MACHINE TRANSLATION AND MINORITY LANGUAGES

Harold L. Somers

Department of Language Engineering, UMIST,
PO Box 88, Manchester M60 1QD

Language Engineering (LE) products and resources for the world's "major" languages, including Machine Translation systems, CAT systems, on-line dictionaries, thesauri, and so on, are steadily increasing, but there remains a major gap as regards less widely-used languages. This paper considers the need for LE support for minority languages. The current situation regarding LE resources for the languages in question is reviewed. Some proposals for rectifying this situation are made, including techniques based on adapting existing resources and "knowledge extraction" techniques from machine-readable corpora.

1. INTRODUCTION

While the availability of Language Engineering (LE) products and resources for the world's "major" languages steadily increases, including Machine Translation systems, 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 the "orthographically challenged" of the world's languages, even word-processing software!

Because of accidents of world politics as much as anything else, the world's languages fall into three or four league divisions, reflecting the computational resources available for them. This paper will identify which languages are more or less badly served, 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, both indigenous and non-indigenous.

The paper will also try to make some proposals for what we can do 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 customizing from resources for related languages, the possible use of software localization tools, and the use of "knowledge extraction" techniques from machine-readable corpora.

2. 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 United Kingdom 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 6% 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. Table 1 shows the main language spoken by Indians, Pakistanis and Bangladeshis in England, and Table 2 shows the self-rated level of literacy in English for these groups.[1]

Table 1. Languages spoken by members of "Asian" ethnic groups, percentages broken down by sex and age. *Source*: Health Education Authority [1]

Ethnic			Women			Men		
group	Language	All	16-29	30-49	50-47	16-29	30-49	50-47
Indian	Gujerati	36	28	45	45	18	44	37
	English	32	51	18	8	62	25	19
	Punjabi	24	16	29	33	16	24	29
	Urdu	3	3	3	2	1	4	5
	Hindi	2	0	1	4	1	1	6
Pakistani	Punjabi	48	34	66	84	28	51	57
	English	24	37	3	0	58	15	7
	Urdu	22	23	26	16	10	27	30
Bangladeshi	Bangla	73	72	72	71	68	76	83
	Sylheti	17	14	27	28	7	17	17
	English	10	14	1	0	25	7	0

Table 2. Self-rated English reading ability of members of "Asian" ethnic groups, percentages broken down by sex and age. *Source*: Health Education Authority [2]

			Very	Fairly			Don't
Ethnic group	Subgroup		well	well	A little	None	know
Indian	All		47	14	6	24	9
	Women	16-29	64	12	3	12	9
		30-49	35	17	7	33	8
		50-74	19	4	10	66	1
	Men	16-29	68	10	1	4	17
		30-49	49	19	6	17	9
		50-74	31	16	14	29	10
Pakistani	All		31	21	5	37	6
	Women	16-29	44	21	4	23	8
		30-49	11	13	5	69	3
		50-74	0	2	2	93	0
	Men	16-29	51	24	3	9	13
		30-49	31	35	9	23	2
		50-74	24	21	7	46	2
Bangladeshi	All		24	15	9	48	4
	Women	16-29	39	13	7	36	5
		30-49	2	2	8	85	3
		50-74	0	2	2	96	0
	Men	16-29	53	24	6	10	7
		30-49	14	26	15	40	5
		50-74	6	14	18	62	0

[1] Some languages are differently named in various sources (e.g. Bengali, Bangla), and spelling sometimes varies. In this paper such differences have been standardized in tables and quotes.

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.

Many local councils, particularly in urban areas, recognize this, and maintain language departments to provide translation and interpreting services with in-house staff, as well as lists of free-lance translators.[2] 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.

2.1. 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 in-house translators to cover Urdu (5 people), Cantonese and Bangla (2 each), Punjabi, Hindi (1 each) and Gujurati (0.5), plus Somali (1), Vietnamese, Arabic and Bosnian (0.5 each) for the needs of refugees [3]. 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 [4]. 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 3 shows the totals for the top 20 languages.[3]

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 range into which translators typically translate for the business community.

2.2. 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

[2] Rhoderick Chalmers, Language Service Manager at the London Borough of Camden suggests that, because of the way this activity is funded, in-house translation staff are likely to be replaced by free-lancers in the future.

[3] I am grateful to Rhoderick Chalmers for providing these figures.

Table 3. Translating and interpreting jobs by the London Borough of Camden Language Services, January to September 1997: top 20 languages. *Source*: Camden Language Service Statistics.

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

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.

Word-processing software is generally available for most of the world's languages, at least as far as provision of fonts for the writing system, allowing texts to be composed on a word-processor and printed, rather than hand-written. As we shall see, many of the other computational features associated with word-processing, that English users are accustomed to, are simply not available for NIMLs. For the kinds of repetitive translations mentioned in the last paragraph, for example, "Translation memory" software would clearly be a great advantage.

3. COMPUTATIONAL RESOURCES FOR "EXOTIC" LANGUAGES

Language-relevant computational resources are certainly on the increase. The US-based magazine *Multilingual Communications & Technology* regularly lists new products and advances in existing products, and the software resources guide that it periodically includes grows bigger each issue. The translators' magazine *Language International*, recently taken over by Benjamins, has a similar "Language Technology" section. 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, German, Spanish, Italian, Russian) plus Japanese, Chinese (i.e. Mandarin), Korean and, to a certain extent, Arabic. Their concern 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.[4]

The situation regarding provision of language technology for minority languages mirrors the equally dismal picture in language *planning* in the UK discussed by Stubbs (1991) [5]: he reports that "In England some 5 per cent of children are bilingual, but in many schools over 60 per cent of the children speak the same language other than English, and in some schools it is over 90 per cent" [6]. Furthermore, he stresses, "such children are not from migrant worker or immigrant families. . . . They are second or third generation British citizens." Stubbs draws attention to the "overt rhetoric . . . of ethnic diversity and multiculturalism . . . always held in check by ethnocentric and assimilationist assumptions" [7]. Young bilingual children quickly learn that their language skills are deprecated if their "other" language is not European. Amma Ntiriwaa, a 13-year old Ghanaian, is quoted as follows in a recent newspaper article: [8]

> I came to England when I was eight and in my primary school there were a couple of students who were French bilinguals. The teacher made a great fuss of this, but when I tried to speak my language, Twi, to my teacher, she said "What a funny language!" I soon realised that my language was not as important as French . . .

A recently published directory of LE resources [9] lists over 1200 software products, and includes a useful index on a language-by-language basis. Table 4 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 4 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 *World Language Resources*, they are only listed under fonts and word-processors.

Let us consider in a little more detail each of the categories listed in Table 4. In the next section, we will return to each of these categories and consider how we could go about providing the missing resources.

3.1. Word-processing, Hyphenation and Fonts

As mentioned above, 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

[4] The situation is slightly different in the US and Canada, where these superleague languages are also important community languages; but there are still many other languages — an even wider range in North America — spoken by large numbers of immigrants which are not of commercial interest.

Table 4. Provision of computational resources for "exotic" languages, as listed in Hearn (1996) [10] and/or *World Language Resources* (1997) [11].

Language	Word-Processor	Hyphenation	Fonts	Spell-checker	Style-checker	Dictionary (mono)	Dictionary (biling.)	Dictionary (multiling.)	Thesaurus	Terminology	CAT	MT
Albanian	•		•									
Arabic	•		•	•			•	•				•
Bangla	•		•									
Bosnian	not listed – see Croatian, Serbocroat											
Cantonese	not listed separately – see Chinese											
Chinese	•		•			•	•			•	•	•
Croatian	•		•				•				•	
Farsi	•		•									
Greek	•	•	•			•	•			•	•	•
Gujerati	•		•									
Hindi	•		•									
Polish	•	•	•					•		•	•	•
Punjabi	•		•									
Serbocroat	•	•	•			•	•					•
Somali	not listed											
Sylheti	not listed											
Vietnamese	•		•				•					
Urdu	•		•									
Welsh	•	•										

must be especially provided for. For non-Roman script languages of course, hyphenation may not be an issue. The equivalent, for Arabic-script languages, is the provision of variant letter forms.

Chinese is a "first division" 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.

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 less facilities and are harder to use than standard word-processing software.

3.2. Spell-checking, 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.

As just mentioned, dictionaries 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 4 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 4, we should be aware that these are often very small (typically around 40k entries) and unsophisticated (just one translation given for each word).

3.3. 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 practise, "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).

3.4. 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 standardizing 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 standardized. 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.

3.5. 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 4 shows only too clearly that this kind of software is simply not available for most of the languages we are interested in.

4. DEVELOPING NEW LANGUAGE ENGINEERING 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 "first division" languages are starting to take for granted.

4.1. Word-processing, Hyphenation and Fonts

At this low level, as we have seen, provision is not too bad. Arabic word-processing packages can generally accommodate the different letter forms that printing requires (e.g. for justified text, letters are stretched so as to avoid hyphenation), even for Urdu which has a number of extra letters customized from the Devanagari writing system used for Hindi — essentially the same language, though spoken by a different political and religious group — to cover Urdu sounds not found in Arabic. Even more "exotic" languages not listed in Table 4 are usually covered as far as fonts are concerned, and in the worst case the committed translator can get software for developing original fonts.

4.2. Extracting Monolingual Word-lists from Existing Texts

From the point of view of the computer, fonts are simply surface representations of internal strings of character codes, so building up a dictionary 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. 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 customized 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 word-list.

As mentioned above, spell-checkers rely on a word-list plus language-specific heuristics. "Spelling" is in any case an alphabetocentric notion almost entirely meaningless for ideographic writing systems like Chinese and Japanese, 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.

4.3. Dictionaries and Thesauri

Monolingual dictionaries, i.e. word-lists with associated definitions, or *thesauri* in the sense of lists of words organized according to similarity or relatedness of meaning, are a completely different matter. While the procedure described above could be used to generate a list of "attested word forms", it is only the smallest first step towards developing a dictionary in the sense understood by humans. It is not 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 categorize 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. Of course 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 utilize 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.

On the positive side, a search of the World Wide Web reveals a number of sites of possible interest. For example, an English–Urdu dictionary [12], albeit in transcription, is being developed by Waseem Siddiqi, a student at KTH, Stockholm, and other similar projects appear to be in progress, at various American universities. However there is no indication of how long these resources will be maintained, or even remain available, and in any case they tend to be several orders of magnitude too small.[5]

4.4. Use of 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

[5] Siddiqi's dictionary downloads as 278k bytes, and, to take a randomly chosen example, there are just 87 words listed under 'N'.

languages, though a number of researchers have also looked at Chinese and Japanese. Fung and McKeown (1997) summarize the work done on this task [13]. 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 (cf. [14]).

Besides extracting everyday bilingual vocabulary, attention has been focussed on identifying and collected technical vocabulary, *terminology*. Fung and McKeown describe how technical terms are extracted from their English–Chinese bilingual corpus [15]. Dagan and Church (1997) describe a semi-automatic tool for constructing bilingual glossaries [16]. Fung et al. (1996) show how the linguistic properties of certain languages can make this task more straightforward [17].

4.5. Developing Linguistic Descriptions

For most other purposes, a fuller linguistic description of the language is necessary. Sophisticated grammar checkers, and certainly CAT or MT tools, are usually based on some sort of linguistic rule-base. Although some work has been done on automatically extracting linguistic rules from corpora, nothing of a significant scale has been achieved. 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. Another difference is that the set of "tags", or syntactic categories, recognised by a tagger is much more fine-grained than those used by a parser. For example a tagger might distinguish singular and plural nouns, transitive and intransitive verbs, predicative and attributive adjectives, and so on.

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. For more on tagging see Barnbrook (1996) [18].

4.6. Example-based MT

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 (e.g. Somers and Jones 1991 [19]; Collins et al. 1996 [20]). 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.

5. CONCLUSIONS

Thirty years ago, one of the main reasons given in the infamous ALPAC report [21] 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 (see Hutchins, 1996 [22]). 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 paper 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 enquiry suggested here.

REFERENCES

1. Rudat, Kai, 1994, *Black and Minority Ethnic Groups in England*, London, Health Education Authority, p. 27.
2. *Ibid.*, p. 31.
3. Figures supplied by M.Y. Nizami, Translation and Interpretation Service, Manchester City Council.
4. *Words: A Newsletter from the Translation & Interpretation Service*, Manchester City Council, Autumn Issue 1996; and personal communication.
5. Stubbs, Michael, 1991, "Educational language planning in England and Wales: Multicultural rhetoric and assimilationist assumptions", in Florian Coulmas (ed.) *A Language Policy for the European Community: Prospects and Quandaries*, Berlin, Mouton de Gruyter, 215–239.
6. *Ibid.*, p. 229.
7. *Ibid.*, p. 231.
8. Upton, Debbie, 1997, "New power of Babel", *Guardian Education*, October 7th, 1997, p. 6.
9. Hearn, Paul M., 1996, *The Language Engineering Directory*, Madrid, Language & Technology.
10. *Ibid.*, pp. 301–329.
11. *World Language Resources: International Software Buyers Guide*, Vol. 5 (1997), Los Angeles, California. Advertising supplement issued with *Multilingual Communications & Technology* magazine.
12. The URL is `http://www.student.nada.kth.se/~d94-fsi/dictionary/`.
13. Fung, Pascale and Kathleen McKeown, 1997, "A Technical Word- and Term-Translation Aid Using Noisy Parallel Corpora across Language Groups", *Machine Translation* 12: 53–87.
14. Isabelle, Pierre and Susan Warwick-Armstrong, 1993, "Les corpus bilingues: une nouvelle ressource pour le traducteur", in Pierrette Bouillon and André Clas (eds) *La Traductique: Études et recherche de traduction par ordinateur*, Montréal, Les Presses de l'Université de Montréal, 288–306.
15. *op. cit.*, pp. 79ff.
16. Dagan, Ido and Ken Church, 1997, "*Termight*: Coordinating Humans and Machines in Bilingual Terminology Acquisition", *Machine Translation* 12: 89–107.
17. Fung, Pascale, Min-yen Kan and Yurie Horita, 1996, "Extracting Japanese Domain and Technical Terms is Relatively Easy", *NeMLaP-2: Proceedings of the Second International Conference on New Methods in Language Processing*, Bilkent University, Ankara, Turkey, 148–159.

18. Barnbrook, Geoff, 1996, *Language and Computers: A Practical Introduction to the Computer Analysis of Language*, Edinburgh, Edinburgh University Press. See particularly Chapter 6.
19. Somers, Harold and Danny Jones, 1992, "Machine Translation Seen as Interactive Multilingual Text Generation", *Translating and the Computer 13: A Marriage of Convenience?* London, Aslib, 153–165.
20. Collins, Bróna, Pádraig Cunningham and Tony Veale, 1996, "An Example-Based Approach to Machine Translation", *Expanding MT Horizons: Proceedings of the Second Conference of the Association for Machine Translation in the Americas*, Montreal, Quebec, Canada, 1–13.
21. ALPAC, 1966, *Language and Machines: Computers in Translation and Linguistics*, Washington DC, Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Research Council. National Academy of Sciences.
22. Hutchins, John, 1996, "ALPAC: The (In)famous Report", *MT News International* 14: 9–12.