Creating a Parallel Corpus from the "Book of 2000 Tongues"

Philip Resnik, Mari Broman Olsen, Mona Diab
University of Maryland Department of Linguistics and
Institute for Advanced Computer Studies
{*resnik,molsen,mdiab*}*@umiacs.umd.edu*

1 Abstract

This paper reports on a project to annotate biblical texts in order to create an aligned multilingual Bible corpus for linguistic research, particularly computational linguistics, including automatically creating and evaluating translation lexicons and semantically tagged texts. The output of this project will enable researchers to take advantage of parallel translations across a wider number of languages than previously available, providing, with relatively little effort, a corpus that contains careful translations and reliable alignment at the near-sentence level. We discuss the nature of the text, our annotation process, and intended uses for the corpus, and we point out relevant aspects and potential limitations of the current draft of the Corpus Encoding Standard with respect to this corpus.

2 Why this text?

2.1 The nature of the text

The Bible is a widely available, representative sample of carefully translated texts in a variety of styles in a wide range of languages. These properties uniquely suit our research purposes, which include construction of translation lexicons, and evaluation of semantic tagging for multilingual machine translation and other natural language processing applications. The text is a single cohesive document comprising 66 books by 30-40 authors in a variety of text styles. The corpus provides a representative sample of language styles in the source texts, including narrative, poetry, and correspondence. The New Testament corpus alone "compares favourably in size to other major collections analysed by scholars ... approximately as large as if not larger than the corpus of Homer's Iliad, of Homer's Odyssey, of Sophocles, of Aeschylus, of Herodotus ... [with] individual books ... comparable in size to other well-known classical texts: e.g. Plato's Apology approximates the size of Paul's Romans or 1 Corinthians" (Porter, 1989).

As a resource for research using corpus-based statistical methods in computational linguistics, the Bible is small by current standards (e.g. see (Church and Mercer, 1993)); with some variation for language and translation, it is typically on the order of 800,000 words and 4-5 megabytes. However, this is on the order of some monolingual corpora widely used for corpus-based research, such as the Brown Corpus of American English (Kučera and Francis, 1967), and the breadth across multiple languages offers an opportunity for research not generally available with the larger corpora in use today.

2.2 Availability

Originally written in Hebrew, Aramaic, and Greek between 1000 B.C. and 100 A.D., the Bible is the world's most translated book. The first complete translation, the Latin "Vulgate," or common version, was made in the 4th century by Jerome. By 1804 there were 67 languages with at least one book translated; almost 200 years later that number tops 2100, with more than 350 complete Bibles, and 880 New Testaments.[1] Additional translations are in process, including almost 500 new languages by United Bible Society personnel alone[2]. Many versions are now available electronically, either as text in the public domain, or bundled with search software for a modest fee. See for example, Bibleworks[3] and Techflow biblical software,[4] as well as Section 3.5.

The United Bible Society maintains a site that provides information on whether and when translations were first made in a given language.[5] Sample output is given below. Each response includes a link to the local Bible Society, from which (certain) texts may be ordered (at least in print).

```
You searched for Warlpiri
This language is sometimes called Wailbri or Walpiri.
Warlpiri is spoken in Australia (Northern Terr., Hooker Creek)
 First publication of:
 A single book of the Bible                      1985

You searched for sorbish
sorbish is referred to in this database as Sorbian: Upper.
Sorbian: Upper is spoken in Germany (SE, Upper Saxony.)
 First publication of:
 A single book of the Bible                      1670
 The New Testament                               1706
 The Bible                                       1728

You searched for Kung
That language name could not be found in the database. . . .
Possible Matches: Languages in the database that closely match your query :
   kung: ekoka
   kung: tsumkwe
Possible Matches: Variant language names that match your query : tsumkwe kung
```

[1]⟨http://www.lib.cam.ac.uk/Handbook/Guide15.html#tag1, http://www.biblesociety.org/trans-gr.htm⟩
[2]⟨http://www.biblesociety.org/translat.htm⟩
[3]⟨http://www.bibleworks.com/contacts.htm, http://www.omroep.nl/eo/bible/⟩
[4]⟨http://www.techflow.com.au/Bible.htm⟩
[5]⟨http://www.biblesociety.org/translat.htm⟩

The American Bible Society is experimenting with HTML markup and web distribution of more comprehensive information on all translations. The English sample below is from their *Book of 2000 Tongues* project, an update of their *Book of 1000 Tongues* (Liana Lupas and Erroll F. Rhodes, eds., 1939 and 1972).[6]

```
ENGLISH

Speakers: 450,000,000 first language speakers (1991 est.);
800,000,000 total including second language speakers (Ethn12).

Location: United Kingdom, United States, international.

Kinship: Indo-European / Germanic / West / North Sea / English.

1526 New Testament [Repr.+1836, +1837, +1989; Facs.  1862, +1976]
Peter Schoeffer, Worms

1530 +Pentateuch [Repr. +1967, +1992] Hans Luft, Marburg
(= J. Hoochstraten, Antwerp)

1531 Jonah [Facs. +1863] Martin de Kayser, Antwerp?  Translated by
William Tyndale (Hychyns). Only 10 sheets of Matthew were printed in
1525 by P.  Quentell, Cologne, when work was interrupted to be
resumed afresh at Worms. Revisions of the New Testament by Tyndale
himself appeared in 1534 [Repr. +1938] and 1535 (often reprinted);
the revised Pentateuch was published in 1534.

...
```

A number of versions the Bible are available in electronic media and on the Web, including multiple language versions – for example, the Bible Gateway web site makes available multiple English translations (NIV, NASB, RSV, KJV, Darby, YLT) and versions in German, Swedish, Latin, French, Spanish, and Tagalog.[7] However, our review of versions of the Bible available on-line indicates that, while most versions are in a format useful for browsing and searching, there is no *parallel corpus* of the Bible, in the sense of a collection of documents that is both marked up monolingually according to a standard set of conventions and also explicitly aligned across languages.

For example, the Bible Gateway site makes it possible for a user to retrieve particular passages, and even entire chapters at once. However, the markup in the retrieved text is presentational, and does not make the document structure explicit enough for automatically identifying the same verse

[6]⟨http://www.americanbible.org/2000.html⟩ and ⟨http://www.americanbible.org/2000txt.html⟩
[7]⟨http://bible.gospelcom.net/bible/⟩

in multiple languages. As a case in point, Genesis 1:3-4 in an English (NIV) and French version are encoded as follows:[8]

```
<DT>3<DD>And God said, "Let there be light," and there was light.
<DT>4<DD>God saw that the light was good, and he separated the light
        from the darkness.


<DT>3<DD>Dieu dit: Que la lumi\`{e}re soit! Et la lumi\`{e}re fut.
<DT>4<DD>Dieu vit que la lumi\`{e}re \'{e}tait bonne; et Dieu
        s\'{e}para la lumi\`{e}re d'avec les t\'{e}n\`{e}bres.
```

In summary, the Bible is available in print form for a huge range of languages, and in on-line form for a respectable and growing subset of those languages. However, to our knowledge the project of creating a *parallel corpus* for the Bible has not been previously attempted.

2.3 Careful translation

Because to translators the Bible represents God's Word to the faithful, it is not only among the world's most translated texts but also among the most carefully translated. As suggested by the explosive increase in number of translations described above, wide use and acceptance of translations is a relatively recent phenomenon. A large part of the history of translation theory in general is directly or indirectly due to past and present biblical translation (Nida, 1964; Nida and Taber, 1969; Robinson, 1991). Much current theory and practice grew out of the zeal of Protestant missionaries of the past two decades, eager to present the Bible in the language of the people (Bassnet-McGuire, 1980).

Numerous publications on Bible translation discuss nuances and difficulties of translation at a microscopic level, sometimes providing sentence-by-sentence guidance to translators at the linguistic, semantic, and pragmatic levels (see, for example, (Beekman and Callow, 1974; Blight, 1992; Deibler, 1993; Moore, 1993; deWaard and Smalley, 1979)). Similar care is taken in the introduction of new translations: a recent attempt to introduce gender neutral language in the New International Version (NIV), last revised in 1983, met with a great outcry in some segments of the population, leading the publisher to abandon the project (LeBlanc, 1997). Translations in established languages, therefore, tend to be conservative. And first translations in a given language are subject to rigorous scrutiny at every stage. We may therefore be confident that the texts we have are as accurate as humanly possible.

In languages with multiple translations, texts could also be paired according to age and style of translation: there are many translation contemporaries of the King James Version, for example. These include Luther's German and the Spanish Reina de Valera, which have a literary feel and may be too formal for some purposes. In contrast, versions published by the United Bible Societies and

[8]The French source is encoded using ISO-8859-1 (Latin-1); for readability here we have replaced accented characters with their LaTeX encodings, which transparently associate an accent with a character, e.g. a backquote for a grave accent.

the Summer Institute of Linguistics tend to follow "dynamic equivalence" theories of translation (Nida, 1964; Nida and Taber, 1969), attempting to make the impact of the original text idiomatic for today. For example, an original Greek phrase in Colossians 1:20 translates literally as "the blood of his cross," whereas the Good News Bible has "God made peace through *his Son's sacrificial death on the cross*" (GNB, 1976), emphasis added. Alignment of such translations would therefore serve as an important source for pairs of idioms and figures of speech.[9]

2.4 Standard structure and verse alignment

One of the difficulties with parallel corpora is that most often they are not explicitly aligned — for example, a considerable amount of work has been done attempting to automatically align the Canadian Hansards, parliamentary proceedings in English and French, at the sentence (and sometimes word) level.[10] Because the Bible's structure is fully standardized in terms of books, chapters, and verses, alignment at the verse level comes essentially for free, and in fact the main aim of this project is to represent that standardized structure in a consistent format.

Within verses, of course, there is considerable variation in translations, and so research requiring finer-grained alignments, e.g. word-aligned parallel corpora, will still require further work. However, the consistent verse-level alignments provide appropriate training material for algorithms that learn to do lower-level alignments on the basis of correctly aligned text, e.g. (Melamed, 1996a), and can also be used as a source of test material for algorithms that attempt to produce sentence-level alignments.

The structure inherent in the Bible also eliminates some problems of omissions in parallel text. Later translations eliminate, relocate, or footnote passages found in the King James Version (as well as its contemporaries and descendants). For example, the last part of Mark 16:8 and John 7:53-8:11 are contested: both are attested in the plurality of manuscripts, but not in the oldest texts, discovered since the KJV translation. Verse alignment limits the impact of such omissions, as such cases would result simply in null pairings. (See (Melamed, 1996c) for discussion of automatic methods for detecting omissions in translations.)

3 Annotation

3.1 Intermediate Format

Annotating individual language versions of the 66 books of the Bible (or, in some cases, the New Testament) requires only a simple 3-level hierarchy of text elements (book, chapter, verse). In our initial pass through the annotation process (see below), we are labeling elements as `b` (book), `c`

[9] Special issues arise in automatically creating translation lexicons that include non-compositional pairs. See, e.g. (Melamed, 1997) and Section 4.2.

[10] ⟨http://www-rali.iro.umontreal.ca/TransSearch/TS-simple-uen.cgi⟩

(chapter), and `v` (verse), producing an intermediate representation that captures the major structural levels without conforming to any particular DTD. The following examples show a single verse, Matthew 1:7, in 9 languages:[11]

ENGLISH: <v id="MAT:1:7">And Solomon begat Roboam; and Roboam begat Abia; and Abia begat Asa;</v>

FRENCH: <v id="MAT:1:7">Salomon engendra Roboam; Roboam engendra Abia; Abia engendra Asa;</v>

DANISH: <v id="MAT:1:7">og Salomon avlede Roboam; og Roboam avlede Abia; og Abia avlede Asa;</v>

FINNISH: <v id="MAT:1:7">Salomolle syntyi Rehabeam, Rehabeamille syntyi Abia, Abialle syntyi Aasa; </v>

GREEK: <v id="MAT:1:7">solomwn de egennhsen ton roboam roboam de egennhsen ton abia abia de egennhsen ton asa</v>

LATIN: <v id="MAT:1:7">Salomon autem genuit Roboam Roboam autem genuit Abiam Abia autem genuit Asa </v>

SWEDISH: <v id="MAT:01:7">Salomo födde Roboam, Roboam födde Abia. Abia födde Asaf;</v>

SPANISH: <v id="MAT:1:7">Salomón Engendró a Roboam; Roboam Engendró a Abas; Abás Engendró a Asa;</v>

VIETNAMESE: <v id="MAT:1:7">Salomôn sinh Roboam, Roboam sinh Abya, Abya sinh Asa, </v>

In all these cases, the intermediate encoding for book and chapter elements are identical:

```
<b id="MAT">
<c id="MAT:1">
...
</c>
</b>
```

The labels (`id` attributes) for elements make it possible to identify verses in a context-independent way by including the book and chapter in the label, e.g. "GEN:1:1" for Genesis, chapter 1, verse 1. This will allow users to take advantage of simple tools such as Unix 'grep' for simple day-to-day manipulation (for example, needing to look up a particular verse) while also being able to utilize more powerful SGML-based tools.

[11] In the final version, accented characters will of course be encoded according to SGML conventions; here they are left in a simpler form for readability.

3.2 Target Format

For our target annotation format, we are aiming for verse-level annotation of structure conforming to the Corpus Encoding Standard (CES) subset of the TEI (Ide, 1996). In many respects, creating a parallel corpus from multiple versions of the Bible is an excellent match for the CES. Since the corpus is being created primarily for use in corpus-based computational linguistics research, the restrictions imposed by the CES, inn comparison to the full generality of the TEI, are suited to the task (CES Sec. 0.2.3). Moreover, the CES contains useful and explicit guidelines for the encoding, and the consistent structure and content of Bible text should make it straightforward to achieve not only Level 1 but Level 2 conformance to those guidelines, using fully automatic conversion of original files. (Level 2 conformance goes beyond the minimum by requiring both correct paragraph-level markup and consistent marking of some sub-paragraph elements; Level 3 conformance is not a goal since reliable identification of all the specified sub-paragraph elements, particularly names, would require significant manual effort.)

Finally, the cesAlign encoding conventions for parallel corpora are ideally suited for the present task, since they permit an arbitrary degree of parallelism, and because the recording of alignment information in an external document makes it trivial to work with a monolingual subset or any n-way parallel subset.[12] CesAlign specifies the form for a separate alignment document linking existing documents; the alignment document can be created for a pair of Bible versions trivially by encoding one-to-one links between book/chapter/verse labels, as illustrated here:

```
<link xtargets="GEN:1:1 ; GEN:1:1">
```

Alignment at the sub-verse level would be considerably more tricky, of course, since different translations reflect different decisions about how verses are broken into clauses, etc. We leave this as a potential problem for future work.

Despite the fact that the CES is well suited for our task in many ways, we can suggest two ways in which the current CES draft may be problematic for our purposes. The first is merely a question of its scope, and may be remedied as the standard develops: one goal of our corpus-based research using the Bible is to investigate word sense and semantic issues, and these are explicitly outside the purview of the current CES draft (CES Sec. 0.2.4).

Second, and more important, we observe that the verse structure of the Bible does not respect the linguistic subdivisions chosen in the CES, at least with regard to the encoding standard for primary data (cesDoc). Built into the standard are basic elements of paragraph, sentence, token; however, verses can contain material above sentence level, as in (1), as well as sub-sentential units, as found in the two verses in (2).

```
(1) <v id="GEN:1:31">And God saw every thing that he had made, and,
    behold, it was very good. And the evening and the morning were the
    sixth day. </v>
```

[12]⟨http://www.cs.vassar.edu/CES/CES1-5.html⟩

```
(2) <v id="GEN:10:13">And Mizraim begat Ludim, and Anamim, and Lehabim,
    and Naphtuhim, </v>

    <v id="GEN:10:14">And Pathrusim, and Casluhim,
    (out of whom came Philistim,) and Caphtorim. </v>
```

One alternative would be to use the cesDoc DTD, [13] using the `div` element to identify chapters and treating verses as paragraph-level elements — this would produce a "Level 1 CES-conformant" encoding. However, identifying Bible verses with either paragraph- or sentence-level elements would sacrifice standardization at the semantic level (CES Sec. 1.3.3), since "sentence" or "paragraph" for this corpus would mean something different than the conventional meaning of those terms for other corpora.

Another alternative would be to utilize the notion of a *chunk* in the cesAna DTD for encoding linguistic annotation, annotating each verse as a chunk comprising the series of tokens within that verse. This would preserve adherence to the standard at the semantic level, but would sacrifice the notion of *verse* as a meaningful structural element at the level of the primary encoding — instead shifting the burden to the linguistic level of encoding (following the cesAna DTD). This seems fairly unnatural.

We observe that this problem is potentially more general than the specific application of the CES to annotating Bible text. For example, we expect that encoding speech data will present similar problems. The basic structural element in conversational speech is the turn (e.g. see the Child Language Data Exchange System database, (MacWhinney, 1991), which contains transcripts of conversations), and like verses, turns may comprise material both below and above the level of the sentence.

3.3 Input Formats

Within a particular electronic version of the Bible, we have observed that data formats are fairly consistent. And once low-level character set issues are dealt with — some pertaining to non-Latin character sets, and some involving the transition from a PC to Unix platform — the input formats seem to group according to a reasonably small set of dimensions. These include:

1. Line breaks: whether or not verses are implicitly delimited by appearing one per line, or broken across lines and delimited in another fashion.
2. Labels: whether book labels appear explicitly in a file or are implicit in the file name; whether verses are explicitly numbered and, if so, whether those labels also include chapter and verse (e.g. "1" vs. "1:1").
3. Header information: whether files contain information regarding the edition, translation, etc.

[13]⟨http://www.cs.vassar.edu/CES/CES1-4.5.html⟩

4. Formatting codes: whether the documents are essentially in plain-text format or contain embedded formatting.

An on-line Swahili version of the New Testament, for example, illustrates embedded formatting, with separate marking for chapters and verses (Matthew 2:1-2):

```
\c 2
\s Wageni kutoka mashariki
\p
\v 1 Yesu alizaliwa mjini Bethlehemu, mkoani Yudea, wakati Herode
alipokuwa mfalme. Punde tu baada ya kuzaliwa kwake, wataalamu wa nyota
kutoka mashariki walifika Yerusalemu,
\v 2 wakauliza, <<Yuko wapi mtoto, Mfalme wa Wayahudi, aliyezaliwa?
Tumeiona nyota yake ilipotokea mashariki, tukaja kumwabudu.>>
\p
```

A French version illustrates plain text with one verse per line, as well as the name of the book being repeated with each chapter heading (Matthew 2:1-2):[14]

```
Matthieu   2
 1. J\'esus \'etant n\'e \`a Bethl\'ehem en Jud\'ee, au temps du \
 roi H\'erode, voici des mages d'Orient arriv\`erent \`a            \
 J\'erusalem,
 2 et dirent: O\`u est le roi des Juifs qui vient de na\^itre? car   \
 nous avons vu son \'etoile en Orient, et nous sommes venus pour      \
 l'adorer.
```

The simple, uniform structure of the source text appears to greatly reduce the variation in document encoding for the on-line source documents. Minor variation within a version does occur, for example verse numbers sometimes being followed by a period and sometimes not, but these are easily handled. By organizing the annotated versions book by book, we eliminate potential problems in reordering — for example, the book of Hebrews is the 58th book in the English Bible, and the 63rd in the German Bible, although the relative order of every other book is identical.

3.4 The Annotation Process

This project could not have been undertaken if the input or target formats required any but the most minimal manual effort. Fortunately, it appears that a simple perl script suffices for each of the versions we have looked at so far, and once a script has been written for one language it is adapted fairly easily for others. We do not yet have firm data on how long it takes to write or adapt scripts, but in simple cases scripts have been written and executed in less than an hour; we

[14]We use the backslash here to indicate line continuations, and although the character encoding is ISO-8859-1 (Latin-1) we use the LaTeX encoding of accented characters here for readability.

are confident that an input version in any given language should not require more than a few days' effort at the outside. While code fragments are generally not suitable for inclusion in a paper of this kind, the simplicity of the program's main loop should be apparent even for those not familiar with the perl programming language:

```
while (<STDIN>)
{
  if (($chapter, $verse, $line) = /^(\d+):(\d+)\s+(.*)\r$/)
  {
    # Possibly deal with boundary between chapters
    if ($chapter != $current_chapter)
    {
       # Close old chapter element if there is one
       # and upate current chapter
       if ($current_chapter != 0)
       {
          print "</c>\n";
       }

       # Open new chapter element
       $current_chapter = $chapter;
       print "<c id=\"$book:$chapter\">\n";
    }

    # Print verse element
    print "<v id=\"$book:$chapter:$verse\">$line</v>\n";

  }
}
```

3.5 Status

As of this writing, English, French, Danish, Chinese, Finnish, Greek, Latin, Swedish, Spanish, and Vietnamese versions of the complete 66-book Bible have been annotated in the intermediate book/chapter/verse format. Our next main goal is to begin conversion to a suitable CES target format; in view of the issues raised in Section 3.2, we welcome the feedback of the community as to the format of the end result. We hope to make rapid progress through the remainder of current inventory of electronic source versions; currently our total set includes Arabic, Chinese, Danish, English, Finnish, French, German, Greek, Hungarian, Korean, Latin, Quechua, Swahili, Swedish, Tagalog, Turkish, Vietnamese, and Warlpiri.

We are in the process of reviewing the status of each Bible version with respect to restrictions on its redistribution. As most of the versions we have looked at were publicly available on the World Wide Web, we are optimistic about the possibility of our making our annotated versions available. Even if we are unable to redistribute our versions of the text, however, an alternative mechanism exists for making available TEI-encoded versions of even commercial versions of the on-line Bible: we will release the annotation scripts and instructions for their use, and members of the community can acquire source versions themselves (via downloading of publicly available versions, or purchase of commercial versions), creating an annotated version of the text for their own use with our software.

4 Research Uses for Aligned Bibles

4.1 Translation and Comparative Linguistics

An important part of the translation process is the comparison of previous translations, into the same or different translations. Miles Coverdale, translator of the first printed English Bible (in 1526) writes: "one translation declareth, openeth and illustrateth another, and ... in many cases one is a plain commentary unto another" (quoted in the introduction to (Vaughan, 1967)). The KJV translators also stood firmly on the shoulders of the translation giants, including in their subtitle ...Translated out of the Original Tongues and with the Former Translations Diligently Compared and Revised (KJV,).

Having an aligned text of the type we describe facilitates research in the original languages of the Bible, as well as in comparative linguistics more generally. Particularly in langauges with no living speakers, the subtleties of the text are revealed primarily through examination of a variety of translated forms by expert scholars. The work of Olsen (1997) illustrates this methodology. In that research, Olsen explores, among other things, the potential translations of certain New Testament Greek grammatical forms as a way of discovering the range of meanings conveyed by such forms. The resulting data were used for a theoretical separation between the semantic (uncancelable) and pragmatic (cancelable and variable) meaning, a distinction important for computational lexicons.

4.2 Resource acquisition for natural language processing

Parallel corpora are increasingly of interest in natural language processing, with applications in cross-language information retrieval (Hull and Oard, 1997), machine translation (e.g. (Brown et al., 1990)), in approaches to word sense disambiguation (Brown et al., 1991), and in computational lexicography (Melamed, 1996b) . However, corpora reliably aligned at the word or even the sentence level are difficult to obtain even for commonly found language pairs, and for "low density" languages – those for which few resources exist – parallel corpora are even mopre difficult to find.

The Bible is an interesting alternative to investigate: as discussed above, it can potentially yield a multi-way parallel corpus with representation from every language family, with the content carefully translated and nearly sentence-level alignment included. Although it is not the largest of corpora, parallel corpora of significantly smaller size have yielded useful results, e.g. (Resnik and Melamed, 1997), and although its content is more specialized than, say, contemporary newspaper text, it does cover a very wide range of linguistic phenomena and domains of world knowledge; for example, see the range of conceptual categories in the Louw-Nida thesaurus for the New Testament (Louw and Nida, 1989).

We plan to investigate parallel versions of the Bible as a possible resource for bootstrapping natural language resources, especially for work in machine translation, first by applying the techniques described by Resnik and Melamed for extracting and assessing word correspondences, and then using techniques for identifying multi-word units (Melamed, 1997). We also plan to evaluate the coverage of the Bible with respect to vocabulary and conceptual content by comparing it with existing lexicons for interlingual machine translation (Dorr, To appear; Dorr and Olsen, 1997) and thesauri such as WordNet (Miller, 1990).

In particular, we would like to compare the translation lexicons we create to bilingual lexicons automatically acquired by other means. (We have available to us Spanish-English and Arabic-English, with Korean-English in process, with other projects planned (Dorr, To appear)). We would like to investigate (i) the extent to which the biblically-based lexicons could be considered a "core" or "seed" lexicons, and (ii) what would be needed (in terms of coverage and resources) to scale up the biblically-based lexicons.

5 Conclusion

We have reported on a project to annotate biblical texts in order to create an aligned multilingual Bible corpus for research purposes. At present, we have implemented a standard intermediate-level annotation, delimiting book, chapter, and verse, for a growing collection of languages. The availability of on-line versions of the Bible leads us to be optimistic about the prospect of creating a resource that covers a wide variety of languages and will be valuable to specialists in translation, linguistics, and the computational analysis of language.

6 Acknowledgments

This work was supported, in part, by Department of Defense contract MDA90496C1250, DARPA/ITO Contract N66001-97-C-8540, and a grant from Sun Microsystems Laboratories. We are grateful to Dan Melamed and two anonymous reviewers for helpful comments relating to this work.

References

Bassnet-McGuire, Susan. 1980. *Translation Studies.* Methuen, New York.

Beekman, John and John Callow. 1974. *Translating the Word of God.* Zondervan, Grand Rapids, MI.

Blight, Richard C. 1992. *Translation Problems from A to Z.* Summer Institute of Linguistics.

Brown, P., J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, R. Mercer, and P. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

Brown, P., S. Della Pietra, V. Della Pietra, and R. Mercer. 1991. A statistical approach to sense disambiguation in machine translation. In *Fourth DARPA Workshop on Speech and Natural Language*, Pacific Grove, CA, February.

Church, Kenneth W. and Robert Mercer. 1993. Introduction to the special issue on computational linguistics using large corpora. *Computational Linguistics*, 19(1):1–24.

Deibler, Ellis. 1993. *An index of implicit information in the Gospels.* Summer Institute of Linguistics.

deWaard, Jan and William A. Smalley. 1979. *A Translator's Handook to the Book of Amos.* United Bible Society.

Dorr, Bonnie J. To appear. Large-Scale Dictionary Construction for Foreign Language Tutoring and Interlingual Machine Translation. *Machine Translation*, 12(1).

Dorr, Bonnie J. and Mari Broman Olsen. 1997. Deriving Verbal and Compositional Lexical Aspect for NLP Applications. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL-97)*, pages 151–158, Madrid, Spain, July 7-12.

1976. *Good News Bible: The Bible in Today's English Version.* American Bible Society, New York.

Hull, David A. and Douglas W. Oard. 1997. Symposium on cross-language text and speech retrieval. Technical Report SS-97-04, American Association for Artificial Intelligence, Menlo Park, CA, March.

Ide, N. 1996. Corpus encoding standard. Available at http://www.cs.vassar.edu/CES/.

The Holy Bible, Authorized King James Version. World, Cleveland and New York.

Kučera, H. and W. Francis. 1967. *Computational analysis of present-day American English.* Brown University Press: Providence, R.I.

LeBlanc, Doug. 1997. Hands off my NIV! *Christianity Today*, June.

Louw, Johannes P. and Eugene A. Nida. 1989. *Greek-English lexicon of the New Testament based on semantic domains.* United Bible Societies, New York. 2nd edition.

MacWhinney, Brian. 1991. *The CHILDES project: tools for analyzing talk.* Erlbaum.

Melamed, I. D. 1996a. A geometric approach to mapping bitext correspondence. In *Conference on Empirical Methods in Natural Language Processing*, Philadelphia, Pennsylvania.

Melamed, I. Dan. 1996b. Automatic construction of clean broad-coverage translation lexicons. In *Proceedings of the 2nd Conference of the Association for Machine Translation in the Americas*, Montreal, Canada.

Melamed, I. Dan. 1996c. Automatic detection of omissions in translations. In *Proceedings of the 16th Annual Conference on Computational Linguistics (COLING-96)*, Copenhagen.

Melamed, I. Dan. 1997. Automatic discovery of non-compositional compounds in parallel data. In *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing (EMNLP-97)*, Brown University, August.

Miller, George. 1990. WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4). (Special Issue).

Moore, Bruce R. 1993. *Doublets in the New Testament.* Summer Institute of Linguistics.

Nida, Eugene A. 1964. *Towards a Science of Translating.* E.J. Brill, Leiden.

Nida, Eugene A. and Charles R. Taber. 1969. *The Theory and Practice of Translation.*

Olsen, Mari Broman. 1997. *A Semantic and Pragmatic Model of Lexical and Grammatical Aspect.* Garland, New York.

Porter, Stanley. 1989. Verbal Aspect in the Greek of the New Testament, with Reference to Tense and Mood. In D.A. Carson, editor, *Studies in Biblical Greek, Vol. 1.* Peter Lang, New York.

Resnik, Philip and I. Dan Melamed. 1997. Semi-automatic acquisition of domain-specific translation lexicons. In *Fifth Conference on Applied Natural Language Processing*, Washington, D.C.

Robinson, Douglas. 1991. *The Translator's Turn.* The Johns Hopkins University Press, Baltimore and London.

Vaughan, Curtis. 1967. *The New Testament from 26 Translations.* Zondervan, Grand Rapids, MI.