Automatic Concordance Creation for Texts in Any Language

N. W. Rees and J. D. Riding

Linguistic Computing at The British & Foreign Bible Society

Abstract

The work of the Bible Societies is to translate and publish scripture in the mother tongue languages of the world. Many translations are prepared for constituencies where the general knowledge of the text may be quite low. Identifying the key pericopes of scripture in an extent of more than 1,000 pages can be challenging. Help for the reader is most commonly given in the form of a concordance which lists the important narratives and other key areas of the text indexed by the key words which most closely represent the themes of the text.

The British & Foreign Bible Society in the UK has developed a system to allow an existing concordance to the Bible to be used as a model for a similar concordance in another language. Built upon the automatic glossing technology developed by BFBS, the system is wholly language independent and automates the vast proportion of the task. Key words are first glossed from the model text to the new target. Sets of quotation references for each word are then compiled. These sets are then subset by reference to the quotation sets selected for the original key words in the model concordance.

1 Definition

Concordances to large works or collections of works are not uncommon. Any large corpus of text which is regularly studied is likely to be a candidate for a concordance. The particular focus of this work is that of the Bible Societies' task to translate, publish and distribute the Bible as widely possible through the world. Biblical concordances have a long history and are probably the most common context in which concordances are to be found. In this context concordances generally take the form of an extensive table of references listed under key words which are likely to be search indexes for a reference. Key words are listed alphabetically, usually in their citation form, and then references for each key word follow, typically in the preferred canonical order of the original text. Each reference is identified by a book name, chapter and verse number followed by a portion of the verse which includes the key word. The key word is often emboldened or italicised for ease of identification. The number of key

words, references and the length of the extracts from the verses are typically determined by the nature and scale of the concordance. This is in turn influenced by the target constituency, the number of pages available in the product and the size of the page itself.

2 Different styles of concordances

Concordances fall into one of two groups distinguished by the nature of the indexing system used to reference the entries. Most traditional concordances, and all semi-exhaustive and exhaustive concordances, are primarily linguistic documents. Increasingly common for biblical concordances is the shorter thematic style of concordance. Commonly between one hundred and two hundred and fifty pages in length these documents are placed at the back of their Bible in the same binding and may also be published separately. Each type has its strengths and weaknesses. A short thematic concordance will clearly be less comprehensive than a near exhaustive concordance listing nearly every word in the text together with its references. Some problems, however, are common to both. Both types require a lemmatised text so that cognate forms can be readily identified, both are likely to require the disambiguation of homonyms and the cross referencing of synonyms and other related terms.

2.1 Type 1: Fully exhaustive and semi-exhaustive concordances

For this type of concordance the indexing is linguistic. The process of compilation begins with the complete lemmatisation of the whole text in order to identify related surface forms and preferred citation forms. Whilst some cross referencing may be applied at a later stage to identify close cognates, the indexing system is primarily linguistic. Exhaustive concordances may well group entries under individual surface forms and then group these sets of references in turn under the general citation form for the lemma. It is not uncommon for semi-exhaustive concordances to the Bible to be years in the making. The first concordance to an English Bible to become generally known was prepared by Alexander Cruden 1699-1770 [1, 5] in the mid-eighteenth century. Cruden spent almost fifty years working on various editions of his concordance during which time he was regularly admitted to institutions in an attempt stabilise his mental health. The compilation of exhaustive concordances is a huge task. Faced with the need to build large semi-exhaustive concordances [10, 4] to both the Good News Bible [2] and Y Beibl Cymraeg Newydd (The New Welsh Bible) [3] the British and Foreign Bible Society began exploring ways to shorten the process by automating the task of lemmata identification.

2.2 Type 2: Shorter thematic concordances

Thematic concordances have many similarities with larger exhaustive concordances but differ in two important ways. First of all, whilst the indexing system

may appear to be based upon key words, closer inspection reveals that in many cases a particular citation form may not only group entries including different surface forms of that lemma, it may also include entries indexed by closely related words which share a broadly similar semantic. Often these related words will be close synonyms. The second way in which thematic concordances differ from exhaustive concordances is the much more restricted set of entries which are given for each key word. Strictly speaking this is nothing to do with the thematic nature of the concordance, but in practice these concordances tend to be much smaller and need to subset their entries very significantly to fit the available space. Having developed systems to assist with the creation of large semi-exhaustive concordances, the team at BFBS began extending and adapting their earlier work to handle shorter thematic concordances. The first concordance produced using these techniques was a short Concordance for the Good News Bible [11] of 118 pages designed to be bound with the Bible in a single block.

Whilst early work focused on identifying the key terms in the text and listing the key references for each of those terms, in time a different perspective was developed. Thematic concordances, unlike larger concordances, are not really linguistic documents. They have a very practical application which is to enable the reader to quickly locate important passages within the text. Whilst they are presented as lists of references indexed by key word the individual key words are in reality secondary to the set of references. A shorter concordance is fundamentally a set of important locations in the text which the majority of readers are likely to want to find. The primary selection of content is not, therefore, linguistic. Only when those key references have been identified can the editor move on to the question of how best to index them. This may seem counter intuitive but it is the basis for a good concordance which will serve the needs of the readers.

This type of shorter concordance proved very popular and not only in the home market in the UK. The Good News translation is widely used throughout the anglophone developing world and the availability of an affordable, albeit short, concordance to the text was welcomed. National Bible Societies from Africa and South America were soon knocking at the door asking if the Good News Concordance could be translated for use with their indigenous texts. This raised a number of important questions. Whereas a semi-exhaustive concordance can be produced in the same way for different languages thematic concordances are not so straightforward. One man's key index may be another man's irrelevance. A single example can demonstrate this. Consider the verse from the gospel of John: "I am the way, the truth, and the life". Let us suppose that only one index were permitted for this verse. At first glance most readers would imagine there are three options, in fact there are four. Biblical exegesis identifies a number of verses in John's gospel as containing *I am* sayings. These verses are important for the reader and so in addition to *way*, *truth* and *life* we have a fourth indexing option, *I am*. But if we must choose only one index the index we select is likely to depend on the perspective of the target constituency. For some the abstracts *I am* and *truth* will be most important, for others *way*

will be most important and others again will prefer *life*. Likewise, problems arose in considering the selection of entries to be included under each key word. In some East African cultures for example the importance of *blood* as a symbol for life itself required more entries to be added to the original English list for *blood*. It soon became clear that simply reproducing the original indexing risked exporting a whole set of cultural dependencies which might have little relevance in a different constituency.

To deal with these issues a methodology evolved which used the power of the automatic lemmatisation system to short cut much of the process but gave opportunity for changes to be made to the generated concordance. This combination of automatic compilation and manual redaction ensures that the concordance can be closely tailored to the needs of the readers.

3 The concordance builder methodology

A pre-requisite for what has become the Concordance Builder [7] system was the ability to utilise work already done creating concordances in other languages. Time is rarely available for an editing team to start from scratch. More often, the editors already have in mind an example concordance. The concept of an example or *model* concordance was established early in the development. The first task is to reproduce, so far as possible, the model concordance indexing and entry selection. This part of the process is entirely automatic and uses the BFBS automatic glossing technology [9] to identify equivalent terms between the model text and the target text. This process will typically recreate 85% the model concordance in a different language in about two minutes, depending on the speed of the processor.

The second stage is for the editors to review and approve the automatically generated entries and to add or remove any references as required by the target constituency. Having completed the indexing process a portion of each entry must now be selected for display in the concordance after the verse reference. Once again, this is largely automatic. Those entries where a portion cannot be automatically selected, typically a small minority, are then reviewed and a suitable section selected manually by the editors. Once this work is complete all that remains is for the editors to supply a suitable preface if required. The concordance pages are then created automatically as a press ready PDF file or as XML.

3.1 Automatic index generation

The process of reproducing a model concordance in a new language requires the original indexing to be translated for the new text. At first sight this looks straightforward but it can in practice become quite complex. Each head word in the model concordance must be analysed to identify the lemma and its various surface forms present in the model text. In English the list of forms will be short. In more highly inflected languages the list of forms can run into hundreds. The

original text for which the model concordance was created is then taken together with the new text for which the new concordance is to be prepared. Using the glossing technology the best equivalence is identified within the target text for each head word in the model concordance.

The BFBS glossing technology is fundamentally a statistical process. Once a pair of texts have been selected the two text streams are aligned. Clause alignment can often present major difficulties but the existence of a common schema for the Bible which divides the text into short sections provides a useful structure. Each section, or verse, rarely contains more than twenty words. Different canonical traditions can cause difficulties, but these differences are now recognised and well understood [6] and solutions are in place to ensure that the corresponding portions of text align correctly. In its simplest form the glossing program creates a map of locations in the model text where a particular lemma is found, The system then examines the same locations in the target text for terms that occur in these places. Scores for each term found are adjusted for global occurrence throughout the text and the best score at each location chosen as the most likely match.

The single greatest problem faced by the glossing technology, after the issue of clause alignment has been overcome, is that of complex morphology. Many languages, a disproportionate percentage of developing world vernaculars in particular, exhibit much higher inflection rates than most Western European languages. This can make the identification of cognate forms of a lemma much more difficult. The glossing technology which drives the concordance creation includes an automatic morpheme analysis process [8]. This process first makes an analysis of the target language morphology, identifying stem lemmata and common morpheme structures. Whilst the results from this process are unlikely to be complete the benefit of even a 95% analysis of word formation is very significant and pays dividends when the glossing process runs.

Sometimes, particularly if the two languages are reasonably close, a single target lemma will be found that maps closely to the original. Languages which are more distant may generate results in which two or more target lemmata map to the original lemma from the model. The automatic glossing technology can handle these complexities. Once target lemmata have been identified the various surface forms present in the text can be listed and the key words in each reference identified. In cases where a model headword has glossed to synonyms or closely related lemmata in the target, the editor must now decide how best to catalogue the entries for each lemma. It may be appropriate to list all together with a multiple head word but it is more likely that each lemma will be listed independently with appropriate cross references to the other related head words. The final stage of head word creation is to review the head words and select the most helpful citation form for each one.

3.2 Reference list creation

The model concordance has a set of references under each head word where instances of forms of that head word are found in the text. In semi-exhaustive

concordances these lists may be comprehensive, giving every place in the model text where a form of the lemma is found. For smaller concordances the list will almost invariably be a subset of these instances. As discussed above, the critical issue is not 'what head words should be included' but 'what are the key references that should be present'? The automatic indexing process will generate indexes for the majority of the reference list, typically between 80% - 90% of the whole. These indexes will be broadly equivalent to those in the model concordance. The model concordance supplies not only a the set of model key words, but also the set of model references which correspond to the references in the text most likely to assist the reader. This key reference list is linked by the model with the most helpful key words to index the references. Thus a reference set for a particular key word will contain only a subset of the references in the text to that particular lemma, but those present should be the key references for the user.

Taking the example of a model lemma that maps closely to a lemma in the target text the same reference list is created from the target text and where a form of the target lemma appears in a reference it is automatically indexed under that key word. All of this work is handled automatically and generally takes less than five minutes to run. The editor is left with the task of indexing the remaining verses which the automatic process has failed to handle. Typically they will amount to about fifteen percent of the whole. The editor is provided with a simple display which lists the currently unindexed verses. Each verse must be reviewed and an appropriate index word selected. In some cases the word selected will be a form of a lemma which is already present as a head word. In these circumstances the reference is added to the list for that head word unless the editor directs otherwise. Where the index word selected does not already exist in the head word list a new head word is created and the verse listed beneath that head word. Occasionally editors may feel that a particular reference is not important for their particular constituency and delete it from the concordance. Conversely, particular cultural contexts may require more emphasis on certain concepts and editors may wish to add additional references to ensure adequate coverage.

By the end of the indexing process each reference is indexed against one or more head words and the structure of the concordance is complete. The remaining tasks are geared towards presentation.

3.3 Selecting portions of references for display

Having a set of indexed references for a concordance is in fact only half the battle. Once the indexing is complete each reference must be examined and a portion of the text selected which includes the key word and gives sufficient context for the reader to be able to identify the pericope. The amount of text which can be selected is limited by the typography of the concordance. Most small concordances are set in two columns. Since the vast majority are bound with the text the reference the page size is identical. A typical Bible page measures approximately 130mm by 200mm. Allowing space for margins and

gutters the length left for each one line reference is about 40mm. At 6pt with $\frac{1}{4}$pt or $\frac{1}{2}$pt leading this works out in a language like English at something like seven or eight words in a line. Finding a suitable portion of text that is clear and includes the keyword can be a challenge. In the past some publishers have used ellipses to illustrate lacunae in a selection in an attempt to provide the most helpful portion of text within the available space. Another common device is to truncate the key word to the initial letter. The Bible Societies have not found either of these devices to be particularly helpful. Much of our work is in the context of the developing world where literacy levels are sometimes low. Removing part of a reference has been found to cause at least as many problems for the reader as it solves for the editor. Abbreviated head words are a very European concept. European languages tend to inflect their words primarily by suffixes. Many other languages prefer prefixes. In these cases the key words in most of the references will tend to abbreviate to one or a handful of initial letters which represent the common prefix morphemes. The alternative is to use the initial letter of the lemma or citation form but this again causes problems where literacy levels are lower.

BFBS have developed an automatic system to select portions of references. The system uses punctuation (if present) and words and phrases likely to mark clause boundaries. These words and short phrases are identified by the system aitomatically by an analysis of words which occur at known phrase boundaries unambiguously marked by punctuation. Given these parameters much of the task is automated. All automatic selections may be reviewed by the editors reviewed by the editors and, if necessary, the portion selected can be adjusted manually. Automatic selections are ranked for confidence and presented for review ordered according to confidence rankings to allow the editors to focus on potential problems. Once the selection review is complete the concordance is ready for typesetting.

3.4 Handling proper names

Given that the object of a small concordance is to identify references which guide the reader to key pericopes in the text, proper-names have an important rôle in this. Most biblical narrative is comprised of stories about people and places. Often a reader may remember a narrative by the name of the protagonist or another key character in the who appears in the story. A concordance needs, therefore, to provide a comprehensive list of proper names and the references within the text where they can be found. Sub-setting the reference list for a proper name is difficult. As an alternative the BFBS systems creates a list of all the names which appear more than a particular number of times in the text, usually three times. The list is generated using the same glossing technology as for the main concordance. All references to each name are retained but conflated to indicate pericopes within which a character is mentioned many times. For example, the character Aaron appears in the Good News Bible in 327 verses. The first 51 of these references fall within chapters 4 to 12 of the book of Exodus. These references are conflated into a single line in the names index. References

to the same name in adjacent verses are linked thus '2 Chron 3:9-10 x4', those in adjacent chapters to 'Ex 4-12 x51'. These 'bridge' references identify the key passages where references to that character or place are found. No portion of text is displayed.

3.5 Output to press

Once the indexing and portion selection is complete, all that remains is to typeset the concordance ready for press. Concordances are fundamentally tabular data. As such they are well suited to programmatic typesetting. After many years of compiling concordances the Bible Societies have identified a set of typography which works well for these products. Within this general framework there are many options for page layout depending on the size of the page and other issues such as average word length in the language. As part of the Concordance Builder system the user is provided with a simple interface which allows him to select the typography options preferred. PDF output is then created automatically by exporting the concordance data directly into Adobe InDesign where the typography required is automatically applied and the pages created.

4 Outcomes in the field

Within the Bible translation community this technology has revolutionised the presentation of new translations. The Bible is a large text, typically of up to 1,500 pages or more. In communities where literacy may be low it can be a difficult text for readers to fully engage with. The benefits offered by helps such as glossaries and concordances can be very significant. Prior to the development of the Concordance Builder, biblical concordances were rarely available outside the western world. Such as were available took many years to compile and were often prohibitively expensive for most readers. Using Concordance Builder, the time needed to create a short concordance suitable for publishing in the back of a Bible has fallen from years to weeks. With the reduction in time taken has come a similar reduction in cost. First editions of new translations can now be printed with concordances at very little cost and great benefit to readers. In the last few years concordances have been created for English (UK & US), Spanish, Swahili, Latvian, Russian, Portuguese (Brazil & Europe), Albanian, Solomon Islands Pidgin, Burmese languages, Tzotzil, Quechua, Ayamara. Work is presently in hand with support from BFBS on concordances in Kinyarwanda, Chichewa, Lithuanian and French. The system has also been released generally to the Bible translation community and is now in use world-wide.

The method described above is very closely tailored to creating biblical concordances quickly and easily. Nevertheless the technologies which automate much of the process are not really specific to the Bible translation environment. The key process is that of automatic glossing. Where large corpora of text must be made available in more than one language this technology offers the prospect of automating the creation of indexes and glossaries without the need for supplied lexica and linguistic tables. The key pre-requisite is that it is possible

to align the translated text reasonable closely. Technical and legal documents often provide a clearly defined structure of sections and paragraphs which can be mapped across translations in much the same way that the chapter and verse structure is common to all translations of the bible. There is, in principle, no reason why this technology could not be applied in a wider context.

For more information on the work of the Linguistic Computing team at British & Foreign Bible Society visit http://lc.bfbs.org.uk.

References

[1] A Cruden. *Cruden's Complete Concordance.* Hendrickson Publishers Inc, 1980.

[2] BFBS Eds. *Good News Bible.* Bible Society, 1976.

[3] BFBS Eds. *Y Beibl Cymraeg Newydd.* Bible Society, 1988.

[4] OE Evans. *Mynegair I'r Beibl Cymraeg Newydd.* Bible Society, 1998.

[5] J Keay. *Alexander the Corrector, the tormented genius whose Cruden's Concordance unwrote the Bible.* Overlook Press, 2005.

[6] NW Rees. Analysis of canonical issues in paratext. http://lc.bfbs.org.uk/request.php?canonicalissuesinparatext.pdf, June 2009.

[7] NW Rees. Paratext concordance builder. http://paratext.ubs-translations.org/about/cb, July 2009.

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

[9] JD Riding. Statistical glossing, language independent analysis in bible translation. In *Translating and the Computer 30.* ASLIB, ASLIB/IMI, 2008.

[10] DWC Robinson. *Concordance to the Good News Bible.* Bible Society, 1983.

[11] DWC Robinson. *Good News Bible with Concordance.* Bible Society, 1987.