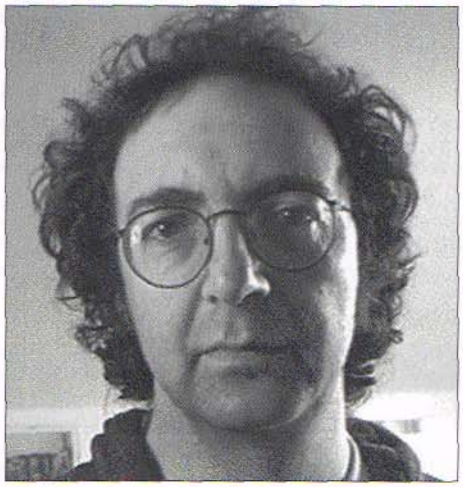
Language Resources and Minority Languages

Harold Somers (Department of Language Engineering, UMIST, Manchester)

examines the problems of technological disadvantage

INTRODUCTION

While the availability of Language Engineering (LE) products and resources for the world's "major" languages steadily increases, including Machine Translation (MT) systems, Computer-Aided Translation (CAT) systems, on-line dictionaries, thesauri, and so on, there remains a major gap as regards less widely-used languages. Missing are not only these kinds of products, but even simple tools like spelling- and grammar-checkers and, for languages which use idiosyncratic writing systems, even word-processing software.



Harold Somers - Department of Language Engineering, UMIST, Manchester

Because of accidents of world politics as much as anything else, languages fall into three or four "league divisions", reflecting the computational resources available for them. These divisions correspond to a certain extent to the ranking of languages according to their world-wide "influence", suggested by George Weber recently in these pages, though the inclusion of Hindi/Urdu in Weber's list is of special interest for our purposes.

This article will identify which languages are more or less badly served by LE, and some of the reasons behind this. It will also consider the sociological impact of "LE imperialism" in relation to minority languages in the UK. We will also try to make some proposals for what can be done about the situation. Recognising that the development of LE products for a new language is rarely a trivial matter, we will investigate some techniques that can make the task more manageable, or more feasible, including customising from resources for related languages, the possible use of software localisation tools, and the use of "knowledge extraction" techniques from machine-readable corpora.

MINORITY LANGUAGES IN THE UK

The UK is nominally an English-speaking country, with small regions where the indigenous Celtic languages are more or less widely spoken as the first-language. However, a more realistic linguistic profile of the UK must take into account the large areas of the country where there are significant groups of people speaking non-indigenous minority languages (NIMLs). According to the 1991 Census, ethnic minorities form about six per cent of the population of Great Britain. Across the country, languages from the Indian subcontinent, as well as Cantonese, are widely spoken; other NIMLs are more regionally concentrated, e.g. Greek and Turkish in London. While second- and third-generation immigrants are largely proficient in English, having received their schooling in this country, new immigrants as well as older members of the immigrant communities - especially women - are often functionally illiterate in English, even if they are long-term residents. For example, in a 1994 study, Kai Rudat reported that the numbers of Asians who rate their own English reading ability as "None" is 24% for Indians, 37% for Pakistanis, and 48% for Bangladeshis, but figures for certain subgroups according to sex and age were as high as 93% (Pakistani women over the age of 50) and 96% (Bangladeshi women over 50).

Many local councils, particularly in urban areas, recognise this, and maintain language departments to provide translation and interpreting services with in-house staff, as well as lists of freelance translators. Their work includes translating information leaflets about community services, but also one-off jobs where individuals are involved, for example, in court proceedings. Apart from serving the immigrant communities, refugees and, particularly in the capital, asylum seekers, bring with them language needs that are being addressed by local government agencies.

Internal Translation Needs in the UK

The range of languages handled by these agencies is impressively large. While the NIMLs account for a large percentage of the volume, there is an increasing volume of translation work relating to refugees and asylum seekers. For example, Manchester City Council employs five in-house translators to cover Urdu, two each for Cantonese and Bangla, one each for Punjabi and Hindi, and a part-timer for Gujurati. In addition, translators for Somali, Vietnamese, Arabic and Bosnian are employed for

the needs of refugees. In the 1995/96 accounting year just over half a million words of English were translated into various languages, this figure rising to more than 870,000 words in 1996/97. In Newcastle-upon-Tyne, the top seven languages translated by the Council's Translation Services are Bangla, Hindi, Punjabi, Urdu, Cantonese, Arabic and Farsi. The London Borough of Camden Language Services section had 1,803 translation and interpreting "jobs" in the period January to September 1997, involving 38 different languages. Table 1 shows the totals for the top 20 languages. Presumably, most of the work involves translation *into* these languages, though there may be a certain amount of translation into English too. In addition, the Language Services provide interpreters when necessary.

A comparison of the situation in Manchester, Newcastle and Camden shows how much regional variation there is. In Camden, Urdu was required only three times in the period covered by our statistics, while in Manchester it is by a long way the most needed language. On the other hand, Polish accounts for nearly a fifth of the demand in Camden, almost equal with Bangla. But what is striking in each case, and in other authorities we have spoken to, is the range of languages, and in particular the need for languages outside the usual subset in which translators typically translate for the business community.

Computational Requirements for NIMLs

Just like translations in the private sector, "public service" translations come in all shapes and sizes. Some are one-off jobs relating to legal proceedings or the provision of social services; others concern the dissemination of information to the general public. Again, just as in the private sector, some texts may amount to updates of previously translated material, may contain passages that are similar or identical to other texts that have already been translated, or may be internally quite repetitive.

Apart from printed documents, texts can be found on computerised media such as the information screens on bank Automatic Teller Machines, often provided in a variety of European languages, presumably for the benefit of tourists, but rarely, if ever, in NIMLs; community information screens in Town Halls; job availability announcements, now available in computerised systems in some places, and so on.

COMPUTATIONAL RESOURCES FOR "EXOTIC" LANGUAGES

Language-relevant computational resources are certainly on the increase. Various magazines aimed at translators, such as this one, regularly list new products and advances in existing products. Some periodically include software resources guides. But just a glance at these publications reveals an overwhelming concentration on the "superleague" languages which are seen as important for world-wide trade: the major European languages (French, Russian, Italian, German, Spanish, sometimes affectionately known as "FRIGS") plus Japanese, Chinese (i.e. Mandarin), Korean and, to a certain extent, Arabic. Interestingly, along with English, all of these bar Italian and Korean are in Weber's "top 10" languages, together with Portuguese (for the Brazilian market) and, as mentioned before, Hindi/Urdu, which Weber treats as a single language. Their market is the translation of documentation for products, commercial communications, and, especially recently, web-pages. Of course translation, like any other service industry, must be governed by market forces; but the languages that are of interest to commerce form an almost empty intersection with those of interest to government agencies dealing with the ethnic communities, refugees and asylum seekers.

LANGUAGE	TOTAL	%
Bangla	385	21.35
Polish	327	18.14
French	172	9.54
Somali	165	9.15
Romanes	97	5.38
Spanish	87	4.83
Albanian	84	4.66
Russian	60	3.33
Arabic	53	2.94
Farsi	44	2.44
Lingala	37	2.05
Czech	35	1.94
Tigrignan	29	1.61
Portuguese	27	1.50
Turkish	27	1.50
Sylheti	26	1.44
Greek	22	1.22
Chinese	20	1.11
Italian	17	0.94
Romanian	15	0.83

TABLE 1
Translating and interpreting jobs by the London Borough of Camden Language Services, January to September 1997: top 20 languages. Source: Camden Language Service Statistics.

A recently published directory of LE resources lists over 1200 software products, and includes a useful index on a language-by-language basis. Table 2 shows the provision of translation-relevant LE resources for some of the languages identified above as being of interest to us. What is immediately noticeable from Table 2 is the number of languages for which the provision is largely limited to the obvious non-language-specific, such as fonts and word-processors for Serbocroat and Welsh, for example, which need only to have the Roman alphabet and a few diacritics. Notably, Urdu and Hindi, which are among the top three significant UK NIMLs are not explicitly provided for: they are not even listed in Hearn, while in a recent LE software advertising supplement, they are only listed under fonts and word-processors.

Let us consider in a little more detail each of the categories listed in Table 2, and how we could go about providing the missing resources.

DEVELOPING NEW LE RESOURCES

In this section we will review the prospects of developing LE resources for all these languages and consider the steps that can be taken to make available to translators of NIMLs some of the kinds of resources that translators working in the "superleague" languages are starting to take for granted.

Word-processing, Hyphenation and Fonts

Word-processing and font provision is more or less trivial for languages using the Roman alphabet, though in some cases (e.g. Vietnamese) the requirement for unusual diacritics may be a challenge. Hyphenation rules differ hugely from language to language (and even between varieties of the same language), and so must be especially provided for. For non-Roman script languages of course, hyphenation may not be an issue. Chinese is a "superleague" language and so is well provided for in terms of word-processing software. It should be noted however that software that goes beyond provision of character handling but is based on Mandarin may be unsuitable for Cantonese and other languages (often incorrectly identified as Chinese "dialects") using Chinese characters. It should not be forgotten also that high-quality systems for less popular languages are correspondingly more expensive. Hussein Shakir of Newcastle-upon-Tyne City Council told me that there are several quite good DTP packages available for Urdu which provide good quality output, but they are expensive - around £1000 per copy - and have fewer facilities and are harder to use than standard word-processing software. Even more "exotic" languages not listed in Table 2 are usually covered as far as fonts are concerned, and in the worst case the committed translator can get software for developing original fonts.

Language	WP	Hyph	Font	SpCh	StCh	Dic 1	Dic 2	Dic n	Thes	Term	CAT	MT
Albanian	•		•									
Arabic	•		•	•			•	•				•
Bangla	•		•									
Chinese	•		•			•	•			•	•	•
Croatian	•		•				•				•	
Farsi	•		•									
Greek	•	•	•			•	•			•	•	•
Gujerati	•		•									
Hindi	•		•									
Polish	•	•	•					•		•	•	•
Punjabi	•		•									
Serbocroat	•	•	•			•	•					•
Vietnamese	•		•				•					
Urdu	•		•									
Welsh	•	•										

Bosnian, Cantonese, Somali, Sylheti - not listed

TABLE 2
Provision of computational resources for "exotic" languages, as listed in Hearn (1996) and/or World Language Resources (1997).
KEY:
WP - word-processor,
Hyph - hyphenation program,
SpCh - spell-checker,
StCh - style-checker,
Dic 1 - monolingual dictionary,
Dic 2 - bilingual dictionary,
Dic n - multilingual dictionary,
Thes - thesaurus,
Term - terminology.

Spell-checkers, Dictionaries and Thesauri

Modern spell-checkers rely on a word-list (which is not the same as a dictionary, as it simply lists all the words, including their inflections, without distinguishing different word senses), as well as rules - or at least heuristics - for calculating the proposed corrections when a word is not found in the dictionary. Note that for some languages with agglutinative morphology, it is effectively impossible to list all the possible word- forms. These heuristics may be based on the orthographic (and morphological) "rules" of the language concerned, or may take into account the physical layout of the keyboard. Alternatively, they may simply try a large number of permutations of the letters typed in, allowing also for insertions and deletions, and look these up in the word-list.

From the point of view of the computer, the internal representation of a character is more or less independent of the font used to represent it on the screen or printer, which means that all texts are stored internally as strings of character codes. Thus building up a word-list of acceptable strings for a given language can be done independently of the writing system it uses. It is not difficult (only time consuming) to take megabytes of correctly typed Hindi, say, and extract from it and sort into some useful order (e.g. alphabetical order of the character codes) all the "words" that occur in the texts: we just need a working definition of "word boundary" for that language, which may in the simplest case merely involve recognising the "space" character. Such a corpus of text could easily be collected by translators who work on a word-processor.

Assuming that spell-checking algorithms are to some extent independent of the data (i.e. word-lists) that they use, it should not be too difficult to develop customised spell checkers. Indeed, many word-processors permit the user to specify which word-lists or "dictionaries" are to be used, including the user's own, and this can then be extended as it is used, by the normal procedure whereby users are allowed to add new words to their spell-checker's wordlist. It should be noted that "spelling" is in any case an "alphabetocentric" notion almost entirely meaningless for ideographic writing systems like Chinese and Japanese (irrespective of the word-boundary problem already mentioned), and of arguable interpretation for syllabic or semi-syllabic writing systems. In addition, languages differ in the degree of proscription regarding spelling, especially for example in the case of transliterations of loan words or proper names.

Dictionaries, in the normal sense of the word, are much more than word-lists: as well as distinguishing different word senses, they will usually offer some grammatical information. In one sense they are also something less than a word-list, since they usually do not list explicitly all the inflected or derived forms of the words. As Table 2 implies, it is useful to distinguish monolingual, bilingual and multilingual dictionaries. We include here also "thesauri", where we use the term in its non-technical sense of "dictionary of synonyms". Although bilingual dictionaries are listed for many of the languages in Table 2, we should be aware that these are often very small (typically around 40,000 entries) and unsophisticated (just one translation given for each word).


www.logos.it/language_today

Compared to word-lists for spell-checkers, lists of words with associated definitions or organised according to similarity or relatedness of meaning, are a completely different matter. The generation of a list of word forms is only the smallest first step towards developing a dictionary in the sense understood by humans, and it is not at all obvious how to associate word meanings with different word forms automatically. The best one could do would be to create and analyse concordances of the words, which would categorise them according to their immediate contexts, but this again is only a tool in the essentially human process of identifying word meanings and cataloguing them. Of course, for many languages this has been done by lexicographers. Published dictionaries do exist for many of the languages we are interested in, and here there is a small glimmer of hope. Many dictionaries nowadays are computer-typeset: this means that publishers have machine-readable versions of their dictionary, admittedly with type-setting and printing codes indicating lay-out and type-face changes and so on. It is not an impossible task however to develop software that can extract from these the information that is needed for an on-line resource that is useful for translators. There is a major obstacle of intellectual ownership and copyright, but for certain languages, both monolingual and bilingual dictionaries are in some sense available in computer-friendly form, if only the will to utilise them is there.

Unfortunately, this situation does not apply to all the languages we are interested in. For languages of the minority interest, dictionaries are often published only in the country where the language is spoken, where the publication methods are typically more old-fashioned, including traditional lead type-setting or even copying camera-ready type-written pages. To convert these into machine-readable form by scanning them with OCR equipment implies a massive amount of work which is surely impractical.

Another line of attack is to use bilingual corpora: like the (monolingual) corpus mentioned above, a parallel bilingual corpus could be built up by collecting material from translators, though in this case there would be the requirement that the original (source text) material was also in word-processor format. There has been considerable research recently on extracting from such resources lexical, terminological and even syntactic information. Before any information can be extracted from a bilingual corpus, the two texts must first be aligned, i.e. the sentences and paragraphs which are translations of each other must be explicitly linked. Of course this may be more or less trivial, depending on the language pair and the nature of the text. Quite a lot of research has been done recently on this problem. Much of it has concerned aligning corpora of related Western languages, though a number of researchers have also looked at Chinese and Japanese. Fung and McKeown summarise the work done on this task. Of particular interest is work done on Chinese, where translations are rarely very "literal", so that the parallel corpora are quite "noisy". Fung and McKeown have developed a number of approaches to this particular problem. One drawback is that even the best of these methods with the "cleanest" of corpora can only hope to extract much less than 50% of the vocabulary actually present in the particular corpus. With languages that are highly inflected, even this figure may be very optimistic.

On the other hand, an aligned bilingual corpus presents an additional tool for the translator in the form of a Translation Memory. Even if this cannot be actually used by commercially available Translation Memory software, in the sense of searching and pasting entire sentences which match the source text up to an agreed threshold, an aligned bilingual corpus can also be consulted on a word-by-word basis, where the translator wants to get some ideas of how a particular word or phrase has previously been translated. Besides extracting everyday bilingual vocabulary, attention has been focused on identifying and collecting technical vocabulary and terminology.

Style- and Grammar-checking

Style and grammar-checking at its best involves sophisticated computational linguistics software which will spot grammatical infelicities and even permit grammar-sensitive editing (e.g. search-and-replace which also changes grammatical agreement). In practice, "style-checking" tends to be little more than text-based statistics of average sentence length, word repetition, words and phrases marked as inappropriate (too colloquial), and use of certain words in certain positions (e.g. words marked as unsuitable for starting or ending sentences). Good grammar-checkers have only been developed from a very small number of languages: English of course, and French, but probably not even all of our "superleague" languages. Grammar checkers require sophisticated computational linguistic grammars, and although some work has been done on automatically extracting syntactic information from corpora, nothing of a significant scale has been achieved. A similar bottleneck faces development of MT systems proper, which we will discuss further below.

Terminology Management

In technical translation, whatever the field, consistency and accuracy of terminology is very important. Terminological thesauri have been developed for many of the "major" languages in a variety of fields with the aim of standardising terminology, and providing a reference for translators and technical writers. A

characteristic of NIMLs however is that they are often associated with less technologically developed nations, and so both the terminology itself and, it follows, collections of the terminology are simply not available. A similar problem arises from the use of a language in new cultural surroundings. For example, a leaflet explaining residents' rights and obligations with respect to registering to vote or paying local taxes may not necessarily be very "technical" in some sense, but it will involve the translation of terminology relating to local laws which would certainly need to be standardised. If one thinks of the number of agencies involved in this type of translation - every (urban) borough or city council in the country, plus nationwide support agencies - then the danger of translators inventing conflicting terminology is obvious. Again, although terminology management tools exist, there is the problem of "loading" the terminology, which in turn presupposes that the terminology itself exists.

CAT and MT

After an initially disastrous launch in the 1980s, commercially viable CAT and MT software is now a reality: developers are more honest about its capabilities, and users are better informed about its applicability. But Table 2 shows only too clearly that this kind of software is simply not available for most of the languages we are interested in. As mentioned above, sophisticated grammar checkers as well as CAT and MT systems are usually based on some sort of linguistic rule-base, so a linguistic description of the language is necessary. Because the work done on automatically extracting linguistic rules from corpora has not yet produced significant results, a more viable alternative might be to try to develop linguistic resources by adapting existing grammars. This might be particularly plausible where the new language belongs to the same language family as a more established language: a Bosnian grammar, for example, could perhaps be developed on the basis of Russian or Czech.

An alternative to full linguistic analysis is tagging. This term is used to indicate a process whereby words are labelled for syntactic category, but further structural analysis is not attempted. Tagging differs from the traditional parsing of computational linguistics also in the methodology usually adopted: whereas parsing operates according to linguistic rules, tagging is usually on the basis of probabilities with reference to immediate context. A tagged corpus is a useful resource, because it can be used to help linguists write the grammars that are needed for more sophisticated tools like MT. Developing a tagger for a "new" language is usually done by "training" with a corpus: a linguist marks up the tags on a training corpus, and then the software uses this as a model from which to derive its own rules. Researchers have generally reported a fairly clear correlation between the amount of text given as training data and the overall accuracy of the tagger, as might be expected. But this is a plausible route for developing sophisticated LE resources for NIMLs, always assuming that a linguist with the appropriate language background can be found to mark up the initial training corpus.

A final avenue that might be worth exploring is Example-based MT (EBMT). In this approach to MT, the main database is a set of previously translated segment pairs. Translation of a new text proceeds by searching this database for a closely matching example, and then using it as a model for the new translation. As a translator's aid, this approach is known as "Translation memory" of course, but there has been some research on developing EBMT as a fully automatic approach to MT. The main problems in EBMT, assuming of course that an aligned bilingual corpus has been obtained and that its coverage is suitably broad, concern the manipulation of partial matches, for example where the sentence to be translated is a bit like two or more examples in the database, but not exactly like any of them: the question is how to "clone" the new translation from the matched bits, i.e. how do we know how to glue together the fragments? Current thinking in EBMT circles seems to be that a hybrid of EBMT and traditional rule-based MT is appropriate for this case, which brings us back to the problem of developing grammars for our NIMLs.

CONCLUSIONS

Thirty years ago, one of the main reasons given in the infamous ALPAC report for cutting back the development of MT was simply that there was not the demand for translation. Indeed, the report asked whether perhaps there was too much translation going on. It is fairly clear that no such conclusion could be drawn today. But just as there is plenty of work for translators into and out of the major commercial languages of the world, there is an ever-growing need for translation into (more often than out of) NIMLs. This article has discussed the grave lack of computational resources to aid translators working with NIMLs, and has attempted to identify some means by which this lack could be quickly addressed. The road will certainly be a long one, not least because the funding to support research in computational linguistics related to NIMLs will only come from government agencies, unless the private sector sees this as an area where it can make charitable donations. Obviously, at least for the time being, there is no commercial interest in these languages. However, mere difficulty has never been a serious obstacle in basic research and development, and this author, at least, will be making efforts to pursue some of the lines of inquiry suggested here. ■