Statistical Glossing - Language Independent Analysis in Bible Translation

J D Riding
Linguistic Computing
at
British & Foreign Bible Society

October 20, 2008

Abstract

Bible translation sets a number of particular challenges for machine translation and analysis. The nature of the work is such that translators are often working with local vernacular languages for which there are little in the way of lexica and other linguistic databases. Commercial text processing and translation systems are rarely able to contribute. Translation Consultants (TCs) charged with advising translation teams often have few computer based aids to assist them in their work. This paper describes the development of the Statistical Glossing Tool (SGT) which ships with the Paratext translation editor. SGT is a language independent method for the analysis of bible translations. It provides an objective assessment which TCs can use to help them identify key issues in a new translation where further work may be needed. SGT requires no information about target languages other than the text itself.

1 The Great Divide

1.1 Commercial Translation

The importance of lingua franca inevitably focuses resources onto a handful of commercially key languages. Commercial languages benefit from huge investment in a number of contexts. Many of them have a long history as written languages with the consequence that generations of study have made available detailed analyses and lexica. These resources are available to system developers and with their help it is possible to construct first draft translation systems. These perform well on individual language pairs by utilising the extensive linguistic knowledge bases. In addition to these databases it is often the case that the particular context in which a system is required to operate encourages the developers to limit the scope of the system. This is typically done by restricting the vocabulary sets and sometimes the grammatical tables from which the

system generates its translations. It is invariably the case that difficult problems become more tractable when their scope is limited in some way and these sensible techniques contribute to the success of many of the current generation of machine translation and translation assistance systems.

The majority of the world's 6,912 languages [2] do not, however, fall into this context. Only a small fraction of natural languages are served by existing systems. The remainder have little or no support even for standard text processing functions let alone credible MT solutions. Least well supported are the thousands of vernacular languages of the developing world. In this environment detailed lexical and grammatical knowledge bases simply don't exist or if they do are unlikely to be in a form which can be easily used by knowledge based systems.

1.2 Developing World Vernaculars

Within the developing world it is rarely the case that the authoring aids taken for granted elsewhere, such as spell-checkers, and even basic text processing of lesser used scripts are available. Non-roman scripts have long been a particular problem although the advent of Unicode [1] has helped to resolve issues of scripting to a large extent. It is not, however, unusual for a bible translation project to be at the forefront of defining a language's orthography and for glyphs to be encoded within the Private Use Area (PUA) to provide symbols which have not been included in the Unicode standard. Nor is it unusual for rendering engines to provide less than a 100% solution for such orthographies. All of this contributes to a situation where translation teams find themselves unable to use standard text processing software and lack the support given by text and word processing packages which is taken for granted in the developed world.

2 Bible Translation

Within the Bible translation community the context for translation is radically different from that of commerce. A typical bible translation project has as its target language a local vernacular for which there is unlikely to be any existing repository of linguistic information. The text to be translated covers the widest range of content and styles and it not, therefore, susceptible to being limited to particular vocabulary sets or syntax. It is also unlikely that there are many trained linguists in the community. It is entirely possible that the translation will be the first major text to be written down and printed in the language. The nature of the text is such that the quality of translation is critical. Unlike much commercial translation which is often geared to legislation, technical documents and news publications a bible translation can expect to have a shelf life measured in decades rather than weeks or months. Accuracy is clearly important as is a translation style which is accessible and natural to the target constituency. The size of the text also creates particular problems for translators. When a project may take twenty five years to complete issues of consistency become

critical. Even if the membership of the translation team does not change over the period it is inevitable that the style of translation will develop over the years. This in turn will introduce inconsistencies in the way particular concepts are handled across the breadth of the text. Traditionally bible translation has been seen as a mission orientated task. In the past it was often led by ex-patriate missionaries who committed their lives to living and working in a community for many years. This is becoming less common as a model. Within the United Bible Societies the typical model is a translation team of mother-tongue speakers of the target language drawn from the community for whom the translation is being prepared. Such teams are inter-confessional, representing each of the major Christian denominations in the community. This approach ensures that the translation will be accessible to the local constituency and will speak clearly to their needs in the context of their culture and language.

2.1 Biblical Text and MT

If one were to set the challenge of finding a text which the majority of people would feel to be least well suited to automatic translation it is likely the bible would feature strongly in most lists. Whilst most of us are relatively relaxed about opening a box of flat-packed furniture and encountering a set of instructions written in the vocabulary of one language but in the style and syntax of another this becomes more and more problematic as the importance of the text to the reader increases. At a purely practical level, the bible contains narrative, poetry and song. Much of the meaning of the text is dependent upon niceties of structure which in their original form are often quite invisible in translation but which the translators must nevertheless strive to reflect in their work. Persuading a bible translation team that any form of automatic system might be able to assist them is not easy. Nevertheless, the bible has some characteristics which make it well suited to some types of automatic analysis. In its favour is its sheer size. Disregarding proper-names a modern English translation is likely to contain 12,500 different surface forms of words. A more typical figure for a developing world vernacular is about 40,000 words and word lists in excess of 70,000 words are not unusual [3, 31]. This represents a vast amount of data from which much useful information can be derived and used to assist translators and TCs in their work.

The various translations of the Bible which have been made represent a huge set of parallel corpora well suited to automatic analysis. Research at BFBS since the early 1990s has focused on ways in which automatic analysis of a developing text might benefit the translator and translation consultant. Early attempts at automatic glossing proved effective in providing an objective analysis of individual glosses across the breadth of a text and in generating lemmata lists from which concordances could be created. Two problems prevented the technology being made generally available to the field: disparate encoding standards made generic solutions difficult to engineer and, more importantly, the computing power required to generate useful results over a large corpus was not available. Both these limitations are now gone.

2.2 Discourse and Verses

If an automatic system is to make progress with biblical text it needs be able to identify structure within the corpus. It is sometimes possible to identify elements of discourse. Unfortunately many such identifications are dependent upon punctuation which may not work identically in other languages. Similar problems exist when text in translation needs to be radically recast for stylistic reasons and there is always the problem of anaphora [5] raising ambiguities where we might never have imagined they exist. All in all, discourse analysis in free text such as biblical poetry and prose is difficult to make progress with. Providentially we do have a way of aligning translations. It is often arbitrary but it is, if not universal, sufficiently well understood to allow us to use it with confidence to reference elements of the text across different languages.

The system of chapters and verses into which books of the bible are divided in the form we have it today is derived from the work of the 16th century French printer Robert Estienne (Stephanus). His editions of the Greek New Testament introduced the system of divisions still in use today. Every bible published uses one of a handful of well-known variants of these divisions. Verse divisions can sometimes appear to fall in odd places but generally they serve the purpose of dividing the text into small sections the content of which is largely common between translations. This provides those working with computers and biblical text with a reliable system of reference to exploit.

3 Helping Translators

In Bible translation, local teams of translators are supported by Translation Consultants (TCs) with specialist knowledge in biblical languages and formal linguistics. It is not unusual for TCs to have little knowledge of a target language they support yet they must still enable their translators to review and critique their work. The need for tools to assist with this task has been clear for many years. In 2007 an automatic glossing system, derived from the early research at BFBS [6], shipped with the translators' principal text processor 'Paratext' [7]. The Statistical Glossing Tool (SGT) [4] gives TCs the opportunity to make objective assessments of the consistency and style of translations. This in turn enables the TCs to focus their teams on areas of the translation that might benefit from review.

Bible translations need continual review throughout the lifetime of a project to ensure that there is consistency in style across the text as a whole. This is the work of the translation consultant (TC) who is tasked with working with the local translation team and providing them with access to in-depth biblical and linguistic scholarship to ensure their translation is of the highest possible standard. The typical translation consultant is a trained linguist and biblical scholar and is based at a local bible society or perhaps regional centre from where he supports many projects, typically about twelve or more. It is possible that he speaks the languages of the projects for which he is responsible but it

is certainly not guaranteed. Most likely, he may have some knowledge of two or three languages but the majority are unknown to him other than in general terms. Nevertheless, he is required to assist the translation teams in their work, advising on particular difficulties inherent in different parts of the text and critiquing the work done. Assessing a translation into a language unknown to the assessor is clearly a challenge. traditionally it is done by back-translating the new text into a language common to the TC and the translators. Ideally the back-translation is made by a different set of translators from those carrying out the work but this is not always possible. The problems inherent in this approach are obvious. It is not easy for a TC to make an objective assessment of a translation using this method. Whilst individual passages can be assessed in this way it remains very difficult to assess the consistency of the work as a whole, particularly as time passes and the amount of completed text grows. Until 2007, however, this was the only method available to the TC.

3.1 Analysis without pre-requisites

Within the Bible translation community the resources do not exist to construct detailed information describing each natural language. The resources needed to construct knowledge base systems are such as to make this approach generally impractical. For a system to be generally useful it must be largely language independent. SGT requires no input from the user other than to indicate the general family of inflection (affixal, prefixal, suffixal etc...) to which the target language belongs. No lexicon is required nor is there any need to supply complex rule sets describing the transformations encountered in surface forms of words.

Many of the difficulties discussed thus far can be dealt with by providing information about how the source and target languages of a translation work. Lexica, grammars and translation memory systems can all be used to build a system which can attempt a first cut translation or perhaps attempt some form of consistency checking. Unfortunately, any knowledge-based approach is dependent on the resources to create the knowledge bases being available. This is simply not the case in the developing world. We cannot assume that any linguistic data will be available. We must hope that the target language has a defined orthography which is encoded with Unicode. What we can rely upon as a project progresses is a growing body of text forming a parallel corpus with the model text from which the translation is made. Moreover, both these copora share a common structure imposed by their common chapter and verse markers.

3.2 Language Independent Processing

It is clear that any system developed to assist translators and TCs must be capable of handling whatever languages the TC requires of it. This tends to rule out any attempt at knowledge-based processing. A solution is required that needs no pre-defined lexica and that returns useful results across the broadest possible set of languages. This ability is essential in the context of bible translation. In recent years a number of significant steps forward have been made. These

include the definition of a universally accepted system for tagging biblical text in place of the disparate set of regional solutions used before; the development of the Paratext translation authoring platform by United Bible Societies, described in more detail below; and the development of the Unicode standard for encoding scripts. These three developments have provided common standards and allowed the development of a common platform, accessible to all translators and TCs. Prior to this the practical difficulties of working with disparate encodings of scripts and structure were such as to make the development of language independent systems almost impossible.

4 The Statistical Glossing Tool (SGT)

The ability to identify equivalent terms across parallel corpora brings great benefits to those tasked with reviewing new translations. An essential part of the manual review of new translations is the investigation of key terms within the text to ensure that they have been rendered accurately and consistently into the target language. Traditionally this has been done by the TC and the translators going through lists of key terms and their locations in the text and reviewing the translation in each of these locations. This method is both time consuming and difficult to apply objectively but until now it has been the only way this could be done. 2007 saw the beta release of the Statistical Glossing Tool (SGT). Originally developed as an aid to building concordances, SGT is a development of research carried out at British & Foreign Bible Society initially by David Robinson and later by the author. SGT uses the common standards for encoding script and text described above and the universal system of verse markers to generate glosses between a pair of texts. Early work in the late 1980s (notably the MALACHI system (Machine Analysis of Lemmata And Closed-corpus Heuristic Indexing)) from which this system is derived demonstrated that this was indeed possible and subsequent developments in automatic morphology analysis, together with the greatly increased power of modern PCs have allowed the development of the current system.

Early models required a very large corpus of text, at least the extent of a New Testament, and performed well identifying glosses between clearly defined semantics. More similar languages tended to give better results and the more complex morphologies typical of many vernaculars contributed to poorer results for these languages. Later systems attempted to address both the problem of complex morphologies and the limited amount of processing available. The Augustus system, released in 1997, was able to handle languages with more complex morphology better but attempts to allow glossing using smaller sections of the corpus failed to give good enough results. This limitation together with the lack of coherent encoding standards prohibited wider release of the system in this form. The development of Paratext, particularly the release of Paratext 5 at last dealt with the various encoding problems and the wider availability of powerful PCs encouraged the subsequent development of SGT.

Throughout these various incarnations the basic principle remained the same.

The user selects a word from the model text, for example *temple*. If lemmata tables exist for the model text these can be used to identify related surface forms such as *temples temple's* and *temples'*. A map of verses in the model where the word can be found is then made. The same verse map is acquired from the target text and the words found in those verses listed. The occurrence of each of those words within the target language verse map *tm* is noted together with their occurrence globally in text as a whole *tg*. We can derive a simple probability of relationship R between our model word and each of the target words thus: $R = \frac{tm}{tg}$. The closer this value approaches 1 the more likely that the word is equivalent. Earlier refinements included synonym handling by taking the best match found and removing those verses from the target map before recalculating the occurrences for the remaining words. For example, an attempt to gloss *bread* may well generate a partial but strong result. On removing those verses from the map an alternative such as *loaf* may well come to the fore and account for most or even all of the remainder.

More complex morphologies were first addressed by analysing not only the words found in the target verse map but sequences of characters within words. In the case of languages such as those of the Bantu group with rich morphologies it is not uncommon to find that a map of fifty verses may contain twenty, thirty or forty difference surface forms of a lemma. Processing slices of words allowed common radices to be identified as well as providing helpful morphological analysis. Ultimately the problem of complex morphology is being addressed by a pre-process which runs once when the target language is first loaded and generates a table of lemmata and surface forms [3]. The glossing process then uses this information to improve the results.

4.1 Paratext (PT) as an editing and processing environment

The Paratext editing environment was conceived originally by Renier de Blois (UBS TC) and developed further by United Bible Societies (UBS) under the leadership of Nathan Miles (UBS Software Development). The principle of Paratext is very simple. The translator is presented with a window containing a set of frames each one of which displays a particular translation. As the user moves their text cursor to a particular verse in one of the frames, all the other frames scroll automatically to the same verse. This allows the translator to type their translation into an empty frame (previous populated with verse tags) and see the corresponding verse displayed in the model text and, if they choose, the original Greek or Hebrew as they type. Paratext also provides lexical data for the base texts via the Source Language Tools sub-system also developed by de Blois. This simple model revolutionised the work of translators. No longer was it necessary for translators and TCs to carry heavy sets of reference books with them. The information they needed was presented automatically on screen as they worked. The same systems also provided a comprehensive set of checks to ensure the structure of the text was maintained.

The general adoption of Unicode across the bible translation community

and the development of Paratext as a standard authoring and encoding system for bible translations provided a platform which handled successfully common text processing issues for the vast majority languages. The importance of the Paratext system to the subsequent development of SGT and related technologies cannot be over emphasised. Prior to Paratext the time required to validate the structure of texts was such that there was little scope for anything more helpful beyond checking that the verse structure was in place and generating word lists. Texts created within Paratext can be relied upon to be structural consistent with their models and to have all the necessary structure tags in place to allow their verse structures to be successfully navigated. Moreover, Paratext provides a set of objects which expose the text to other systems. A related system simply requests a verse or verses of a text from Paratext and Paratext returns the text, with or without structural tags as required.

4.2 SGT and Paratext

As PCs became more powerful the limitations which had prevented processing in the field largely disappeared. Processes which in the 1990s could only realistically be run on dedicated RISC machines at BFBS began to look more practical on PCs in the field. As the possibilities widened UBS asked BFBS if the glossing technology could be made more widely available via Paratext. This development began in 2005 as a collaboration between the author and Clayton Grassick of the Canadian Bible Society. As a member of the Paratext development team Grassick was able to review glossing technology and determine the most effective way that it might be coupled with Paratext. The original specification called for a Key Term Glosser which would determine the best equivalent term for each of the entries on the UBS Key Terms list. As the system developed it became clear that there were wider possibilities including automatically generated interlinear displays. Grassick reviewed the original simple but robust match algorithm and concluded that whilst it provided a good solution for glossing single terms, there were a number of weaknesses when it was used to attempt interlinear alignments. Where there were very few occurrences of a term within a corpus problems arose. Likewise, very frequently occurring terms might generate spurious matches by simple coincidence.

For example, in the case of very frequent words such as *the*, a gloss attempt to Spanish using the original algorithm would indicate a strong relationship R between *the* and the Spanish conjunction y (and) since almost every verse in an English bible contains *the* and the same is true in Spanish for y. The problem was resolved by calculating what the strength of a relationship would be between the terms purely by coincidence R_c. This allowed R to be adjusted by $R = R - R_c$ and the result rescaled to fall between 0 and 1.

The original algorithm can also produce spurious results with infrequent words. For example, if the word *perro* occurred only twice in a Spanish text and in both cases the English text had *dog* in the corresponding verse this method would conclude that $R = 1$ (certainty). This is, of course, entirely possible. It is, however, equally possible that $R = 0.5$ and we were lucky twice. Given the

available information all that can be done with confidence is to plot a curve which gives the likelihood of every particular value of R. This curve proves to be an Incomplete Beta Function with parameters $m + 1$and $s - m + 1$where m is the number of matches and s is the total number of verses where the source word appears. Rather than estimate R directly we try to find the value of r, the number for which we are 95% certain that $R > r$. In the *perro* example, after 2/2 matches, we are 95% certain that $R > 0.368$.

A further improvement was made in calculating the inverse of a gloss. Where a lemmata table is available for the model text it is common to find strong matches for particularly common surface forms in the target text with a particular lemma in the model. For example, the attempt to gloss *were* from English to Spanish gives a strong signal for *era*. The Spanish *era*, however, will map strongly to the English lemma [*to be*]. This is a good gloss but loses much of the information of tense, and person implicit in *era*. We adjust for this by calculating the result of an exponential function $RInv^{\frac{1}{\gamma}}$ and multiplying the match score by this result. Gamma is selectable by the user from values 1-20 using a slider on the interlinear display.

The outcome of these changes to the original algorithm was the ability to generate useful alignments between model and target texts which provided TCs and translators with helpful information on the degree of close correspondence between the two texts in a particular verse. If might be assumed that the closer the correspondence the better but this is not necessarily so. Whilst some passages may well generate closely parallel translations others, particularly those in which more abstract ideas and metaphors are present, are less likely to encourage such close translation. One of the group of TCs who formed the early testing panel commented that the interlinear display had allowed him to review areas of the text known to be strong in the use of metaphor and which, in consequence, present particular challenges for translators. He discovered that, as expected, these passages tended to generate poorer results in the interlinear display with one or two exceptions. The exceptions interested him and on investigation he concluded that in those places the translators might have stayed closer to the original metaphorical imagery in the model text than might be entirely helpful for their readers. This is an interesting demonstration of the value of this sort of processing. Much of the output from SGT, whether in Key Term analysis or interlinear alignment will simply confirm what the translator and TC believe to be true about their text. It is the areas where the results are not as expected that will generally prove worth investigating.

4.3 Concordances

The original work from which SGT is derived was geared towards generating concordances. In a text the size of the bible, concordances and glossaries become important aids for the reader. With SGT in place as part of the Paratext software suite the generation of concordances in the field became a practical proposition. Prior to SGT such projects were measured in years. The team at BFBS have developed a method for the creation of short concordances, typically

between 100 and 300 pages in length, which can be bound with a bible. Usually referred to as 'Back of the Bible' (BoB) concordances these products are an essential guide for those studying the text, both lay and ordained.

The SGT processing has been harnessed within the Concordance Builder (CB) program to enable translation teams to produce concordances to their work quickly and easily. The system uses SGT glossing technology to gloss each head word in a model concordance against the target text, find the best equivalent, list the verses in which a form of that equivalent is found and then subset that verse list against the verses listed in the model under the model headword. Editors are free to reject the suggested glosses but early testing suggests this is not often necessary. Using punctuation and clause boundary information supplied by the editors, CB can automatically select from each verse the portion of the verse containing the key word which best fits the final typography. The system has reduced the time taken to construct a concordance to a new bible translation from years to a matter of a few weeks. As part of the Paratext suite CB is able to take advantage of the Paratext Publishing Assistant which automatically typesets the finished concordance via Adobe InDesign and without the need for on-screen editing.

5 Further Developments

Current weaknesses in SGT are largely in processing languages with particularly complex morphologies. The current pre-process which analyses target language morphology performs well enough to give good results with languages of similar type and complexity of Swahili, i.e. highly inflected languages where word-formation takes place by prefix and/or suffix affixation. The original morphology processor performs less well with extremely highly inflectional languages, particularly where significant changes occur in stem radices as a consequence of morpho-phonological change at stem-morpheme boundaries and languages where a significant degree of infixal change takes place. A much imporved process has been developed which will be ported to SGT as time allows, hopefully during the next eighteen months. Work is in hand at present towards developing a morphology analyser for non-concatenative morphologies. Experiments are also underway to investigate the possibility of working with parallel corpora which do not have the benefit of unambiguous tagging systems such as the biblical chapter and verse divisions.

Lastly, I must record my thanks to the British & Foreign Bible Society whose persistence in supporting this work made possible the development of these systems. Thanks are also due to United Bible Societies for their readiness to trial a highly experimental process and their enthusiasm and support for its subsequent development. Likewise, the support of the UBS Paratext team and the Institute for Computer Assisted Publishing at the Canadian Bible Society was key in the task of porting the original BFBS research into the Paratext platform.

J D Riding
Linguistic Computing at
British & Foreign Bible Society
October 2008

References

[1] Unicode Consortium. *The Unicode Standard, Version 5.0.* Addison-Wesley Professional, 2006.

[2] Raymond G Gordon. *Ethnologue.* SIL International, 15th edition, 2005.

[3] JD Riding. A relational method for the automatic analysis of highly-inflectional agglutinative morphologies. Master's thesis, Oxford Brookes University (MPhil), 2007.

[4] JD Riding. Paratext glossing tool. http://lc.bfbs.org.uk/news.php?item.28.3, February 2008.

[5] DWC Robinson. The problem of anaphora. Technical report, British & Foreign Bible Society, 1985.

[6] DWC Robinson and PJ Robinson. Remembrance of things parsed. *The Computer Bulletin*, 3:22–24, 1991.

[7] United Bible Societies. Paratext software. http://paratext.ubs-translations.org/, March 2006.