Sources of Linguistic Knowledge for Minority Languages

Harold L. Somers
Department of Language Engineering
UMIST, Manchester, England

harold@ccl.umist.ac.uk

Abstract

Language Engineering (LE) products and resources for the world's "major" languages are steadily increasing, but there remains a major gap as regards less widely-used languages. This paper considers the current situation regarding LE resources for some of the languages in question, and some proposals for rectifying this situation are made, including techniques based on adapting existing resources and "knowledge extraction" techniques from machine-readable corpora.

Key words: minority languages, corpora, knowledge extraction

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.[1] Missing are not only these kinds of products, but even simple tools like spelling- and grammar-checkers. Because of accidents of world politics as much as anything else, the world's languages fall into three or four ranks, reflecting the computational resources available for them. This paper will identify some languages which are more or less badly served, and will 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 suggest some techniques that can make the task more manageable, or more feasible.

[1] An earlier version of this paper appeared in the proceedings of *Translating and the Computer 19* (Aslib).

2. Minority Languages

The notion of "minority language" is relative, depending on the geographical stand-point of the observer. We can define the term from an LE perspective (see below), or else from a local point of view. This latter option is relevant, since our proposed solution to the problem of linguistic knowledge acquisition relies on there being a community of professional linguists servicing the minority language speaking community.

The UK is nominally an English-speaking country, with small regions where the indigenous Celtic languages are more or less widely spoken. However, a more realistic linguistic profile of the UK must take into account that there are significant groups of people speaking *non-indigenous minority languages* (NIMLs). 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. In other countries, the picture will be different, but only in the details.

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 (Rudat, 1994). 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 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 major cities, asylum seekers, bring with them language needs that are being addressed by local government agencies.

Just like translations in the private sector, "public service" translations come in all shapes and sizes. 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.

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 users of the world's major languages are accustomed to, are simply not available for NIMLs, nor is there much evidence — e.g. at the recent AMTA Panel Session (Gerber, 1998) — that the major providers of LE software will turn their attention towards NIMLs.

3. 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* has a similar "Language Technology" section. But just a glance at these publications reveals an overwhelming concentration on the few 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.

A recently published directory of LE resources (Hearn, 1996) lists over 1200 software products, and includes a useful index on a language-by-language basis. Table 1 shows the provision of translation-relevant LE resources for a selection of the NIMLs of significance in the UK, and hence of interest to us. What is immediately noticeable from Table 1 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 (1996) while in *World Language Resources* (1997), they are only listed under fonts and word-processors.

Let us consider in a little more detail each of the categories listed in Table 1. In Section 4, 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 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 rank" 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, and may have less facilities and be harder to use than standard word-processing software.

In fact, at this level, provision is not too bad. Arabic word-processing packages can generally accommodate the different letter forms that printing

Table 1. Provision of computational resources for "exotic" languages

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

Source: Hearn (1996), *World Language Resources* (1997)

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 1 are usually covered as far as fonts are concerned, and in the worst case the committed translator can get software for developing original fonts.

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 1 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 1, we should be aware that these are often very small (typically around ~40,000 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).

3.5 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 UK, 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 1 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 these kinds of language 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 "major" languages are starting to take for granted

4.1 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 no: 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.

4.2 Dictionaries and Thesauri

Monolingual dictionaries, 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 so that machine-readable dictionaries are available, although they may include type-setting and printing codes and so on. Software that can extract from these the information that is needed for an on-line resource that is useful for translators has been widely reported (Boguraev et al., 1987; Farwell et al., 1993; Walker & Amsler, 1986), and Mágan Muñoz (1998) discusses this tactic specifically for a minority language.

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.

4.3 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 (Dagan & Church, 1994; Fung & McKeown, 1997; Gale & Church, 1991; van der Eijk, 1993). Before any information can be extracted from a bilingual corpus, the two texts must first be aligned. Of course this may be more or less trivial, depending on the language pair and the nature of the text. Again, much research has been done recently on this problem, much of it concerning corpora of related Western languages, though a number of researchers have also looked at Chinese and Japanese. Fung & McKeown (1997) summarize 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 & 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, 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 (Isabelle & Warwick-Armstrong, 1993).

Besides extracting everyday bilingual vocabulary, attention has been focussed on identifying and collected technical vocabulary and terminology. Fung & McKeown (1997) describe how technical terms are extracted from their English-Chinese bilingual corpus. Dagan & Church (1994) describe a semi-automatic tool for constructing bilingual glossaries. Fung et al. (1996) show how the linguistic properties of certain languages can make this task more straightforward.

4.4 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 (Brent, 1993), nothing of a significant scale has been achieved without the help of a rule-based parser or an existing tree-bank (Briscoe & Carroll, 1997; Grishman & Sterling, 1992; Manning, 1993). Two proposals directly related to developing MT systems for low-density languages describe software involving sophisticated interaction with a bilingual human expert (Jones & Havrilla, 1998; Nirenburg & Raskin, 1998).

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. 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. Tagging has the advantage of needing only a representative corpus with which to train the tagger. 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.

4.5 Example-based MT

A final avenue that might be worth exploring is Example-based MT (EBMT), in its purest form requiring only a set of aligned previously translated segment pairs (Collins et al., 1996; Somers, 1998). The main problems in EBMT, assuming 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 little 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

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. At least for the time being, there is no commercial interest in these languages. It is to be hoped that at least some of the lines of enquiry suggested here will prove fruitful in the short term.

References

Boguraev, Bran, Ted Briscoe, John Carroll, David Carter & Claire Grover. 1987. The Derivation of a Grammatically Indexed Lexicon from the Longman Dictionary of Contemporary English. *25th Annual Meeting of the Association for Computational Linguistics,* Stanford, California, 193-200.

Brent, Michael R. 1993. From Grammar to Lexicon: Unsupervised Learning of Lexical Syntax. *Computational Linguistics* 19, 243-262.

Briscoe, Ted & John Carroll. 1997. Automatic extraction of Subcategorization from Corpora. *Fifth Conference on Applied Natural Language Processing,* Washington, DC, 356-363.

Collins, Bróna, Pádraig Cunningham & 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, 1-13.

Dagan, Ido & Kenneth Church. 1994. Termight: Identifying and Translating Technical Terminology. *4th Conference on Applied Natural Language Processing,* Stuttgart, Germany, 34—40.

Farwell, David, Laurie Gerber & Eduard Hovy (eds). 1998. *Machine Translation and the Information Soup,* Berlin: Springer.

Farwell, David, Louise Guthrie & Yorick Wilks. 1993. Automatically Creating Lexical Entries for ULTRA, a Multilingual MT System. *Machine Translation* 8, 127-145.

Fung, Pascale, Min-yen Kan & Yurie Horita. 1996. Extracting Japanese Domain and Technical Terms is Relatively Easy. *NeMLaP2: Proceedings of the Second International Conference on New Methods in Language Processing,* Ankara, Turkey, 148-159.

Fung, Pascale & Kathleen McKeown. 1997. A Technical Word- and Term-Translation Aid Using Noisy Parallel Corpora across Language Groups. *Machine Translation* 12, 53-87.

Gale, William A. & Kenneth W. Church. 1991. Identifying Word Correspondences in Parallel Text. *Workshop on Speech and Natural Language,* Asilomar, Calif., 152-157.

Gerber, Laurie. 1998. The Forgotten Minority: Neglected Languages (Panel at the Third Conference of the Association for Machine Translation in the Americas, AMTA '98, Langhorne, PA). In Farwell et al. (1998), p. xii.

Grishman, Ralph & John Sterling. 1992. Acquisition of Selectional Patterns. *Proceedings of the fifteenth* [sic] *International Conference on Computational Linguistics, COLING-92,* Nantes, France, 658-664.

Hearn, Paul M. 1996. *The Language Engineering Directory,* Madrid: Language & Technology.

Isabelle, Pierre & Susan Warwick-Armstrong. 1993. Les corpus bilingues: une nouvelle ressource pour le traducteur. In Pierette Bouillon & André Clas (eds) *La Traductique: Etudes et recherches de traduction par ordinateur,* Montréal: Les Presses de l'Université de Montréal, pp. 288-306.

Jones, Douglas & Rick Havrilla. 1998. Twisted Pair Grammar: Support for Rapid Development of Machine Translation for Low Density Languages. In Farwell et al. (1998), pp. 318-332.

Mágan Muñoz, Fernando. 1998 Towards the Creation of New Galician Language Resources: From a Printed Dictionary to the Galician WordNet. *Proceedings of the Workshop on Language Resources for European Minority Languages, First International Conference on Language Resources and Evaluation (LREC '98),* Granada, Spain.

Manning, Christopher D. 1993. Automatic Acquisition of a Large Subcategorization Dictionary from Corpora. *31st Annual Meeting of the Association for Computational Linguistics,* Columbus, Ohio, 235-242.

Nirenburg, Sergei & Viktor Raskin. 1998. Universal Grammar and Lexis for Quick Ramp-up of MT Systems. *COLING-ACL '98: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics,* Montreal, Quebec, 975-979.

Rudat, Kai. 1994. *Black and Minority Ethnic Groups in England,* London: Health Education Authority.

Somers, Harold. 1998. "New Paradigms" in MT: the state of play now that the dust has settled. *10th European Summer School in Logic, Language and Information, Workshop on Machine Translation,* Saarbrücken, Germany, 22-33.

van der Eijk, Pim. 1993. Automating the Acquisition of Bilingual Terminology. *Sixth Conference of the European Chapter of the Association for Computational Linguistics,* Utrecht, The Netherlands, 113-119.

Walker, D. & R. Amsler. 1986. The Use of Machine-Readable Dictionaries in Sublanguage Analysis. In R. Grishman & R. Kittredge *(eds) Analyzing Language in Restricted Domains,* Hillsdale NJ: Lawrence Erlbaum.

World Language Resources: International Software Buyers Guide, Vol. 5. 1997. Los Angeles, California: Advertising supplement issued with *Multilingual Communications & Technology* magazine.