# MT & MAT and Developing World Vernacular Languages

**Jon D. Riding**

Linguistic Computing at British & Foreign Bible Society
http://lc.bfbs.org.uk

## Abstract

Amongst the world's 6,912 distinct natural languages a tiny handful, about fifty, account for the vast majority of MT & MAT systems. Most of these languages belong to the developed world; examples include: English, Russian, Spanish, Chinese & Arabic. Only one major African *lingua franca,* Swahili, is represented in the list. This focus on a handful of languages is the reality of commerce. The greatest resource is available where the greatest return for investment can be found.

For the remaining 6,872 languages of the world the MT/MAT cupboard is largely bare. This paper argues that the nature of MT/MAT development since the last Cranfield Conference and before has made the focus on commercial languages inevitable. The outcome of much of the work of the last twenty-five years has been excellent progress in the design and construction of knowledge-based systems, typically targeted at restricted linguistic environments such as legislation and technical manuals. This has not encouraged the development of analytical systems which, in the absence of detailed knowledge-bases are the pre-requisite for generative translation machines.

There are, however, grounds for hope. The advent of Translation Memory (TM) systems in more recent years has encouraged the development of systems which place increasing reliance on learning from their context as well as interpreting a supplied data-set. For MT/MAT to make a serious contribution in the developing world it must recognise that the resources do not exist to supply the knowledge-bases required by most current MT systems. MT systems that cannot undertake at least some degree of analysis of the languages they are called to process are of little benefit. A combination of better automatic analysis and heuristics derived from TM-like systems may well represent the best option for MT to serve the wider global community more effectively.

## 1 Introduction

The focus of this paper is upon the progress made in MT and MAT systems in the context of the needs of the developing world. For the last twenty of these years the author has been closely involved with the development and use of MT and MAT techniques in the context of developing world vernacular languages; specifically, the use of these technologies within the Bible translation community. This perspective is both narrower and broader than what we might term commercial or governmental MT/ MAT. It is narrower in the sense that the work is focussed upon a particular corpus of text (although having said that, the Bible is in reality a library of texts covering most literary genres including historical narrative, to allegory, correspondence, philosophy and poetry). Conversely, Bible translation is also a field of work which is linguistically far broader than the contexts for which most MT/MAT systems are designed. It is important to recognise the distinctions between what we might term developed world commercial MT/MAT and the ongoing work to provide MT/MAT systems to aid translators working in the developing world and in particular working with local vernacular languages within that context.

## 1.1 MT/MAT & World Languages – A brief survey

Within the developed world the last twenty five years have seen great progress in MT/ MAT systems. When the author first began working in this field in the late 1980s the number of effective MT/MAT systems was very few. Systems such as Systran/Eurotra within western Europe and the Meteo system in Canada were often cited as examples of the state of the art twenty five years ago. These systems were typical of much of MT/MAT at that time. Both produced good results by limiting the scope of the task by restricting the translation context to particular genres and serving a very limited number of languages. In recent years the growth of the world wide web has provided opportunities for the development of systems which fewer than ten years ago would have been impractical. Pre-eminent amongst these is Google Translate which now offers first-draft translation between pairs of languages for a set of fifty one languages. The list of languages offered by Google Translate as published in October 2009 is:

| | | | | | | |
|---|---|---|---|---|---|---|
| Afrikaans | Croatian | Finnish | Hungarian | Latvian | Polish | Swahili |
| Albanian | Czech | French | Icelandic | Lithuanian | Portuguese | Swedish |
| Arabic | Danish | Galician | Indonesian | Macedonian | Romanian | Thai |
| Belarusian | Dutch | German | Irish | Malay | Russian | Turkish |
| Bulgarian | English | Greek | Italian | Maltese | Serbian | Ukrainian |
| Catalan | Estonian | Hebrew | Japanese | Norwegian | Slovak | Vietnamese |
| Chinese | Filipino | Hindi | Korean | Persian | Slovenian | Welsh |
| | | | | | Spanish | Yiddish |

**Table 1.1.1 – Google Translate Language Set,October 2009**

This is an impressive list. All the principal commercial *lingua franca* are included, together with most European languages and even one or two minority languages. A similar list of languages with at least some measure of support from MT/MAT systems twenty five years ago would probably have included only the nine EU member state languages and available translation would have been limited to particular contexts of document such as legal and commercial texts. Reordering the list on the basis of global regions offers a different perspective:

| Europe | | | Asia | Pacific | Americas | Africa |
|---|---|---|---|---|---|---|
| Afrikaans | French | Norwegian | Chinese | Indonesian | none | Swahili |
| Albanian | Galician | Polish | Arabic | | | |
| Belarusian | German | Portuguese | Filipino | | | |
| Bulgarian | Greek | Romanian | Hebrew | | | |
| Catalan | Hungarian | Russian | Hindi | | | |
| Croatian | Icelandic | Serbian | Japanese | | | |
| Czech | Irish | Slovak | Malay | | | |
| Danish | Italian | Slovenian | Persian | | | |
| Dutch | Korean | Spanish | Thai | | | |
| English | Latvian | Swedish | Vietnamese | | | |
| Estonian | Lithuanian | Turkish | Yiddish | | | |
| Finnish | Macedonian | Ukrainian | | | | |
| | Maltese | Welsh | | | | |

**Table 1.1.2 – Google Translate Language Set by Regions, October 2009**

What is immediately clear from this second table is the preponderance of support for western European languages and the growing support for the major commercial languages of SE Asia and the Far East. Of equal significance is the complete absence of any language indigenous to the Americas and the presence of just one African language – Swahili.[1] Placing this result in the context of a global perspective is equally informative. The Ethnologue (Gordon 2005) lists 6,912 distinct natural languages in the world as a whole. If we examine the data for each of the five regions and calculate the percentage of languages in each region served by Google the results are:

| | **Europe** | **Asia** | **Pacific** | **Americas** | **Africa** |
|---|---|---|---|---|---|
| Languages: | 239 | 2269 | 1310 | 1002 | 2092 |
| Google: | 51 | 11 | 1 | 0 | 1 |
| Coverage: | 21.30% | 0.50% | 0.10% | 0.00% | 0.05% |

**Table 1.1.3 – Percentage of Languages covered by Google in each Region**

Globally, the impact of MT/MAT as a whole has benefited the speakers of about three-quarters of one per cent of the global set of living natural languages. Nevertheless the impact has been great. Of the 6,912 living languages approximately 85 account for 4.5 billion speakers. The remaining 1.5 billion speakers account for the other 6,827 living languages. MT/MAT has on this measure made great progress in providing support for the first language of the vast majority of speakers.

## 2 First-language and Bible translation

An individual's mother-tongue is a crucial component in their individual and collective identity. The very concept speaks to the core of learned behaviour. Whatever the relative strengths of nature and nurture, it cannot be denied that a child's formative years contribute in very great measure to an individual's world-view, place in local society and individual identity. Much if not all of this context is learned through the lens of the child's first-language. The same references presented in a different linguistic contexts can lead to very different understandings.[2] The importance of first-language is therefore difficult to over-emphasise. Such issues become particularly acute in the context of Bible translation.

The business of Bible translation differs significantly from that of commercial translation. In the context of commercial translation the emphasis tends to be upon clear renderings of largely factual information from one language to another. The Bible translator faces a different problem. Whilst there are still realities to be expressed much of the text deals with abstracts and metaphysics. Not only is the subject matter rather less well-defined, the different styles of writing found in the Bible include both prose and poetry with all that that implies for the translation task. With increasing recognition of the cultural dependencies in translating biblical text the last fifty years has seen the establishment of the principle of dynamic equivalence in translation. Dynamic equivalence (Nida and Taber 1974) seeks not to render the text as far as possible word for word into a new language but instead tries to find ways to express concepts in words and forms which are generally equivalent to the intentions of the original text. The likelihood of lower literacy levels amongst speakers of many vernacular languages and a growing preference amongst many translators to avoid technical vocabulary and external pre-conceptions. These difficulties, however, pale into insignificance beside

[1] Afrikaans is closely related to Dutch and therefore listed as a member of the Indo-European group.

[2] The use of the word 'crusade' by some in the western world has conjured diametrically opposed perceptions of a common history shared by the Christian West and the Islamic Middle East.

the fundamental problem facing nearly all Bible translation teams. Put simply, their target languages will almost invariably fall within the set of 6,827 languages not presently served by mainstream MT and MAT systems.

Many Bible translation projects, even at the start of the third millennium, have to begin not with the business of translating text but with the need to develop and define an orthography for the language. If an orthography already exists it is unlikely that there will be much pre-existing literature in the language. Much of the business of defining and standardising the writing system is likely to fall to the translation team. Only when this has been done can the team move on to the business of translation.

## 2.1 Benefits from the mainstream

The very limited number of languages for which MT/MAT support is available is clearly a disappointment from the Bible translation perspective but there are nevertheless many benefits albeit indirectly, from the growth in commercial MT/MAT in recent more years. Even in 1990 DTP systems had become sufficiently competent to handle very large texts and together with affordable laser printers made it possible to typeset vernacular Bibles at much lower cost. Non-standard character sets could now be built and systems developed to apply consistent changes across a dataset. These Consistent Change (CC) Tables were an early forerunner of today's Regular Expression engines. Just as with Regex it soon became clear that anyone could make a mistake but if you really wanted to destroy a text you needed a CC Table.

Without question the single most helpful development from mainstream IT has been the Unicode project (Unicode Consortium 2006). Prior to Unicode, not only did a translation team have to define its own orthography, it was not uncommon for the team to have to build fonts to display and print text. Whilst it is still the case that occasionally particular characters have to be created for some languages, the vast majority of languages now have definitions within Unicode at least to some degree. Generally, characters and glyphs are defined and, provided the translators have access to a font which includes the code points they need, there is no need to define characters. Other issues such as marks of various kinds (typically punctuation) and collation sequences may not be covered so comprehensively but the current situation is already much improved from twenty five years ago.

Alongside the Unicode standard the growth in affordable computing power has made it possible to deliver compute intensive processing to the field. Much of what has become common place in recent years was already under development more than twenty years ago. The limitations imposed by the speed of affordable processors has gradually disappeared during this time. Processing which in the late 1980s took hours or even days of mainframe time now runs in seconds on a modern PC.

# 3 How MT/MAT has changed Bible translation

To appreciate how much use is now made of MT/MAT in Bible translation projects we need do no more than review its use in a typical developing world translation project in the late 1980s in comparison with an active project today. Prior to 1990 very few translation projects in vernacular languages had any kind of IT support. Only regional centres had computers of any kind. Most translators worked with pen and paper and their text was typed for them by support staff. Less mainstream orthographies were handled by using composite character strings to represent letters that were not part of standard ASCII.[3] Although attempts were made to standardise these codes, in practise most teams adapted the standards to their own needs. Simple structure markers were being developed to indicate paragraphs, headings and suchlike but these too were used inconsistently between and often within projects. Some IT based support was available to check completed work to ensure that structure markers had been applied consistently throughout the text but that was as far as it went. Given the plethora of local standards it wasn't long before the text structure codes, Standard Format Markers (SFM), became generally known as Somebody's Format Markers. Working with a text from the field twenty years ago required certainly weeks, sometimes months of man hours checking for coding errors in both structure markers and characters. These problems added greatly to the time taken preparing a text for publication.

By comparison, a similar project today benefits from global encoding standards and a purpose built translation editing suite. The Paratext translation editor (UBS 2006) takes a text right through from creation, review and checking directly into typesetting. Throughout this process MT/MAT processing is offering increasing assistance to the translators. Whereas 25 years ago it was not unusual for projects to take 25 years to complete a Bible translation that time is coming down nearer to 15 years with the help of the new systems.

The problem of working with texts encoded using disparate standards was finally resolved with the creation of the Unified Standard Format Marker System (USFM). The need for this was driven by the advent of the Paratext translation editor which provided a common platform for all Bible translation projects. Paratext enforced common standards and made practical the concept of field deliverable systems able to analyse the text as the work progressed.

As these text preparation tools were coming into use work was also beginning on MT/MAT systems which could assist the translator. A Machine Assisted Translation team was formed at the British & Foreign Bible Society here in the UK. Tasked with developing MAT systems to aid translators. The team soon recognised that the particular context of Bible translation offered both benefits and particular difficulties. As a corpus for MT/MAT research the Bible has much to commend it. It is large and, like other well known parallel corpora such as Canadian Hansard, it is structured into sections and sub-sections which are common to all translations. The effect of this is to limit the scope of text which must be processed when attempting to parse a clause. Given the limitations on affordable computing in 1990 this was a major benefit. Conversely, the Bible translation task is focussed on the needs of vernacular languages. Any systems developed would need to perform well without the benefit of lexica or grammatical tables and work effectively across the widest possible set of natural languages.

## 3.1 MT/MAT developments in Bible translation

Initial research concentrated on developing systems to investigate key terms in a text by automatic glossing (Robinson 1991). A text such as the Bible has many key terms which carry important semantics and must be handled consistently and appropriately throughout the text. In projects which might last up to 25 years or more this is even more important. Even if the translation team remains the same for the duration of the project it is likely that their own understanding of how best to translate such terms will develop over the course of the work. Finding an objective way of analysing how these

[3] For example, the character string /e was often used to represent é.

terms had been translated throughout the text was a high priority. The approach adopted was statistical glossing. The structured nature of biblical text is well-suited to this approach. Given two translations of the Bible the system calculates the probability of any word in a verse of Text A being equivalent to a word in the corresponding verse of Text B. Early versions of the system, limited by available computing power, attempted only single word glosses. This was first expanded to include cognate forms and the latest version now gloss all the available text across the selected language pair, identifying stems and equivalent phrases as well as individual words. For a more detailed discussion of the process see Riding (2008).

Such objective analysis is of great value to Translation Consultants (TCs) supporting translation teams in the field. A TC must often oversee projects working in languages with which he is not always familiar. Enabling a TC to give adequate support a translation team in these circumstances is key to a successful project. Much of the support work involves the TC working with the team to review their text by means of a back translation from the new text. Ideally the back translation is prepared by others but in practice this is not always possible. In these circumstances the team themselves prepare the back translation with the consequence that a significant degree of objectivity is lost.

With the development of glossing technology for the key terms analyser it was clear that there was a real possibility that an automatic interlinear display might be generated. Whilst an interlinear is sometimes less easy to work with than a true back translation the objectivity offered by the analysis proved a siginificant benefit. Using the same statistical technique and with the addition of a morpheme analysis module to improve the stem identification results an interlinear display was constructed to work within the Paratext translation editor which was able to generate an automatic alignment between any pair of languages. The need to work with the widest range of projects required the system to be able to handle any language sufficiently well to generate helpful results. As discussed above, the nature of vernacular languages make it unlikely that any dictionaries of lexica would be available so a pre-requisite was that the system be entirely self-contained and non-language specific. As TCs have begun working with the interlinear and the key term glossing technology they have begun to discover that more often than not it is the failures to gloss or align that provided the most helpful feedback. They are now able to see that a particular piece of text within which they expected to find a key term is, for some reason, failing to gloss the term successfully against the model text. This in turn enables the TC to raise helpful questions with the translation team. Conversely, the interlinear display provides not only the opportunity to check how individual terms had been translated in a particular section of text it also offers the opportunity to review areas where strong one to one alignment might not be expected. The areas of the Bible where there is significant use of metaphor are well documented. As a general rule, one would not expect to find strong one to one relationships between model and target texts in these passages. Where such strong alignment does occur TCs are now able to identify this and raise appropriate questions with the translators.

## 3.2 Hopes for the future

The first versions of the Paratext Interlinear (PI) shipped in 2006. The most recent version was released in June 2009 and in addition to generating interlinear displays of completed translations is also capable of generating first stage translation. The ability to generate first cut translation is a huge step forward. Using the glossing technology together with word shape recognition for pairs of languages which are linguistically closely related it is often possible to generate a first draft of a passage in a few seconds. Whilst it seems unlikely that purely automatic Bible translation will ever be an acceptable reality automatic first drafting is already helping to speed up the process of translation and to provide useful critique of completed portions of text.

A pre-requisite for effective glossing is the ability to identify the components of the target language. The single most difficult problem which must be addressed to achieve this is that of complex morphologies. The glossing technologies within Paratext which underpin key terms analysis and the automatic interlinear display include an automatic morpheme analyser. At present closely linked to the glossing technology, this processing is potentially a stand alone module which could provide automatic word-formation analysis from any available corpus of text. Traditionally MT/MAT has addressed this problem by providing tables which the system uses to identify the components of words. In the context of developing world vernaculars this is rarely an option. The resources do not usually exist to create detailed databases describing the structure of the language. Using technologies developed by Bible Society in the UK (Riding 2007) the glossing technology within Paratext is able to analyse the morphology of a target language automatically and compile lists of stems and morpheme tables. These tables are then used by the glosser to improve the results. Work is in hand to extend its capabilities from purely agglutinative morphologies to non-concatenative word formation systems. It is hoped that this processing will soon enable some measure of intelligent spelling check based on the degree to which a word conforms to the perceived patterns of word formation in the language.

# 4 The developing world challenge to MT/MAT

The problems faced by the Bible translation community have forced them to seek language independent solutions. Systems need to be able to make their own analysis of the target languages. Conversely, for much of the last twenty five years mainstream MT/MAT has preferred the knowledge-based approached. More recent developments however suggest that the two approaches are converging. Translation Memory systems have much in common with the glossing technologies developed by Bible Society. Both rely on a large corpus of text and both are capable of mapping equivalences between two texts. The performance of both can be enhanced by providing tables which describe the structure of of the language. Here perhaps lies the challenge for MT/MAT if it is to serve adequately the remaining 99.9% of world languages. The resources to provide detailed information on the structure and components of a developing world vernacular language are few. In most cases they are non-existent. For modern MT/MAT to work effectively with these languages solutions must be found which can provide coherent analysis of structure and components for themselves. These solutions must be language independent, able to work with the widest possible set of world languages. Many of the fundamental building blocks to enable this have been put in place during the last twenty five years. The next twenty five offer the opportunity to put them together.

# References

Gordon, R. G., ed. (2005), *Ethnologue*, SIL International

Nida, E. A. & Taber, C. R. (1974), 'The Theory and Practice of Translation' in *Helps for Translators*, United Bible Societies

United Bible Societies, (2006), *Paratext Software*, Available on-line at http://paratext.ubs-translations.org/ (accessed 21/10/2009)

Riding, J. D. (2007), *A relational method for the automatic analysis of highly-inflectional agglutinative morphologies*, MPhil thesis, Oxford Brookes University

Riding, J. D. (2008), 'Statistical Glossing – Language Independent Analysis in Bible Translation', *in Translating and the Computer 30*, ASLIB/IMI

Robinson, D. & Robinson, P. (1991), 'Remembrance of things parsed', *The Computer Bulletin* **3**(1), 22-24

Unicode Consortium, (2006), *The Unicode Standard, Version 5.0*, Addison-Wesley Professional