Communication Technology, Linguistic Technology and the Multilingual Individual[1]

Annie Zaenen
Rank Xerox Research Centre
6, chemin de Maupertuis
38240 Meylan, France
e-mail: `Annie.Zaenen@Xerox.fr`

Geoff Nunberg
Palo Alto Research Center

It is a common place observation that we live in a more and more multilingual environment. This is true even if we ignore the fact that people have become more mobile and more people spend some significant part of their lives in linguistic communities different from the one they were born in and focus our attention on people who spend most of their life in their country of origin. Books, movies and television have exposed most people to some foreign language use for some time but now the possibilities created by new communication technologies, embodied for instance in the internet, will change the way multilingualism influences our lives: a greater variety of documents in foreign languages will become available and the boundary between passive and active use of foreign languages will become blurred. As long as books and newspapers were the only documents in a foreign language that one was exposed to at home, a good reading knowledge of that language was sufficient to gain an understanding of the culture and political life or to follow developments in science and business. With movies and television a passive spoken competence became more important, though help was usually provided in the form of titles or dubbing. Now, however, even a Dutchman who stays home in the Netherlands or a Frenchman who never leaves Grenoble is challenged to interact directly with people whose native language is different from his own: to be capable of only reading of whatever information appears on the internet and not be able to take part in discussions is bound to be extremely frustrating. To compound the difficulty, at the same time the type of language that is used in these interchanges is quite different from the language found in books and even in newspapers: at once more colloquial more context-dependent,

[1]This is a revised version of an invited talk given by the first author at the CLIN'94 meeting. Although the material was revised and enriched, no attempt has been made to hide the origin of this contribution as a talk or to turn this into a 'scientific' paper: caveat emptor!

and hence more challenging for non-natives, particularly if they have had only the kind of traditional foreign-language schooling that emphasizes literary and print uses of the language.

Computational linguists have both a practical and theoretical interest in understanding where this evolution will lead us. In what follows, we take the perspective of the individual language user and will not talk about institutions and their multilingual needs. We also concentrate mainly on the European and North American context as we are not at sufficiently familiar with the situation elsewhere. First we will look at the use of multilingualism in scientific discourse in both a traditional situation and in the situation that is created by new communication technology, then we will look at the broader picture of nonprofessional language use. We will close by pointing to some ways in which linguistic technology might contribute to making life easier in a multilingual environment.

We start from two assumptions which seem to us self-evident, and which we will not in any case try to defend:

- scientific, economic and cultural interactions will continue to become more global and more intense
- members of important vernacular speech-communities will not give up their native language.

Twenty (or even a hundred) years from now we will not all speak exclusively English or Japanese or Chinese. For this situation to be guaranteed, no explicit government action is necessary: culture is transmitted from generation to generation in a way that makes it change at a much slower rate than political and economic dominance patterns. The US might already be on its way out as the dominant economic/political power and we have not yet had the time to adapt to the dominance of English even if we had wanted to! In fact the use of English is receding to a slight extent in the third world as a result of increasing literacy in the vernacular, at least as a vehicle for education and literature.

Assuming this state of affairs, several questions can be asked: in which areas will one language dominate, how do we manage bilingualism in those areas; in which areas will no language dominate? How do we manage multilingualism in those areas and will there be areas in which there are intermediate, preferred but not uniquely dominating, languages?

Before going further into the discussion, let's agree on some terminology and distinguish the following types of languages:

a. Vernacular languages. The languages of major national communities with developed commercial and scientific communities which are currently the medium of substantial scientific publication; e.g., Dutch, French, Italian.

b. Minority Languages/Secondary vernacular languages. The languages of sub-national communities or of less dominant national communities (Catalan, Greek, Slovenian)

c. Vehicular languages. Languages with a substantial history of use as a means of international communication. In the Western context these are chiefly English, French, German, Russian; Spanish, Portuguese

Even if we admit the dominance of English (at the present time) in the domain of scientific discourse, this does not mean that all other languages will be equally important or unimportant in that area. One can certainly imagine that around the Pacific, another language, say Japanese, might be used as a vehicular language in that domain, on the one hand because it is perceived as less foreign than English and on the other because the exchange of information is more intense with native speakers of that language than with the scientific community on a world scale.

Is it likely that another European language will play such an intermediate role, in either science or culture? Traditionally, to achieve such a role, the language has to be the language of a dominant economic/political group or the vehicle of an important social project. France from the 17th to the early 19th century could be seen to play the latter role and as a colonial power it has also played the former, though at present its relative importance in both domains is considerably diminished. A new way intermediate vehicular languages might develop is if practical considerations in the EU lead to the creation of a two-tier system where documents are no longer made available in all the languages but where the political weight of France, Germany and some other countries will require the availability of translations in those language while countries like the Netherlands, Denmark, etc. accept that not all documents be available in their vernacular. Such a policy could lead to a situation in which all the 'small' language groups converge on English, but it could also lead to a situation in which, let us say, Northern and Central Europe converge on German and Southern Europe on French, or some kind of Romance-based pidgin that one now witnesses the informal use of in conversations. This scenario has not taken a clear shape in Europe. Electronic communication means might create yet another need for intermediate vehicular languages. We will discuss this possibility below.

1 Traditional Scientific Communication

It seems clear that in the area of traditional communication of scientific and technical results (and of high-level economic exchanges) English[2] will be the dominant language for some time to come, at least in the 'Western' world. But what does this mean? Mainly that important scientific results need to

[2] and in the area of lazy tourism: if you go on a quick tourist trip to let's say Sweden and then three months later on an equally quick trip to Spain, you might as well learn English and not bother with Swedish or Spanish: to order laet oel or paella English will suffice.

be communicated to the international scientific community in English (though this need is felt more strongly in some fields, like astrophysics, than in others, like medicine, where substantial proportions of the literature are still produced in French, Spanish, Italian, and so forth.) Among the multiple reasons for this, we will mention only one that might not immediately come to mind: to protect the non-native English researcher from plagiarism! Does this mean that all scientific activity will take place in English? Of course not. Does it mean that it would be desirable that all scientific activity should take place in English? We think not, for several reasons:

- It can be intellectually rather crippling to try to formulate one's thoughts in a foreign language, where one does not have the full range of lexical and grammatical distinctions available to hand; it is easier to think in one's native language and translate after the fact. Better that something be lost in translation than never be present at all.
- It is mainly in one's own language that one evaluates questions of methodology and epistemology as such discussions tend to take place in the cafeteria or the neighboring cafe.
- As we said, we start from the assumption that these vernaculars will remain, so that there will be a need for scientific discourse in each of these languages, and hence for the linguistic means for this discourse. The proposal once made by a Dutch minister, to conduct all Dutch university courses in English, strikes us as extremely dangerous in that it would make it much more difficult for the community as a whole to keep in touch with scientific activity in a varying degrees of intensity. If there is no instruction in the vernacular there is no need for the publication of textbooks or other basic scientific material, so that the general public would most likely have to do pretty quickly with pure vulgarisation and be excluded from early and unorganised exposure to scientific results and scientific thinking. This would reverse whatever one thought of as the benefits of going from Latin to vernacular languages three centuries ago, when the first vernacular scientific journals like the Philosophical Transactions and the Journal des S cavans first appeared, as vehicles to deliver the results of the new science to the larger public. In all of this in fact there is a curious reversal of history. In the Netherlands, for example, figures like Simon Stevin and Isaac Beeckman argued that a shift from Latin to Dutch as a vehicle for scientific discourse would open up science to craftsman and others who had no formal education; and the universities' failure to adopt this course, while arguably making them more accessible to foreign scholars, also had the effect of closing off science to classes that had no classical education. See Hackmann (1975) and also Dijksterhuis (1970), p. 126–9, based on Stevin's "Uytspraeck over de Weerdicheyt der Duytsche spraeck" (1586).

How much English is or will be used in primary scientific publications will depend on the use that is made of these publications: in some fields most readers are also writers, and if they feel there is an advantage in writing in English they will also be able to read English. In other fields, for instance, medicine, however, a large proportion of the readers are not writers, and so cannot be presumed to be familiar with a vehicular language like English. There are also many fields like economics where publication in the vernacular is crucial because the primary audience is national (e.g., policy makers). In these fields, publications of recent scientific results in the vernacular will remain important.

In fact, in some cases, publication in vernaculars is increasing. The proportion of the chemistry literature in Japanese is over 10% and growing: while more Japanese are writing in English the absolute number of Japanese chemists is growing still faster than that.

2 The Likely Impact of Electronic Media

1. In theory, the electronic distribution of scientific publications increases the efficiencies of the marketplace: the place of publication does not determine circulation as much as with print. E.g., an article in a medical journal published in New Zealand or Paris is more accessible to a reader in Chicago or Singapore. This could militate for the increased use of vehicular languages, and particularly English, since authors have more incentive to direct their attention to the largest linguistic markets, independent of location.

But, there is reason to suspect that the use of English in scientific publications is close to a maximum level in many areas, for several reasons:

- It is not clear that having a larger absolute audience increases the readership of any article, particularly in fields where every reader is an author. (The average audience for a paper in computational linguistics, for example, is calculated by taking the number of papers that the average linguist reads by the number of papers she writes, and this amount is independent of the size of the community). This will most likely not be influenced by the change of the channel of distribution.

- As already noted, in some fields like medicine, there is a relatively large readership that reads only or chiefly in the vernacular. And outside of the top ranks of science, it isn't clear that there are greater professional rewards for publication in English to offset the greater difficulty in writing in that language. (It should be remembered that the average scientist writes primarily to achieve institutional advancement, and that the modal number of citations for any given scientific article is zero.) Again, this will not be influenced by the change of the channel of distribution.

- Certain political factors are likely to favor the continued use of languages like French, German and Russian as vehicular languages for certain kinds of discussion (all the more since applications for funding must be made in the national language, and since there have been moves in some nations, such as France, to require that state-supported research be published in the national language). In addition, the singular growth of the Japanese community favors continued use of that language by Japanese researchers, since for most of them the advantages of publication in English are marginal.

For any given reader, the net effect of the new distribution channels is probably to increase the amount of material available in languages other than her own. As an example, for an American scientist in a print culture, the preponderance of non-English material generally increases inversely with accessibility (e.g., English sources have traditionally been more accessible than journals in French, Japanese, or Czech). These more remote publications are now more accessible. Even if the proportion of English-language publications in French or Japanese journals increases (as is now happening), the increased availability of these journals means that an English-speaking researcher has more non-English material to deal with. The sharpest increases in written scientific communication are likely to take place in communities where reductions in the marginal costs of publication have the most significant effects. E.g., we would expect a greater proportional increase in the amount of scientific and commercial uses of Arabic, Hebrew, or Hungarian than in English, particularly in forms that fall short of traditional, formal publication. (This development is likely to be paralleled in the production printed materials. Printing on demand makes possible smaller press runs and gives new life to forms of publication that have been marginal with traditional print technologies, which again is particularly to the benefit of smaller linguistic communities.)

Note also that this factor increases not just the efficiency of international markets but also of national markets. It makes possible new forms of scientific and commercial publication in many communities, e.g., like Greece or Hungary, where there have heretofore been only limited resources for scientific publication in the vernacular.

2. Electronic distribution also increases the variety of material available. In the scientific world publication is no longer limited to summaries of results and the like. The American Physical Society in the "Vision 2020" report proposes to put on-line experimental notes, working papers, raw data, etc. There are various proposals to make referees comments available and on-line discussions of results are possible This could increase the amount of material available in nonvehicular languages, since the secondary material that supports a publication is more likely to be in the vernacular than the publication itself, for reasons we discussed above.

3. Electronic distribution has created new communicative forms: newsgroups, moderated lists, etc. On the one hand, these groups increase the amount of international communication in vehicular languages, particularly given the preponderance at this time of English users on the net: in July 1994 there were more than 2,000,000 internet addresses in the US and 400,000 more in other anglophone countries, but only 170,000 in Germany and Austria together, 23,000 in Italy, 70,000 in France. Adding half of Canada, Belgium and Switzerland would give a total of 170,000 for the francophone countries.

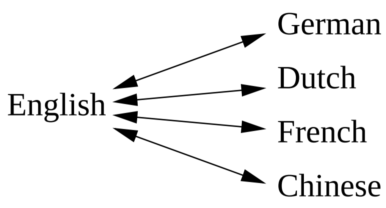
These numbers are misleading, however, as these disparities are not stable. The number of net addresses in the US has increased 38% from Jan 94 to July 94, against 117% in France, 147% in Belgium, 169% in the Czech Republic, 43% in Holland and 142% in Russia.

This increase has two effects on the use of languages. First, it creates a new domain of international communication that people have to deal with, one with very different rhetorical and communicative norms from those of formal publications.

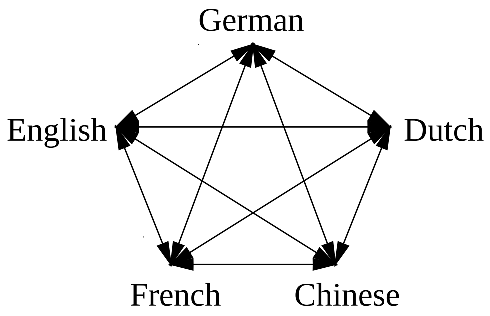
Second, it makes possible the formation of new vernacular communities. Among other things, the net creates a world wide cafeteria. Geography can be ignored and, for the moment, institutionalized channels. But in which language does one discuss in these electronic cafeterias? Here, there is most likely a chance for communities like the 'francophonie', provided enough of their members have access to these new media. Their appeal might not be limited to native speakers: a Flemish- or Italian-speaking researcher who knows French somewhat better than English may have new reasons to participate in the French rather than English net discussions (though he or she will probably still want to publish international papers in English). To a certain extent, then, the net may promote the use of the second-tier vehicular languages (French, German, Russian) as mediums for informal scientific communication, if not for formal publication proper.

3 Non-scientific Communication

Having admitted the need for a large segment of the population to have access in the scientific domain to information in English and to be able to contribute to the information available in that domain in English, should we conclude that the best model for multilingualism is as follows?



or rather as in the following schema.



Obviously the first schema involves less work: the two schemas contrast like the 'interlingua' and the 'transfer' schema in translation and at first blush the arguments that can be made in favor of interlingua carry over. But we have already seen that there might be arguments in favor of secondary vehicular languages which would complicate this schedule. Culturally and socially it also seems to be a very unsatisfying hypothesis: given that communication in a foreign language implies some loss of information, at least of connotation, there will be a double loss if native speakers of e.g. French and Dutch only communicate with each other through English, whereas in direct communication there will be only one loss. Moreover, under the second model the different language groups will stay attuned to each other's cultures in a reasonably equal way. Under the first model the only culture(s) people will know other than their own in a reasonably direct way is/are the one(s) that have English as their mode of expression. This might lead to an impoverished view of the world, and possibly a dangerous one: 'onbekend is onbemind', lack of knowledge leads to insensitivity.

To give just a couple of examples, for an American, unable to read French, it is nearly impossible to get a reasonable idea about the debate on the Islamic veil in France. We are not defending a Whorfian view of language here, but simply noting that a lot of cultural background information about the problem is never given in the English language press. Conversely, for somebody in France who doesn't read English, it is impossible to get a reasonable idea about the debate on political correctness, as again the local press interprets it in a very specific way, which the cultural group affected might think of as unfair. We give these examples between English and French as the cases we are most familiar with but of course similar problems arise when looking at Germany from France or vice versa.

Even in the "exact sciences", moreover, there are epistemological consequences to certain modes of expressions. (The physicist Jean-Marc Lévy-Leblond has made this point graphically by noting that when the first French translations of Heisenberg appeared, his "Unbestimmtheit" was translated, accurately, by "indétermination"; later, under the influence of English, the less accurate term "incertitude" has been used, a translation that leads, as he puts it, to "banal and deceptive interpretations" of the principle.) (Lévy-Leblond 1994) So while there might be a good case to be made for a universal language of science, along with the national languages of science and secondary vehicular languages for scientific discussion (if not so clearly for traditional publications), it is also possible to make a good case for direct interactions based on a variety of foreign languages in the cultural and social domains, which among other things can help to clarify the epistemological issues which surround both the everyday practice of science and its wider cultural reception.

4 To Summarize:

1. Independent of the absolute growth of international scientific and commercial activity (as measured, e.g., by numbers of participants or whatever), there is likely to be a much sharper increase in the overall amount of written material available.

2. The preponderance of this new material will be in vehicular languages, particularly English, in even greater proportions than present scientific publications.

3. The sheer increase of non-English material will be such that English speakers will have a great deal more non-English material to deal with. Speakers of English and other vehicular languages will have to deal with a greater range and amount of non-English material.

4. Speakers whose native language is other than English will have all the more reason to be familiar with English and additionally with some other major vernacular language, and not just the formal registers used in traditional scientific communication, but the informal varieties used on the net, a situation that leads to new communication problems closer to those of real life spoken language.

5. Even in a scientific and a commercial world dominated by one language there remain good arguments for broader forms of multilingualism.

The need for more and more people to be familiar with more and more languages then will not diminish as some people have supposed. In spite of this, however, there is little evidence of any sharp increase in individual multilingualism within our communities. (And in more than a few communities, indeed, individual multilingualism appears to be threatened, partly in consequence of economic pressures on the schools that decrease the resources available for foreign-language teaching, partly as a response to political concerns about immigration — cf. the “English-only” movement in the United States.)

5 Can Computational Linguistics Help in This Respect?

We can envision two extreme solutions to the problems described above: perfect and ubiquously available translation or the education of perfect bi- or rather multilinguals. From the beginning on computational linguistics has tried to realize the first solution. Through the years it has, however, become clear that machine translation is an exceedingly difficult problem. Computer methods for foreign language learning have been less investigated but are clearly not having perfect results, and human teachers aren't either. Does this mean that computational linguistics has to give up all ambitions in the multilingual domain? This depends on what one wants. If the requirement is perfect translation or methods to create perfect multilinguals, the answer is yes, but several other possibilities remain, which fall short of perfection but may go a long way to resolving the difficulties.

Let's summarize the types of needs we have identified:

technical / scientific
- reading reports / articles
- writing / presenting reports / articles
- searching for information
- asking for information (orally / in writing)
- exchanges of opinion about scientific matters, collaborative efforts
- executing instructions

.....

cultural / political
- accessing news / information about current affairs (papers, radio, television, ...)
- discussing 'states of the world'
- reading literature

.....

6 Which Tools Could Be Developed to Help in These Contexts?[3]

1. As the number of non-English publications and texts will sharply increase, even if the relative proportion of English texts also increases, IR tools with multilingual capabilities. The need for more sophisticated IR tools is exacerbated by the enormous variation in the quality and type of electronic publications, and by the absence of a standardized format (e.g., with abstracts, summaries, etc.) that facilitates searching.

[3] We ignore spoken language capabilities because we are completely incompetent in that domain.

2. To the extent that virtually all of the new material produced will be available on-line, it will be amenable to computationalized translation aids, which makes it possible to improve the return on investment for the human foreign language user: with these aids what is learned can become useful much earlier than was the case until now, mainly because sophisticated help can be developed that cuts down on the amount of lexical information that has to be digested before texts in the foreign language can be understood. The same means can also be used to give more information about colloquial language use.

3. Tools that help with the generation of text in a foreign language. Here it is less clear how computers can be useful in the near future but some interactive translation tools can be developed that help the user produce limited texts in a foreign language. With such tools, the user translates her own text getting source language feedback on what the system produces in the target language. This will not take all frustration out of the interaction but can reduce it.

Has computational linguistics played an important role in developing such tools? It seems to us it has played less of a role than it could have. Most of the effort of what is called computational linguistics seems to have been concentrated in the area of high level linguistic analysis. Of course high-level analysis is important in a number of areas that have not been the topic of this paper. But we suspect that one of the reasons it has been privileged is that it is more easily integrated with linguistic theory and with challenging parsing/generation questions which have become the traditional subject matter of computational linguistics. Limiting one's attention in that way, however, betrays a very shortsighted view of what is interesting. For instance, at Xerox we have been developing a foreign language understanding aid of the type alluded to under 2 above. It is clear that this tool doesn't incorporate any deep insights derived from theories on grammatical functions or anaphoric relations, nor does it propose any new variation on the Earley algorithm but it does incorporate the results of ten years of theoretical research in finite-state theory and technology, which in turn is based on a couple of basic if not mainstream insights in the nature of phonology.

What is "theory" and what is "application" is more relative than one tends to think: when one of us was teaching what I call theoretical linguistics, she had a colleague who characterized what she was doing as "applied linguistics" because to him theoretical linguistics was a branch of mathematics. Computational linguists might have more fun if they were less morbidly obsessed with the distinction between the applied and theoretical aspects of language technology.

References

Dijksterhuis, E. J. (1970). *Simon Stevin: Science in the Netherlands around 1600*. The Hague: Nijhoff.

Hackmann, W. D. (1975). The growth of science in the Netherlands. In Crosland, M., editor, *The Emergence of Science in Western Europe*, 89–109. New York: Macmillan.

Lévy-Leblond, J.-M. (1994). La langue tire la science. In *Colloquium on 'Science et Langues en Europe'*. Paris.