 'speak' seems to be meant as equivalent to 'have command of'. I cannot know precisely which languages are intended here, but it is true that I have derived useful, and true, information from at least that many languages while working on that book and others. I cannot 'speak Chinese', but I was able to provide a phonetic transcription of texts from Confucius's *Analects*. Using other materials from the Internet I could gloss passages of Akkadian cuneiform and Egyptian hieroglyphs, locate relevant text in Sumerian, Persian and Portuguese, apply dialect changes to and parse Mexican Nahuatl and Palestinian Aramaic. In none of these languages can I boast any sort of fluency. But, in sum, my point is this: if you embrace the presence of foreign languages, and are interested enough to try to come to grips with them, more and more you will find the wherewithal to do so available to you (usually free of charge) on the Internet.

The set of language tools of which MT is a leading member are not available as a seamless suite which enables English users to look though the obscuring dark glass of foreign language to their crystal-pure meaning beneath, even if, here and there, web-page translation may in some cases be good enough to give this illusion. They are not, and cannot be, the realization of the monolingual dream of MT. But they are very much better than nothing, and – coupled with the right attitude to the point and value of foreign languages – they may be crucial aids to inter-lingual communication.

It is possible to look ahead into this dynamically improving, and enriching, world of inter-lingual electronic media. Just as the print-revolution – and various other social revolutions associated with urbanization – changed the ground-rules of communication among Europeans in the 16th century, so modern electronic technology is set to change the ancient need for a single lingua franca for all who wish to participate directly in the main international conversation. In brief, if electronics can remove the requirement for a human intermediary to interpret or translate, the frustrations of the language barrier may be overcome without any universal shared medium beyond compatible software. Recorded speeches and printed texts will become virtual media, accessible through whatever language the listener or speaker prefers. Machine translation, and language technologies more generally, may yet be very useful to minority languages, promoting and extending both their use and their status.

5 Will there be a Lingua franca?

We turn now to our third question: is the future course of globalization, as we currently perceive it, likely to call for a lingua franca in the long-term, whether it will be English, or some successor to it?

First of all, we can note that the forces making for the spread of English will soon peak, and the sequel will be a long retrenchment, as auxiliary English comes to be used less widely. Power, prestige, position, population, even practicality will never again favour English as they have in the 19th, 20th and early 21st centuries. If the world system remains dynamic, English will very much need to look to its laurels.

English does not even have all the advantages of position that Latin once had. Unlike Latin, once peerless in the world it knew, it does have com-

petitors – vernacular languages with hundreds of millions of speakers and intercontinental spread; and it has peaked in an age before some of their home populations have even reached their economic prime, China, India, Indonesia, Brazil, perhaps even Russia.

It will be strange if a country like India stays loyal to English once there is any serious trickle-down of its new and growing wealth. Already, its objective to double higher education by 2015 (to 15% of the age cohort) is putting pressure on the proportion educated in English. There is an issue here to be resolved, even if the outcome is not clear in advance. Perhaps – like Latin America in the 19th century – it will hold on to the language of its former colonists, and content itself paying lip-service to *indigenismo*, its heroic native roots. But regional languages are entrenched in the government of India, as they never were in Latin America: more likely, as in early modern Europe, it will be the elite language which has to yield. The bonds that tie India to English are far weaker than those of tradition and sentiment which once tied Europe to Latin.

It is often assumed that power politics and the global competition among great states will naturally be reflected linguistically. Hence the current international ubiquity of English is seen as a reflection of US 'unipolarity'. If this is doomed to pass, then it must, it is presumed, be followed by some other common language. The choice falls most obviously on Chinese, since this is already the world language with most speakers, and on current trends the Chinese economy is growing to be the largest in the world. Certainly the international importance of Chinese is very likely to grow, and as the Chinese become richer and more influential internationally, their concern to participate in the world on terms set by Anglo-Saxons will diminish. There is already evidence of this, highly predictable, change. The 2008 Pew Global Attitudes Survey in China reported that "Most Chinese (77%) agree that 'children need to learn English to succeed in the world today,' ... down substantially from 2002, when 92% agreed with this view."

However, this is only a small part of the coming changes. There are many parts of the world where English is not part of the national tradition, and they include the main countries about to increase in population size (sub-Saharan Africa, the Middle East) or relative wealth and influence (China, Russia, Brazil). Such a world is moving not to English or monolingualism, but it is hard to choose among these contenders for future linguistic influence. Very likely, the world is moving towards a much more multilingual, diverse, and potentially incalculable future.

6 Conclusion

But when technological ground is continually being ploughed up, there is cope for interesting new crops to germinate and flourish. Radical multilingualism may be one such crop, in a field-system (or a jungle) of pervasive digital technology. And monolingualism – privileging the stale over the fresh, and the few over the many – may well be an ideology whose time is passing.