Tyndale STEPBible Data development for machine analysis and computational linguistics

David Instone-Brewer
Tyndale House, Cambridge, 36 Selwyn Gardens
Tech@Tyndale.cam.ac.uk

Abstract: The development of STEPBible.org by Tyndale House Cambridge, which aimed to provide study tools for the disadvantaged world, resulted in creating datasets which are useful for many other purposes. In particular, computational linguistics and other machine analysis can benefit from the more stringent and varied ways in which data has been refined and presented. Much of this data resulted from a project supported by ETEN to automatically tag Bibles to Greek and Hebrew. A public repository is gradually being populated with the results of this work.

Keywords: computational linguistics, semantic range, Biblical Hebrew, Biblical Greek

Introduction to STEPBible.org

The purpose of STEPBible.org, developed by Tyndale House Cambridge from 2012, was to provide reliable study tools for Bibles in multiple languages – hence the acronym Scripture Tools for Every Person: STEP. Electronic Bibles are increasingly available for reading, but they have been difficult to compare with each other and Tyndale House especially wanted to encourage comparisons with the original Greek and Hebrew. This would enable readers to understand translation issues and provide tools for intelligent study of their own translations with respect to the underlying text.

The OSIS format developed by Crosswire.org, together with their Java code (the JSword project) was chosen as a foundation, and this is still the core code for packaging and searching Bible data. This was built as web-based tools which allowed users to search and present Bibles in a variety of ways. Searches include words in the translation, topics, and words in the original Hebrew and Greek (in translation, transliteration or Unicode characters). Presentations included interleaving different versions verse by verse, side by side with the option of highlighted differences, and word-based interlinear for any translations that had sufficient tagging.

Multiple language support was critical for most users, so about 50 language interfaces were developed using machine translation with the facility for users to suggest corrections, and even add a language which wasn't available. The first to be added by users was Ukrainian, which was a wonderful way for people who had been forced to speak Russian for decades to celebrate their newfound freedom. There are currently 59 language interfaces, which allows most people to use the software in their mother tongue or major second language.

The first version presented three levels of complexity, with different search tools and presentations depending on whether someone wanted to simply read the text, or study it like a non-scholar or like a scholar. Some search options were so complex that the multitude of drop-down menus for a single lookup couldn't be seen without scrolling down the screen.

The second version merged almost all of this functionality into a single search box which presents the user with multiple types of search that they may be looking for. So someone who starts typing "Heb" is presented with a set of References from the letter to Hebrews, Bibles in Hebrew, Greek

and Hebrew words meaning "Hebrew" listed by translation and in their own alphabets, topics such as Heber, Hebrews and Hebron, and an offer to look up "Heb" in the text of translations that are open. This is initially daunting but undeniably powerful, especially when one realises that these can all be mixed. That is, one can look in Hebrews for the Greek word normally translated "Hebrew" when it occurs in passages concerning Hebrew people and specify that it must be translated as "Hebrew*" (i.e. Hebrew or Hebrews) in at least one of the translations that have been selected. If NIV, ESV and KJV are searched in this way, the only result is the colophon which is present only in KJV:

Figure 1. Complex search at StepBible.org. [1]

The disadvantaged world is a primary target group for these facilities rather than the first world where tools such as BibleWorks, Accordance and Logos provide similar facilities. Because of the risk of intermittent or patchy internet coverage, an offline version was developed which installs the complete code along with Apache to serve any browser with the same facilities when they are offline. This works with PC, Mac and some flavours of Linux.[2]

Initially the goal was to equal the facilities of the commercial tools, but a limiting factor was the available data. Although these software packages were built initially from free datasets, they had all improved them, partly by employing experts though often thanks to corrections sent in by enthusiastic users. The inaccuracies within the free data soon became a limitation for STEPBible, and this was compounded by independent initiatives to correct or expand the data in different ways.

For example, one module may tag Genesis 2.20 with H0121 (a tag for "Adam") while another uses H0120 (a tag for "man"). Neither are wrong, but this fails to make an interlinear link. Also, some tagging systems take into account variants or give separate tags to homonyms – both of which result in a lack of link between the words of one text or translation and another. The Tyndale STEPBible project wanted to enable work from various sources to be studied together, irrespective of their tagging practices and philosophies.

STEPBible tackles this in two ways. First it records the possibility of different types of tags, so that the 'wrong' tag is still recognised. Hence KJV, ESV and Hebrew all tie up in Genesis 2:20 despite disagreements about precise tagging. There are still problems, as the screenshot below shows, but these are due to differences in tagging philosophy. The Hebrew words can be tagged word by word, but translations need to tag either the most significant word (as seen in the ESV tagging) or all the

[1] The repeatable URL for this search is: https://www.stepbible.org/?q=version=ESV|strong=G1445|meanings=hebrew|reference=Heb|text=Hebrew*|version=KJV|version=NIV&options=VHNUG&display=INTERLEAVED

[2] Downloadable from https://www.stepbible.org/downloads.jsp

translated words represented by a single Hebrew word (as seen in the KJV tagging) – though this often fails because of innate differences in the languages.

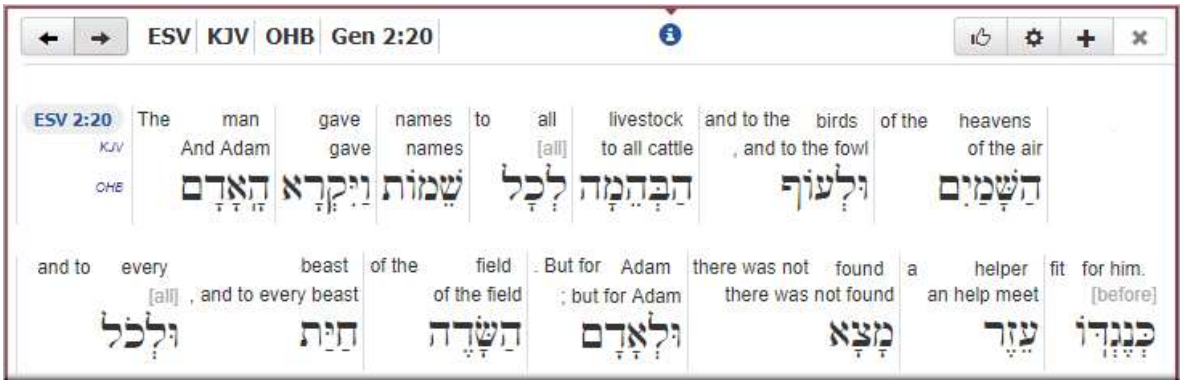

Figure 2. Interlinear lineup with different types of tagging.[3]

Another problem illustrated here is that the top translation determines the tagged words that are visible. If the Hebrew is placed at the top, the missing words *lo* (translated 'not' without a specific tag) becomes visible. Tagging systems are also victims of human vagaries. In this case, whoever tagged the KJV thought that "help meet" was a single word. They didn't realise that "meet" meant "suitable" in older English, so it should be tagged with *kenegdo* as in the ESV tagging.

Figure 3. Interlinear lineup with Hebrew determining the order.[4]

These and other frustrations motivated the development of datasets specifically for Tyndale STEP-Bible which could also be used by other projects, and for data refinement which would enable machine-based tagging, using humans as final correctors, but making initial decisions with consistency and surprising accuracy.

[3] The repeatable URL for this search is: https://www.stepbible.org/?q=version=ESV|version=KJV|reference=Gen.2.20|version=OHB&options=HVGUN&display=INTERLINEAR

[4] The repeatable URL for this search is: https://www.stepbible.org/?q=version=OHB|version=ESV|version=KJV|reference=Gen.2.20&options=UVLHGN&display=INTERLINEAR

The problem of data

Computers are useful agents for helping research, so long as we can provide data that is organised and consistent. In the past this has meant reducing data to ASCII or perhaps to simple numbers in small fields. This kind of structure still gets the best results, but it is difficult to restrict natural language in this way and new ways of organising data are being found.

Data in AI research

Good data is still needed even by AI neural nets, which can find patterns where even humans struggle – and this is remarkable because humans are innately good at recognising patterns. But before neural nets can do this, they have to be trained using well-labelled data. You can't expect a neural net to recognise the common features that make up a dog, if half the pictures presented to it don't contain a dog. And to start with, the images of dogs need to be isolated from other elements within a picture.

The importance of quality data for machine learning was recognised especially by Fei-Fei Li – who is now a Stanford professor and Google Cloud chief scientist. When everyone else was trying to improve algorithms, she realised that this would not work without improved data. She started up ImageNet – a collection that would grow to millions of tagged and classified pictures of everything.

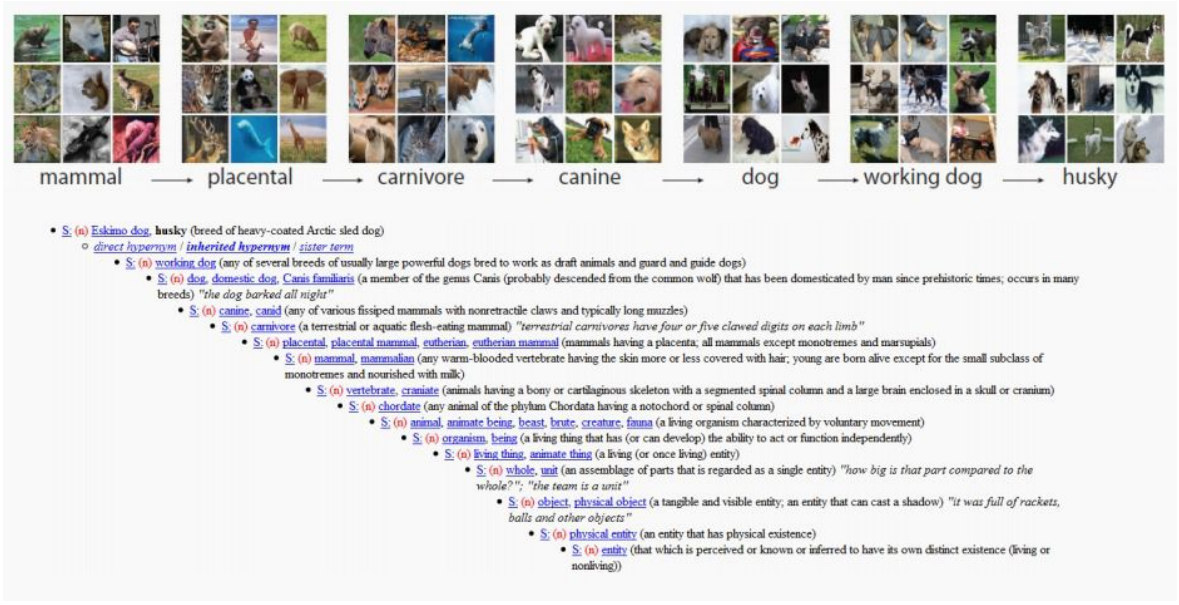

Figure 4. ImageNet classification for a husky dog.

Initially she couldn't get funding, so she used her spare time, volunteers, and whatever physical resources she could beg or borrow. Eventually, when her data started to be used by AI researchers, they saw their recognition rates go up from the 70 percentiles to 90s. Now image recognition software can recognise a dog sitting on someone's lap inside a crowded bus.[5]

[5] For more details see "The data that transformed AI research—and possibly the world" by Dave Gershgorn (Quartz July 26, 2017) at https://qz.com/1034972/

Improving Historic Biblical data

Strong's lexical tags extended for NASB & LXX

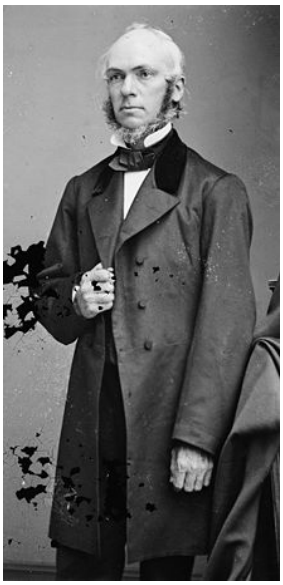

Figure 5. James Strong

The numerical system called Strong's Numbers originated with James Strong, the Mayor of Long Island and enthusiastic Bible scholar who produced his Concordance of the Bible in 1890. He was prescient enough to add a number to each Hebrew word. With the advent of computers, these numbers become an easy way to tag Bibles with the Hebrew and Greek originals.

Unfortunately, Strong was only interested in the Greek behind the King James version. Any Greek words that occurred in variant manuscripts were ignored, which means that most modern Bibles are missing about 120 words which occur in other NT manuscripts, 80 of which are used in most modern Bibles because they are occur in manuscripts that are older than those that were available to the King James translators.

When the NASB was tagged by Crosswire.org at the behest of the Lockman Foundation, they added those 80 words at the end of the Strong's numbers (#G6000-6090 with a few gaps). The others have been added to the STEPBible lexicon (#G6091-6099, G9980-9999).

A much larger gap existed for the Septuagint (LXX). The NT has a vocabulary of about 5500 words, but the LXX has almost 4000 more, not including the names. In the ApostolicBible project these were added as decimal points between existing Strong's numbers to maintain an alphabetic order.[6] In STEPBible these have been changed to simple integers as #G6100-9979. The ApostolicBible does not tag names, but STEPBible tags them with Greek numbers when they also occur in the NT, and with Hebrew numbers if they do not. There are also about 100 extra Greek names in the LXX with no Hebrew equivalent which have been added as extra OT Greek words.

It would be tidier and easier to simply ignore Strong's work and start again with an alphabetic list that contains all Bible Greek. However, STEPBible aims to maintain backward compatibility with all the modules that are already in existence, many of which are unlikely to be updated. The numerical order of lexical entries have little consequence for the user, who need never be concerned with them, unless they wish to use them as a quick way to type in Greek words that they are familiar with. So, untidy as it was, the integers for Strong's Greek will likely exist behind software for many more years.

Developing Greek Lexicons

The so-called 'lexicons' at the back of Strong's Concordance consist of a quick definition and notes on relationships with similar words. This was intended as a simple index to his Concordance – but in the era of software, these indexes became Hebrew and Greek 'lexicons'.

Linking these numbers to other lexicons is much more difficult than one might at first imagine. The Liddel-Scott-Jones lexicon of ancient Greek (LSJ) has about 100,000 entries. Only one tenth of these are used in the OT or NT, and the specific entry is often difficult to identify. The problems

[6] *The Analytical Lexicon of The Apostolic Bible Polyglot* by Charles Van der Pool - https://www.apostolicbible.com/analylex.htm

include differences in spelling (especially in Greek), deciding which form to use as the lexical entry, and when to split a word into more than one entry. Sometimes a word in the NT only occurs in a lexicon as a variant spelling of minority usage of a word with a generally different meaning. Some letters such as 'r' & 'l', or 't' & 's' were interchangeable because ancient Greeks couldn't hear the difference or local pronunciation was reflected in spelling. And Greek, like all languages, changed its meaning with time. Some software manages to make links to LSJ by automatic means, but for the reasons outlined above, this doesn't always work well. The Tyndale STEPBible project put significant time into identifying the correct LSJ lexical entry for every OT and NT Greek word.

Developing Hebrew Lexicons

The problem with Hebrew was much worse, because Hebrew lexicography was only getting started in Strong's day. Due to the remarkable explosion of knowledge of Akkadian, Ugaritic and other cognate languages, the understanding of Hebrew has grown immensely during the last couple of centuries. Just as the English word "bear" comes from two old German words, *beran* ('to carry') and *berô* ('a bear' – the mammal), we now know that many Hebrew words which had a wide semantic range are actually two or more separate words. The first verb in the Bible, *bara* "to create" also means "to fatten" (only used at 1Sam.2.29).

The OpenScriptures.org project has done a superb job of tying up Strong's data to the Brown-Driver-Briggs lexicon (BDB). This was the most significant of the early academic lexicons that benefitted from comparative language studies, and the decisions they made have been reflected in subsequent Hebrew lexicons. It is out of copyright, so this is ideal for software projects. They chose to augment Strong's numbers with "a", "b" etc, so for example *bara* is H1254a 'to create' and H1254b 'to fatten'. This has been adopted in STEPBible because it maintains backward compatibility with modules that have been created without this additional detail. Most software can be made to ignore the final letters and still work. STEPBible works differently because it upgrades all modules on-the-fly. So even if a Bible is tagged in the old way, someone clicking on *bara* at 1 Samuel 2.29 will not be linked to other places where it means 'create'.

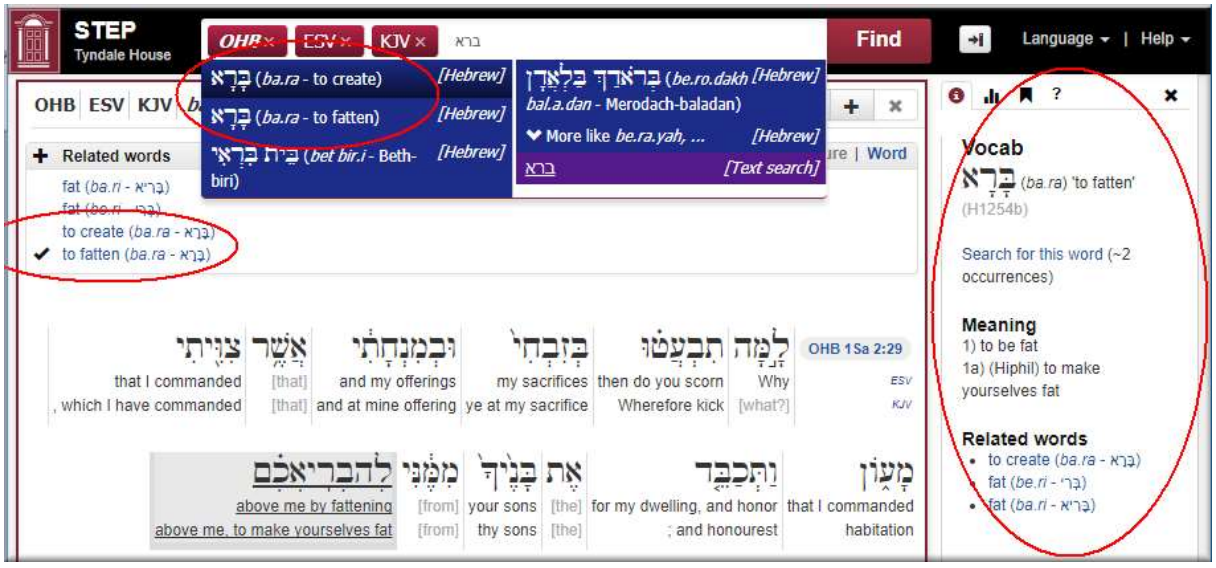

Figure 6. Alignment of *bara* with one of the two lexical entries in BDB.

Developing morphology

When Strong's tagging was developed for computers, some limited morphology codes were added, representing the moods and tenses of verbs. In the Greek numbering they were added at G5627-5799 and for Hebrew at H8686-8853 – both sets after the end of the original word numbering. I haven't been able to trace the originator of these codes, but it was probably OnlineBible.net led by Larry Pierce – an organisation that made the first tagged texts and generally did it very carefully. Although these morphology tags covered only the verbs, and the person and number etc were omitted, they provided exegetically rich information. However, most software ignored them.

In the mean time, morphological tagging carried out by people like Don Carson for OakTree has been developed by many others and the very good version by Tauber is available for free software. Hebrew morphology has been developed from scratch in three projects – by Pennsylvania's Westminster College, Amsterdam Free University, and OpenScriptures.org (though this latter project is not quite finished).

Significant progress

At this stage we should remember how hard Biblical Studies used to be without computers. For example, perhaps we would want to look up the Greek equivalent of Isaiah's phrase about people fainting in the streets "like an antelope in a net" (Isa.50.20). The urban sophisticates of Alexandria, where the Septuagint originated, had never seen an antelope hunt. So the translators used a completely different phrase, saying they were limp and faint "like half-cooked beet (σευτλίον, *seutlion*)" - which is rather like our modern phrase: 'like limp celery'. If you managed to line up the Hebrew with this very different Greek, you'd have problems looking up the meaning of σευτλίον because it is found under τεῦλον – as a variant spelling of the diminutive form. With software like STEPBible, all the work has been done by those who have tagged texts and lined up lexicons.

Figure 7. Lining up LXX & Hebrew with LSJ Lexicon entries.

Datasets produced by Tyndale House for STEPBible

STEPBible.org was initially based on modules created by others in the OSIS library of the Crosswire repository. These were based on data collected, entered and corrected over decades by non-commercial Christian software creators. This data has, on the whole, a high level of accuracy, though limitations were soon discovered. STEPBible.org is therefore built on the foundation of this data, but with significant corrections and developments.

A greater stimulus for creating accurate datasets came from a project generously supported by ETEN to develop an automatic tagging system whereby a translation could be linked to the underlying Greek and Hebrew words. The proof-of-concept for this project was Swahili because this is the most difficult language for a computer to work with, due to its hugely complex system of prefixes and suffixes. This results in almost every verb being unique, so that any automatic alignment has to parse the language at least to some extent. It just happened that a team of Swahili translators were in need of such a tool, so the trial became a required product.

Alignment was achieved by a combination output from the Paratext 'Interlineariser', and the Berkeley Aligner.[7] This, together with a lot of post-processing, produced good alignment on the vast majority of words, but not accurate enough for use by translators (the main target for that project).

Paradoxically, a major source of inaccuracy in this project was the Bible data. Initially it was thought that the Bible data was in a very good state, in that every word was linked to a lexical value that aligned with respectable lexicons, and exact morphology was available. However, the morphology is not very useful for alignment with most translations, because a verb such as "it was created" can be translated with an phrase such as "a created thing". Semantics are much more important, and this is where lexical linkages became problematic.

Words with very similar meanings had completely different lexical values (e.g. *phileō* and *agapē*) while other individual lexical terms could have a multitude of meanings. One difficult word is the Hebrew *paneh* (פָּנֶה) which has a fairly consistent semantic range, meaning "face" which, like in English, can also refer to a 'surface' (as in "the face of the earth"), but idiomatically it has a much wider range of meanings. For example, in Genesis 32.20 (Hebrew 32:21):

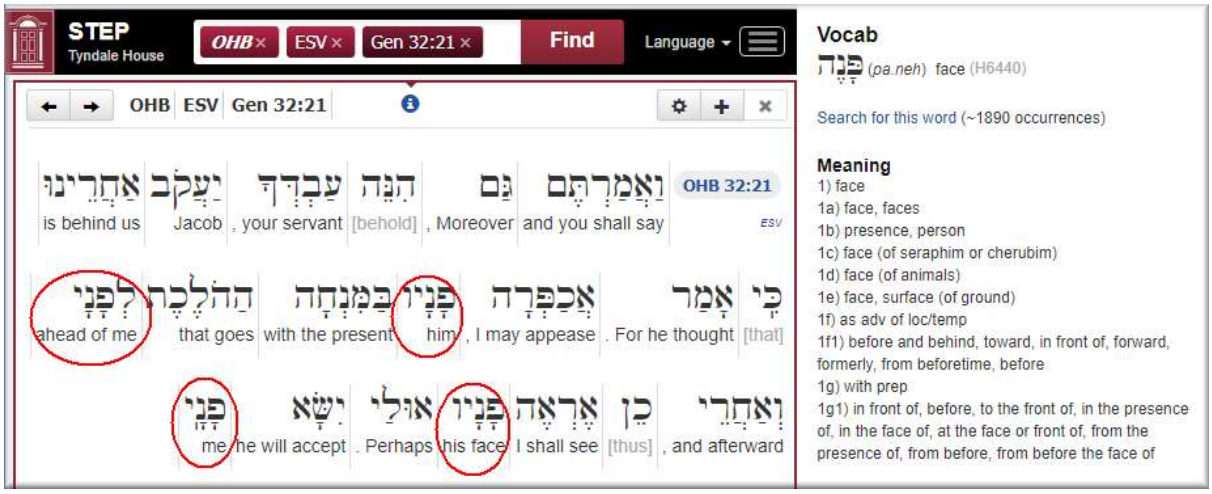

Figure 8. Wide idiomatic range of meaning for Hebrew *paneh*

[7] See a more detailed outline at tinyURL.com/STEPAlign which links to a powerpoint for an overview.

This verse refers to "covering his face" (i.e. appeasing him), going "to his face" (i.e. before him), "seeing his face" (i.e. meeting him) and "lifting my face" (i.e. accepting me). An idiomatic translation such as the NIV may not refer to the word "face" at all, and yet a computer has to somehow figure out how this single word is translated in these four different ways within this verse. Here is what the automatic alignment does with the NIV (without any post-processing):

Figure 9. Automatic computer alignment of *paneh* with NIV text.

This isn't bad, but it isn't good enough – which is why this module is not yet released. It is remarkable how well the computer alignment managed to almost mark the last word of each relevant phrase (which is what it was supposed to do), but it could do better. This was before applying our new data.

Public licence for Tyndale STEPBible data

The new datasets, which are described below, will all be made available on public licence CC BY-NC-ND 4.0 when they reach a suitable level of completion and checking. This licence allows free use by non-commercial projects without any permission, and allows for a negotiated licence for commercial use. There is a no-development restriction, which allows for changes to the format of the data, but any inaccuracies should be reported to Tyndale House so that scholars can access whether or not to implement the change. Updates can then be maintained and distributed centrally so that the datasets remain reliable. The start of this collection can be viewed at Github.[8]

Old Testament Hebrew

The starting point for the Tyndale STEPBible Hebrew text is the excellent Westminster Leningrad Codex freely distributed by the J. Alan Groves Center of Westminster College. This started as a transcription of a printed *Biblia Hebraica Stuttgartensia* (BHS) and was revised towards a facsimile of the Leningrad Codex (Firkovich B19A).[9] The version used is the one curated by OpenScriptures.org. It includes the consonants, vowel pointing and cantillation marks as well as marginal Qere readings – all as found in the Leningrad codex. It also usually adds the accepted Ketiv pointing and divisions demarking suffixes and prefixes. Most importantly, OpenScriptures.org has tagged this text to an augmented version of Strong's numbers which is aligned with BDB. They also usefully added numerical values for levels of conjunction by cantillation which records, in effect, the punctuation of the Leningrad codex.

[8] The start of this collection is at https://github.com/tyndale/STEPBible-Data
[9] See the brief history at https://grovescenter.org/projects/westminster-leningrad-codex.

Tyndale scholars have done a great deal of checking and occasional correction to this text (the results of which are periodically sent to OpenScriptures.org), in particular:

* occasional correction to word divisions and addition of missing Ketiv pointing

* comparison of tagging with other editions with decisions on preferences when they differed. (The most important alternative system is that of SIL. These two systems were rarely wrong, but there were frequent differences of opinion or inconsistencies. Viable options are noted.)

* addition of tags for all prefixes, suffixes and other non-lexical items. These included:
- personal suffixes in three sets: object suffixes found on nouns, and on verbs & prepositions and the subject suffixes.[10]
- prefixes: ב כ ל מ ש (meaning 'in', 'like', 'to', 'from', 'which' etc.)
- the *vav* prefix divided into conjunctive and conversive (i.e. attached to a verb)
- uses of *hé* divided into article, interrogative, directional and paragogic, plus paragogic *nun*
- punctuation markers: *maqqef*, *paseq*, *sof-pasuq*
- paragraph markers: *pe*, *sameq* and inverted *nun*.

* weighting of jussive forms as positive (i.e. the form is different from a normal imperfective), negative (i.e. an imperfective where the jussive would be different) and ambiguous (i.e. an imperfective where the jussive form would be identical, so a morphological decision cannot be made). This and the tags for personal pronouns are necessary to augment the ECTBC morphology.

* differences between BHS, Westminster's transcription, and other electronic Bibles.
The necessity of this data was not anticipated till detailed comparisons of various electronic texts was carried out. Most of these texts were not available for download so comparisons were done on a subset of known problems of transcription. The aim was to find an electronic text that perfectly matched either the BHS or the Leningrad codex, but every dataset examined contained problems. Often these are due to ad-hoc decisions about whether to show the ketiv or qere reading (which are often mixed in the same edition), and sometimes they are simply errors. Some of the test points examined are:

Song.5.16: מַחֲמַדִּים Transcriptions include: מַחֲמַדִּים, מַחֲמַדִּים [11]

Lam.4.3: כְּיָ עֲנִים Transcriptions of this include: כְּיָ עֲנִים, כְּיָ עֲנִים, כְּיָ עֲנִים [12]

Exo.36.21: אַמֹּות Transcriptions of this include: אַמֹּת, אַמֹּת [13]

Figure 10. Contentious transcriptions in the Leningrad B19A facsimile

[10] Subject suffixes are on only a few words such as H3426=יִשׁ, H5750=עוֹד, H2005=הֵן, H0335=אִי.
[11] The dot interpreted as a dagesh is clearly a different colour, so it due to discolouration, not ink. There is another slight discolouration that might account for the vertical stroke of the *metheg*, though this was found in only one transcription so it may be an error.
[12] This is a ketiv where the qere unites them as one word. These transcriptions can therefore be regarded as proposed ketiv pointing.
[13] The circle above the *vav* indicates an alternate spelling without it, so these can be regarded as representing the qere.

* addition of variants and emendations from other MSS, ancient translations and supposition, filtered to include only those that affect the meaning and are considered possible by the UBS Hebrew Old Testament Text Project chaired by D. Bathélemy.[14] Tyndale scholars have added the type of supporting evidence behind these decisions (i.e. Hebrew variants at Qumran and later MSS, ancient translations into Greek, Syriac, Latin, Aramaic or Conjecture).

* addition of word numbers, as divided by spaces and *maqqef*, with corresponding numbers for Qere variants, variant transcriptions and other variants. This, along with book numbers, and three-digit numbering for chapters and verses, makes words and their variants machine-sortable.

* noting Hebrew and English references alongside, with different word numbering on the few occasions when verses commence at different points. For example, the last word of 1 Samuel 20:42 in English versifications is the 5th word of 21:5 in Hebrew versifications, so it is recorded as:

09_1Sa.021.001-05 020.042-27 :/הָ/עִיר H9009#1=ה =art./H5892b#4=עִיר=city\H9016#7=:

The Hebrew numbering for the article is 09_1Sa.021.001-051 and for the word "*'ir"* is …-054. The internal numbering is standardised as #0 for *vav*, #1-3 for prefixes, #4-5 for main words, #6 for suffixes[15] and #7-9 for attached punctuation.[16] This standardisation of the number of digits and internal divisions is designed to make machine searching and comparisons simpler.

Old Testament Greek

The Septuagint in its many forms has a varied and complex textual history. Few transcriptions exist and they are sometimes difficult to tie up with specific manuscripts. Rather than aiming at producing one text with an apparatus, STEPBible Greek OT is based on two texts representing the historical extremes, so that most variants will be included in one or the other text. This is partly a pragmatic decision, based on what is available, but also an historically realistic one, given the long development of this text from the Jewish world into the Christians one. The two texts are:

Rahlfs Edition as approximation of the Old Greek

This text started as a transcription of Rahlfs[17] within the CCAT project[18] and is available in the "LXX" Crosswire module. Ideally a completed Göttingen edition should be used, or even better the text from the Hexapla project, but neither of these are nearing completion. Although Rahlfs based his text on very few manuscripts – mainly Vaticanus with Alexandrinus and Sinaiticus – the resultant work is remarkable, and compares very well with the text of these modern projects which aim to recreate the original text of the Old Greek Bible.

Apostolic Bible for the Ecclesiastical Septuagint

This is an eclectic text created by Charles Van der Pool from the three earliest church printings of the Septuagint.[19] When there were disagreements, his method was to favour whichever text was

[14] Published as *Critique textuelle de l'Ancien Testament*, by D. Barthéley, A. R. Hulst, N. Lohfink, W. D. McHardy, H. P. Rüger and J. A. Sanders (Universitaires Fribourg; Vandenhoeck & Ruprecht, Göttingen.

[15] On two occasions #6 is used for an internally separated word after a *paseq*:
Neh.002.013-17 הֶם/וּ/פְּרוּצֹ֖ים is H1992#4=הֵ֣מָּה=they(masc.)\H9015#5=׀/H6555#6=פָּרַץ=to break through
1Ch.027.012-06 לְ/בָנָ֔/וְ/יְמִינִ֖י is H9005#1=l/H1121a#4=בֵּן=son\H9015#5=׀/H3227b#6=יְמִינִי=Jaminite

[16] On a few occasions all three positions are needed, for example:
Num.010.034-07 ס/ז/:/מַחֲנֶ֖ה/הַ is H9009#1=ה =art./H4264#4=מַחֲנֶה=camp\H9016#7=:\H9019#8=N\H9018#9=S

[17] *Septuaginta*, ed. A. Rahlfs (Stuttgart: Württembergische Bibelanstalt), 1935

[18] The Center for Computer Analysis of Texts directed by Robert Kraft – see http://ccat.sas.upenn.edu/rak/catss.html.

[19] The 1709 Lambert Bos edition which is based on the 1587 Sixtine ed of Vaticanus with lacunae and obvious errors corrected by other MSS, the 1518 Aldine edition by Aldus Manutius, and the 1517 Complutensian Polyglot.

found in two of the three, and if all three disagreed he chose the one closest to the Hebrew.[20] This means that, in effect, he has produced the Greek text with maximum number of changes which were sanctioned by ecclesiastical authorities. These earliest printings were based on manuscripts that had been edited by Christian scribes for centuries and the printed version required an *imprimatur* stamp by the Popes of the day – Leo X (whose expansion of indulgences prompted Luther's dissent) and Sixtus V (who authorised the first printed Vulgate). Also, the methodology of this edition gives preference to majority decisions and also to changes that 'correct' the text towards the Hebrew.

Tagging the Greek Old Testament

The *Apostolic Bible* has been well tagged with an augmented version of Strong's numbers, as described above, by Van der Pool; and the CCAT project added morphological information to its transcription of Rahlfs.[21]

The Tyndale STEPBible Greek OT dataset has merged this data, and will add the missing information for the words that differ between the two texts, so that both texts are tagged morphologically in a way that is compatible with NT morphology, and tagged lexically in a way that is backwardly compatible with Strong's tagging.

A large part of this project was aligning the additional vocabulary of the LXX (which is about double the vocabulary of the NT) to the LSJ lexicon. This alignment means that the NT and LXX can both have access to the same high quality lexicon.

New Testament Greek

The starting point for the STEPBible NT text is the excellent SBLGNT which includes an apparatus of the major variants from the major editions.[22] The choice of variants is very good, in that it includes those which make a significant difference, but the choice of editions is not so useful – they consist of:

* the Majority text of Robinson and Pierpont, which is a good representation of Byzantine MSS
* Tregelles who was the first to take into account newly discovered old MSS such as Siniaticus
* Westcott and Hort's text which was the first of the modern editions using all available MSS
* NIV which is a theoretical text, back-translated from the English, largely conforming to NA

The Tyndale STEPBible data has, in addition, the equivalent information for:

* NA27/UBS4, which is the most recent full edition. Both use an identical text.
* NA28 also known as ECM, which is partly based on computer coherence of MSS[23]
* SBL and Holme's emendations, which are not always evident in the SBL text.
* Tyndale House GNT, which prioritises the earliest MSS and their spelling customs
* Textus Receptus (TR), which is a theoretical text, back-translated from the KJV

The TR is important for historic reasons, in that it behind many older translations. For this edition, a full set of variants are added – not just those selected by Holmes as significant.[24]

---

[20] This methodology is described at https://www.youtube.com/watch?v=KsHaFqSBihM and confirmed to me in private correspondence and interviews.

[21] Robert Kraft gave STEPBible.org permission to use this morphology.

[22] The SBLGNT, edited by Michael W. Holmes is available on a public licence – see http://sblgnt.com. The 'editions' chosen for recording variants is

[23] The results of the Coherence-Based Genealogical Method behind the Editio Critica Maior (ECM) was subject to human evaluation by the same committee that evaluated evidence from the MSS data, so in practice there are very few differences between NA27 and NA28 in the fascicles published so far.

The tagging of the Greek NT is based on the work of many scholars. The version used is that developed and curated by James Tauber,[25] and specifically the version that was attached to the SBLGNT by the Asia Bible Society.[26] The changes in Tyndale STEPBible data consist of:

* a few minor changes to lexical tagging for greater consistency

* corrections to a number of variants in SBLGNT

* increased depth in morphological tagging, with the addition of:
 - identification of Proper Nouns, with P=Person, L=Location, T=Title, G=Gentilic
 - distinguish between Correlative pronouns and Relative pronouns
 - add person numbers to Personal, Reflexive and Possessive pronouns
 - distinguish between Aorist and 2nd Aorist etc.
 - distinguish between Passive, Either middle or passive, and Deponent

* addition of word numbers linked with variants and linked numbers when words have been moved

For example, Hebrews 9.13 either refers to "bulls and goats" or "goats and bulls". Instead of recording this as a variant, it would be ideal to record it as merely a rearrangement of words. Their word-numbers are therefore linked:

59_Heb.009.013-06_10	INSTWH	τράγων
59_Heb.009.013-07_09	INSTW	καὶ
59_Heb.009.013-08	BINRSTWH	ταύρων
59_Heb.009.013-09_07	BR	καὶ
59_Heb.009.013-10_06	BR	τράγων

The words are listed in this order so that the data can be used to create any edition. To create an edition of the Byzantine Majority text, every line is deleted that doesn't list "B" among its editions. When creating a multi-edition source, or when lining up one edition with another, the word-number after an underscore ties up the two editions to produce something like:

εἰ γὰρ τὸ αἷμα τράγων καὶ ταύρων καὶ σποδὸς δαμάλεως ῥαντίζουσα
εἰ γὰρ τὸ αἷμα τράγων καὶ ταύρων καὶ σποδὸς δαμάλεως ῥαντίζουσα

Without the crosslinked numbers, a compared text would appear to have serious variants:

εἰ γὰρ τὸ αἷμα τράγων καὶ ταύρων [ ] καὶ σποδὸς δαμάλεως ῥαντίζουσα
εἰ γὰρ τὸ αἷμα [ ] τράγων καὶ ταύρων καὶ σποδὸς δαμάλεως ῥαντίζουσα

Lexicons

BDB for Old Testament Hebrew

Tyndale STEPBible data is tagged to BDB thanks to the alignment that OpenScriptures created. This work was done very carefully and only a few adustments have been made to increase consistency. OpenScriptures have produced an outline version of BDB, but this is not really usable as a

[24] The 3rd edition of Robert Estienne, known as Stephanus, as found in the 1550 and 1551 printings is generally accepted as the best representation of this genre. The Crosswire module "TR" is based on this, with variants from Scrivener 1894.

[25] The official repository is https://github.com/morphgnt/sblgnt. See allied projects at https://jktauber.com/projects.

[26] The repository at https://github.com/biblicalhumanities/greek-new-testament has the latest version. This was created mainly to experiment with syntax trees – see https://www.ibiblio.org/bgreek/forum/viewtopic.php?f=13&t=1987. Asia Bible Society is now known as Global Bible Initiative.

lexicon.[27] A better abbreviated version of BDB was produced by OnlineBible, and Larry Pierce has given STEPBible.org permission to use this. It was originally created to work with Strong's numbers, and not with the 20% larger vocabulary of BDB. Tyndale scholars adapted this lexicon to work with BDB tagging, but it is a derivative work so it cannot be put out on a public licence.

Instead, the Tyndale STEPBible OT Hebrew lexicion is a full version of BDB, but adapted for easier use. The original BDB is extremely hard-going, partly because of the way the entries are listed, and partly the way they are displayed. The lexicon was organised strictly by the roots, so although a noun may start with a *mem*, it is listed within its three-letter verbal form. This is logical, but difficult, especially when the root is not obvious or certain, so looking up any but the simplest words required a lot of page-turning. The entries were difficult to read because of the amount of bibliographic data included, and the extreme amount of abbreviation required to reduce the physical size of the volume. An electronic edition doesn't suffer these problems, once the initial line-up is made for the complete vocabulary.

In the STEPBible version, the bulk of this data is hidden, and revealed in a hover box. The remaining data is made clearer by making English translations bold, and expanding abbreviations. Previous projects had formated the BDB, with relative success, but the copyright status of these projects proved to be difficult to establish. The STEPBible formatting is a completely fresh approach which represents a considerable investment of time. To visualise the effect of this, see the illustrations for the similar implementation in the LSJ below.

LSJ for Old & New Testament Greek

The LSJ lexicon was released on a public licence by the Perseus project,[28] and has been implemented in various software projects such as BibleWorks, Accordance, Logos etc. as well as sites like Perseus for studying classical literature such as TLG and Philologos.[29] The abridgement known as the "Middle Liddel" is very useful for NT scholars in that it cuts down the number of words and abridges the definitions by cutting out bibliographic details and classical examples. However it misses about 20% of NT words, and about half of LXX words.

The full LSJ is huge, and is popularly referred to as the "Great Scott". It contains about ten times as many words as used in NT and LXX. Many of its entries are also large, because the lexicon is designed to cover all of classical literature. However, this makes it ideal as a lexicon to cover the centuries that span from the LXX to NT, as well as the variety of literature contained in these collections.

The Tyndale STEPBible Greek lexicon is an implementation of the Full LSJ which has:

* Corrected a large number of errors. Extensive checking revealed errors in every implementation of the lexicon. The most accurate version is that use by Logos, and the TLG version at Tufts also had many errors corrected, but probably no version will correct every tiny transcription error in this massive dataset.[30]

---

[27] The repository for this is https://github.com/openscriptures/HebrewLexicon.

[28] The source for this is http://www.perseus.tufts.edu/hopper/opensource/downloads/texts/hopper-texts-GreekRoman.tar.gz. This is actually the Middle Liddel. The full LSJ (often referred to as the Great Scott) is at available http://lsj.translatum.gr/wiki.

[29] The TLG implementation is probably the best, and this is a free portion of the TLG site – see http://stephanus.tlg.uci.edu/lsj. Philologos, now known as Logeion is often faster though less helpful – see http://logeion.uchicago.edu.

[30] For example, even the well-corrected Logos version dates Pseudo-Plutarch's work on the life of Homer (written by a pretender in the 3rd or 4th C AD) as if it was written by the 1st C Plutarch.

* Added dates for all the classic sources. The hundreds of classical authors cited in LSJ makes this lexicon a valuable tool for charting the chronological development of the language, especially as they tend to be cited in chronological order. However, for most users, this is opaque because no dates are given. Therefore dates were added for every author, so that the user can see easily when a particular definition came into use.

* Hidden examples of Greek use. Although the examples from classical literature are so useful, they are also distracting. Therefore they are hidden as a hover box which is linked to the first date in the sources cited. For example the entry may contain "[Refs 5th C BC+]" which indicates that there are examples of a particular usage from the 5th century onward, which are revealed in full when a mouse hovers over the date.

* Abbreviations are expanded. LSJ contains some of the most arcane abbreviations, and sometimes even the expanded version is obscure. For example the abbreviation "A. Ag." has been expanded to "4th-5th c.BC: Aeschylus Tragicus 'Agamemnon'", "dat. pers. et rei" is expanded as "dative of persons and of things" and "sc." has been replaced by "i.e.".

For example, compare the entry for πίστις in the printed and Tyndale STEPBible versions. Hiding distracting information in the latter is illustrated by hovering over a date to reveal the details it refers to.

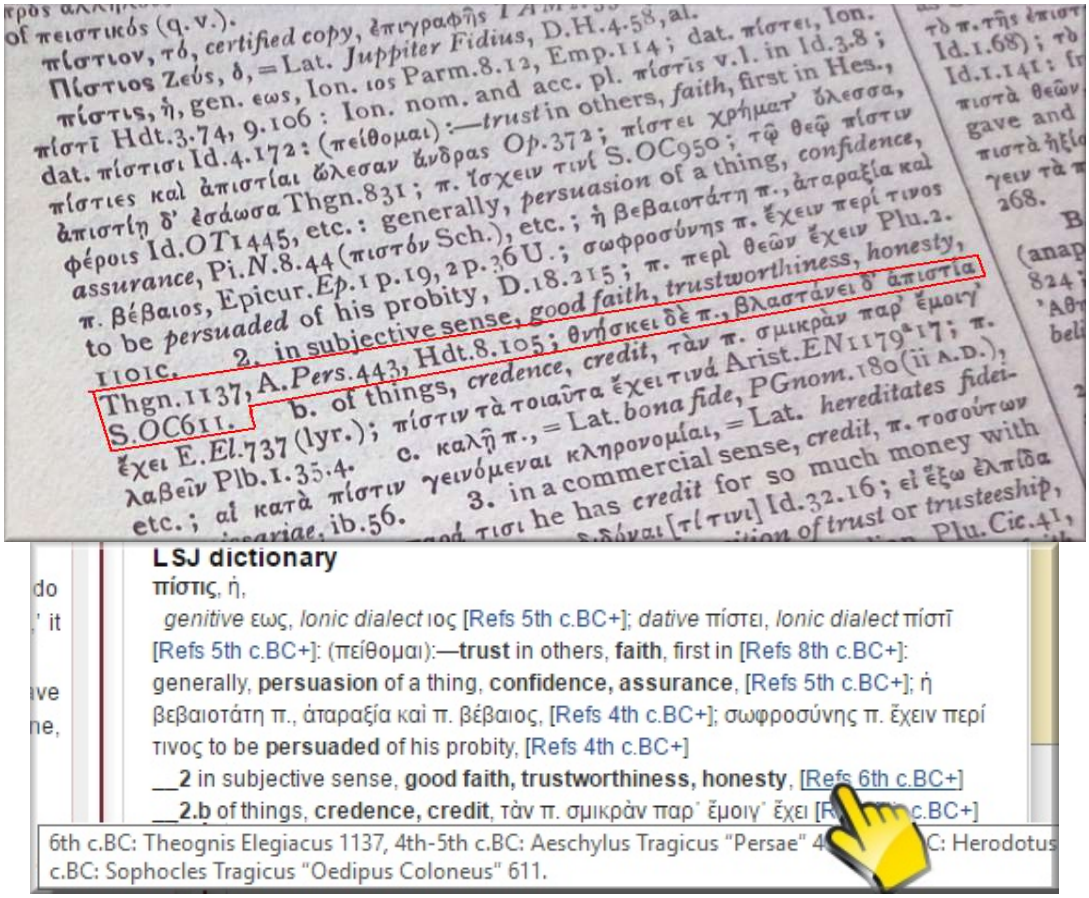

Figure 11. Comparison of the printed LSJ with the Tyndale STEPBible implementation.

Versification

Chapter and verse breaks were added relatively late in the history of the Bible. The Hebrew Bible verse divisions were added by 10th century Masoretes, based on established tradition, but verse numbers were first added by R. Isaac Nathan in 1448 for use in his concordance. A Vulgate in 1528 was the first whole Bible to have verse numbering throughout, but the Apocrypha and NT had long verses – two or three times the length used today. Robert Estienne (aka Stephens or Stephanus) first added the established verse numbers to the NT when he produced a copy of his Greek text in 1551 printed alongside the Latin Vulgate and Erasmus' Latin back-translation from the Greek. The purpose was, presumably, to facilitate the lineup between the different versions, though he was also preparing a concordance (completed by Henry Stephens in 1594). The first full Bible using modern versification was Robert Stephens' 1557 Vulgate, along with a Latin concordance. So both our Hebrew and Greek versification was prompted by the need for data division, which is similar to the modern needs.

However, this apparently neat situation was complicated by many slight variations in versification. At several points in the OT, different Bibles start a new chapter at a different verse, or divide verses differently to create one fewer or more verses, which causes renumbering of subsequent verses. Psalms 9 and 10 are merged in some Bibles, so that every subsequent Psalm is numbered differently, though in these Bibles Psalm 147 is split into two, so that all Bibles end up with 150 Psalms. Except, of course, some Bibles add Psalm 151 which is preserved at Qumran, and many Bibles add Apocryphal books, some of which, in some Bibles, are added as final chapters to Daniel and Esther while in others they are inserted as subverses or interpolated sections within the existing chapters.

The New Testament appears less complex, though not when one looks closely. Although there is virtual unanimity about verse numbering, there are many instances where the point of verse division differs by a few words. This is perhaps because the original divisions were done by Stephanus while travelling from Paris to Lyons on horseback. His lack of concentration, or perhaps the jolting of his pen, resulted in unfortunate decisions that were corrected inconsistently by subsequent editors in just over 100 verses.[31] Most modern Bibles follow the revisions made by Beza in 1565.

Fortunately, the differences in the Old Testament can be narrowed down to three streams: Hebrew, Greek and Latin. Translations tend to follow the versification in their source text, though once this is established in a particular language, subsequent translations from any source tend to follow that versification. As a result, most Eastern European Bibles follow the Greek tradition while Western European Bibles follow the Latin tradition, but modern translations (which often depend more on the Hebrew) tend to keep some or all of the versification of earlier Bibles. For example, the Russian Synodal Bible (1813-76) was proudly based on the Hebrew text instead of the Septuagint used by previous Russian Orthodox Bibles, and yet it kept the traditional Greek versification including two places where Slavonic numbering is unique.[32]

In practice most modern Bibles are an amalgam of different traditions, and even different editions of the same translation can contain different versification.[33] Perhaps the most complex amalgam is that found in most English Bibles which is equally based on all three traditions but which differs

[31] For more information on different editions, and a full list of NT verse differences, with editions that follow them, see Ezra Abbot, *The Authorship of the Fourth Gospel and other Critical Essays*, ch.20: "The verse divisions in the New Testament" (Geo H. Ellis, Cambridge MA, 1888) p.464-477, available at https://archive.org/stream/authorshipoffour00abbo#page/464/mode/2up.

[32] Lev.14.55-56 are merged as one verse and Josh.6 starts one verse earlier so all the verse numbers are different.

[33] For example, the Spanish Reina-Valera continued to partly follow the Vulgate till 1909 but the 1960 revision mostly follows the English style.

from all three in 1123 verses.[34] This English versification is, for good or ill, the de-facto standard because almost all English Bibles follow it, and therefore a large majority of commentaries because they are mostly in English or they wish to reach a wider readership.

Crosswire has attempted to solve this problem by producing versification systems to represent the majority of Bibles, and STEPBible.org has incorporated these to help align Bibles from different traditions. However, this frequently fails to work properly because many Bibles fail to conform to one system completely, and because of the complexity of making different versions align. Therefore a dataset was created with tests for all the sections where variation occurs so that each Bible could be individually marked up in accordance with its own specific mix of versification. This required new data.

Tyndale STEPBible Versification data is based on the English versification – specifically the NRSV (because even English versions contain a few differences) – for OT, NT and Apocrypha. This was a pragmatic decision because so many tools, commentaries, and an increasing number of Bibles are based on this system of versification. Every one of the hundreds of sections where variation occurs is given a rule by which the versification being followed in any specific Bible can be determined. Notes are added to the Bible text to inform the reader of the difference in that specific Bible and how that relates to the traditional versification systems in Hebrew, Greek and Latin Bibles.

One complex example is Psalm 13. Here are the markup and notes seen in the 1917 version of the JPS (Jewish Publication Society) which follows Hebrew versification.

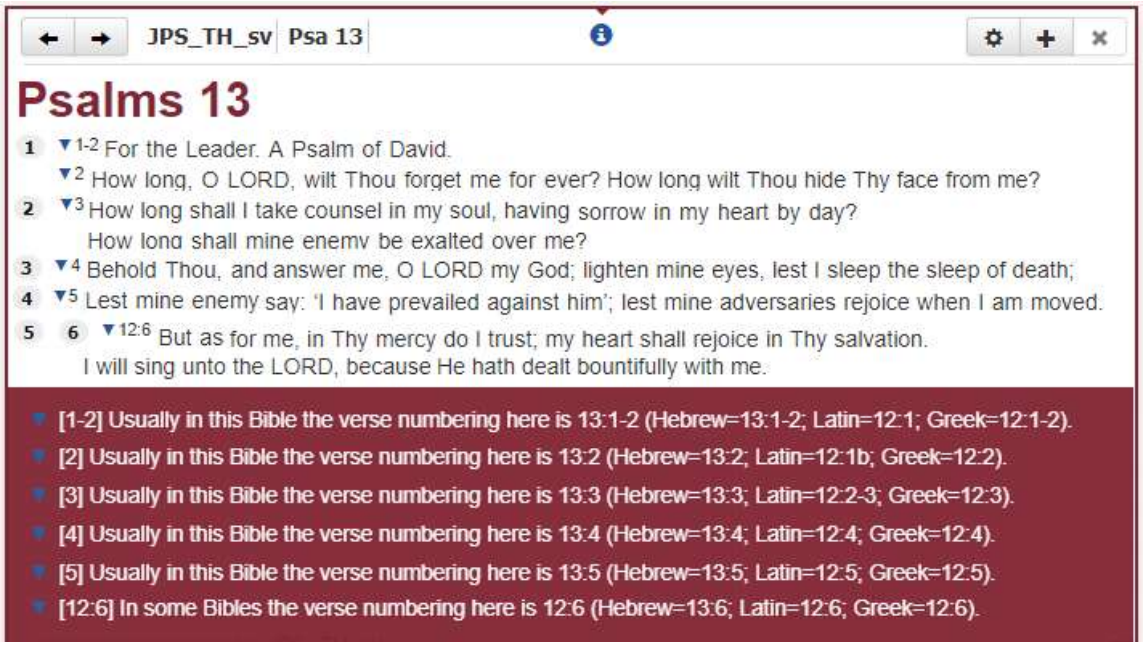

Figure 12. Versification notes and markup for JPS in a complex example.

[34] Most of these (969) are verses in the Psalms that have titles, which in other Bibles are in separate verses, so all subsequent verses are renumbered. Other places where the English OT+Apocrypha deviates from Hebrew, Greek and Latin versification are: Num.29.40—30.16; 1Sam.20.42; 23.29—24.22; 1Ki.22.43-53; Job.41:1-34; Eccl.5.1-20; Dan.4.1-3, 37; Hos.13.16—14.9; Jonah 1.17—2.10; Bar.6.50-73. Usually the English Bible agrees with one or more ancient version. Counting instances instead of verses: it agrees with Hebrew 400x deviating from it 678x; with Greek 424x deviating from it 740x and Latin 573x deviating from it 596x.

Each individual note normally appears only when a user hovers over a note marker (▼). They reveal, to the interested user, that this is Psalm 12 in Greek and Latin traditions – because these Bibles display Psalm 9-10 as a single Psalm. In Hebrew and Latin the first English verse is split into two verses, while Greek splits verse 2 into two verses and English splits the last verse into two – so they all end up with the same number of verses, but with frequent differences. In each version, the markup and notes appear differently, though with consistent features:

* the English verse numbers are shown as the main verse divisions

* the alternate verse numbers are seen in smaller type

* the note indicates whether the alternate numbering occurs "in this Bible" (i.e. the version being read) or in other Bibles.

* the verse numbering is given for the three traditional streams: Hebrew, Greek, and Latin

The downside of this markup, for someone who is used to a different versification system, is that the verse numbers will not always match those that they are used to, though the traditional numbers are still visible in smaller type. The benefit is that their Bible can be easily compared with Bibles that have any other mixture of versification. And when they use a commentary (which is likely to be based on English versification) they can easily identify the verse being discussed or referred to even though their own Bible traditionally uses a different versification. Also, if they quote a verse, they can know whether it would be helpful to cite an alternate verse number for users of other Bibles.

The data for Psalm 13 is:

Rules	English:	Psa.13:6=last & Psa.13:1>13:3
	Hebrew:	Psa.13:6=last & Psa.13:1<13:3
	Greek:	Psa.13:7=last & Psa.12:3<12:4
	Latin	Psa.13:7=last & Psa.12:3>12:4

	English	Hebrew	Greek	Latin
Concatenation	Psa.13:1	Psa.13:1-2	Psa.12:1	Psa.12:1-2
Concatenation	Psa.13:2	Psa.13:3	Psa.12:2-3	Psa.12:3
OneToOne	Psa.13:3-4	Psa.13:4-5	Psa.12:4-5	Psa.12:4-5
DuplicateTarget	Psa.13:5-6	Psa.13:6	Psa.12:6	Psa.12:6

These rules are independent of any information about the Bible – its historical tradition or the words in the translation. It relies only on the number of verses present and the relative length of those verses. This means that it can be applied automatically to any Bible. Wording of notes is standardised for easy translation. Labels such as "Concatenation" indicates the type of changes involved, which informs the program what to do with the text and which form of wording to use.

Meaning tags

Sub-meaning tags

One of the problems identified above is that Hebrew and Greek words can have more than one meaning, either due to a wide semantic range or due to idiomatic use. Biblical Hebrew and Greek have vocabularies of about 8000 and 10000 words respectively, of which about half are rarely used. This compares poorly with the million words now counted as English, and an average adult vocabulary of 42,000 words.[35] The shortfall was made up not by less sophisticated communication but by using the same word to mean several things. In order for computational alignment to have any chance, these uses need to be disambiguated. For example:

בֵּן (*ben* - H1121a) §1 child/son §2 descendant §3 people/men/nation/tribe §4 daughter §5 warrior/son(strong) §6 age[son of](YEARS) §7 [inheriting]son/heir §8 [son of](DESCRIPTOR) §9 calf/lamb/young(ANIMAL) §10 rebel/son(Beli

יָד (*yad* - H3027) §1 hand/arm[anatomy] §2 power/rule/ownership §3 themselves/myself/yourself/[hand of](SUBJECT) §4 by/though[hand of](AGENT) §5 to/with/from/for/beside[hand of](IND.OBJ.) §6 times/parts/(number)[hands] §7 monument/place/station/sign §8 penis §9 tool §10 bank/border §11 spacious/(wide)hand §12 donate/ordain/(fill)hand §13 vow/swear/(raise)hand §14 expend/(reach)hand §15 undertake/(outstretch)hand §16 certainly/[hand to hand] §17 owner/rule/(under)hand §18 swear/allegience/hand(under)

יוֹם (*yom* - H3117) §1 day §2 when/time/(period/life)[of days] §3 year[of days] §4 daily/regularly §5 always/long time §6 today/now §7 old/(full/many)[days]

הָלַךְ (*halakh* - H1980) §1 walk/move §2 come[nearer] §3 went/go[away] §4 take/lead(OBJECT) §5 come[hortative] §6 continue/will be[future] §7 journey §8 follow/behind/go[after §9 send/drive(OBJECT)

יָצָא (*yatsa* - H3318) §1 come/go out/escape §2 send/take out/release §3 extends/ends §4 casting(lot) §5 (sun)rise §6 issue/flow/sprouting/leaping(water/fire/OBJECTS) §7 produce/birth/sprout(LIVING THINGS) §8 regular/accostomed/able(fighter) §9 speak/pronounce §10 surrender/failed(heart)

נֶפֶשׁ (*nephesh* - H5315) §1 soul/heart §2 life/breath §3 myself/yourself/himself §4 person §5 animal §6 appetite/wish §7 dead §8 neck

ψυχή (*psuchē* - G5590) §1 soul/heart §2 life §3 myself/yourself/himself §4 person §5 animal

These examples illustrate the complex nature of the problem. Sometimes the semantic range mirrors that of English, such as the idiom "under the hand" meaning to rule, or a "day" meaning a period such as "the day of judgement". However, this is often due to the influence of the Bible on the language, and this shared range cannot be expected to occur in all languages. Sometimes two words may share part of their range, such as *halakh* and *yatsa* which can both mean "to come" and although one has the major sense of "come to" and the other of "come out", in many languages they

[35] Marc Brysbaert, Michaël Stevens, Paweł Mandera and Emmanuel Keuleers, "How Many Words Do We Know? Practical Estimates of Vocabulary Size Dependent on Word Definition, the Degree of Language Input and the Participant's Age" (Frontiers of Psychology, 29 July 2016 at https://doi.org/10.3389/fpsyg.2016.01116.

will be translated the same in similar contexts. Sometimes words in Hebrew and Greek share a similar semantic range, such as *nephesh* and *psuchē* which can both mean 'soul', 'breath', 'person' etc. This confluence is no doubt due to the influence of the OT on NT Greek. The most important conclusion from these findings is that semantic ranges are not just wide, but disparate – these words need to be regarded as having several separate meanings.

A human can work out which meaning applies where, but creating rules to decide this is extremely difficult, and even Bible translators are sometimes in disagreement. The Watson box (an IBM computer [36]) can now discern this kind of ambiguity in English, though at the expense of a huge amount of training, and in a language which is much less ambiguous. For alignment in a wider range of languages, the words have to be tagged with their meaning according the context at every occurrence.

The Tyndale STEPBible OT and NT data is now tagged in this way. For example, the verb *halakh* occurs with three different meanings in Isaiah 2.3:

> And many peoples shall come (§2), and say: "Come(§5), let us go up to the mountain of the LORD... that we may walk(§1) in his paths."

In English these meanings all appear to be related, but the concepts are very different. Movement is implied in §2 and §1, but one implies approach while the other implies travel along – concepts that are often expressed in very different words, as they are in this translation. The hortatory "Come" of §5 is translated with the same English word as §2, but this will not occur in every language.

Because these ambiguous words tend to be among the most widely used words, this entailed tagging 22% of the words in the OT with these sub-meanings, though only 3.3% of the NT needed the same tagging.

Super-meaning tags

Another problem for computer analysis of Biblical Hebrew, and to a lesser extent Greek, is the number of rare words. The more times a word occurs, the more chance there is to work out how to align it with a translation, but if a word occurs only once (as about 1,500 words do), a computer has very little chance of aligning it correctly.

However, it is possible to link words that have similar meanings using super-meaning tags. These can be applied across language barriers, because as far as a computer is concerned, a Hebrew word meaning "to cry" is little different from a Greek word with the same meaning, and identical linkage means it has more instances to compare with the translation language. So, for example, the following can all have the same super-meaning tag:

בָּכָה (*bakhah*) 'to weep' - 99x	κλαίω (*klaiō*) 'to weep' - 34x
בְּכִית (*bekhit*) 'weeping' - 1x	ἀποκλαίω () 'to weep' LXXx3
בְּכִי (*bekhi*) 'weeping' - 29x	κλαυθμός (*klauthmos*) 'weeping' - 9x
בֶּכֶה (*bekheh*) 'weep bitterly' - 1x	δακρύω (*dakruō*) 'to weep' - 1x
דִּמְעָה (*dimah*) 'tears / weeping' - 22x	δάκρυ / δάκρυον (*dakru*) 'teardrop' - 11x
דָּמַע (*dama*) 'to weep' – 1x	

[36] The Watson box was initially designed to defeat human competitors in a general knowledge quiz, which required it to understand the ambigous and idiomatic communication of real English. See https://www.aaai.org/Magazine/Watson/watson.php. It can now read, understand, and evaluate the likely accuracy and emotive content of a wide range of English communications, from techincal science journals to tweets (see e.g. https://www.ibm.com/watson/services/tone-analyzer).

These can all have the super-meaning tag of "weep" because although they have different grammatical forms, they can all be translated by a related group of terms. For example, the ESV (which aims for greater consistency than most Bibles) translates *bakhah* usually as "wept/weep/weeping" etc. but also as "crying" and "lamentation", while the nouns *dimah* and *dakruon* are usually translated as "tear" but also as "weeping" or "cry". Linking these words helps computer alignment not only with words that occur rarely, but also with the less common translations of common words.

The usefulness of these tags can be envisaged by considering the short verse John 11.35: "Jesus wept". The common name Jesus will be easy to identify, so the computer can also identify the alignment of the other word. One might assume that this would 'teach' the computer the word meaning 'to weep' in that translation. However, this won't happen, because the Greek verb δακρύω ('to weep') occurs only here, and is not even used by the LXX. But when super-meaning tags are used, this alignment can inform all the other places where the same translation occurs with a word tagged with the super-meaning "weep".

Sometimes the super-meaning is linked with only some of the sub-meaning tags. For example, the following can have the super-meaning tag of "man/person":

אָדָם (*'adam*) 'man' - 526x	ἄνθρωπος (*andrōpos*) 'a human' - 556x
אִישׁ (*'ish*) 'man'§1_man – 1075x	ἀνήρ (*anēr*) 'man'§1_man – 176x
אִישׁ (*'ish*) 'man' §3_anyone/someone – 546x	ἀνθρώπινος (*anthrōpinos*) 'human' - 7x
אשׁשׁ (*'ashash*) 'be manly' - 1x	ἀνδρίζω (*andrizō*) 'be manly' – 1x LXXx22
אֱנוֹשׁ / אֱנָשׁ (*'enish* Heb/Aram) 'human' - 59x	ἀνδρειόω (*andreiō*) 'become a man' – LXXx1
גֶּבֶר / גְּבַר (*geber* Heb/Aram) 'strong man' - 84x	ἀνδρόω (*adroō*) 'become a man – LXXx2
מַת (*mat*) 'man' - 20x	ἐπανδρόω (*epandroō*) 'make manly' – 0x

Both *'ish* and *anēr* are sometimes tagged with a sub-meaning §2 "husband" (69 and 53 times respectively). Most languages now distinguish these concepts so the translation is likely to be different in texts where the context implies this different sub-meaning. These instances are therefore not included in the super-meaning tag for "man/person".

Proper Nouns

A significant cause of mis-alignment is the identification of proper names, because Greek and Hebrew commonly contain multiple spellings or names for the same person, but translations standardise these spellings. And sometimes the same name in Hebrew or Greek refers to two different people whom translations refer to with two different spellings. The best-known example is the Greek name Ἰησοῦς which is translated "Jesus" and "Joshua" (Act.7.45 etc.) or "Justus" (Col.4.11). The reason for the latter is that some MSS have the name Ἰοῦστος at Colossians 4.11. However, where this same name occurs at Act.1.23, some MSS have Ἰωσήφ so some translations call him 'Joseph'. All these names are common: 'Joseph' is used for 8 individuals, 'Justus' for 3 and 'Jesus' for 3 individuals.

The Tyndale STEPBible Proper Names dataset identifies individuals by the most commonly translated form of their name, along with the first reference where that individual occurs. If the form is not the same as the common form, this is given too. So individuals above are identified as:

Ἰησοῦς:	Jesus@Mat.1.1	Joshua@Exo.17.9	Jesus|Justus@Col.4.11
Ἰοῦστος:	Justus@Act.18.7	Justus@Col.4.11	Justus|Joseph@Act.1.23

Ἰωσήφ:	Josech@Luk.3.26	Joseph@Act.1.23	Joseph@Gen.30.24	Joseph@Luk.3.24
	Joseph@Luk.3.24	Joseph@Luk.3.30	Joseph@Mat.1.16	Joseph@Mat.27.57

This project has identified 4248 different proper names and linked them to 4275 unique individuals, locations, and titles. However, 800 of these names occur in more than one form in Greek and/or Hebrew, 1400 of these forms are used for more than one person or place, and 208 of these forms are regularly translated in more than one way. All this is confusing for human readers, let alone computers.

For each individual, every occurrence of each form of their name was identified and listed, and then the unique identifier was tagged to that name in the Hebrew and Greek texts. This means computer alignment benefits from being aware of both the form used locally and the individual it represents which is likely to be translated in a uniform way in most Bibles.

This project was massively aided by the prior work done by www.complete-bible-genealogy.com for people and https://www.openbible.info/geo/atlas. The former is based on the KJV whereas the Tyndale STEPBible data is based on ESV so the data is not really compatible, but it provided an invaluable way of checking though also a frequent basis for disagreement. That data, like the STEPBible data, is based on the philosophy of 'medium scepticism' – that is, people with the same name in different passages are assumed to be different individuals unless there are reasonable details with which to link them. Nevertheless there were frequent areas of disagreement. There was less disagreement with regard to the openbible.info data because although precise locations are often difficult to identify, there are fewer instances of confusion.

Translation issues

The number of English Bible translations[37] is evidence of the variety of ways in which the text can be translated. However, most of this variety is due to different styles of translation rather than different meanings.

The Tyndale STEPBible Meanings & Manuscripts dataset aims to pinpoint all the places where the underlying Greek and Hebrew can be translated with a different meaning. This is marked up as an alternate translation of the ESV text, which was chosen because it is based on modern text-critical editions, while aiming to translate in a relatively word-by-word manner. This means that when a single word can be translated with more than one meaning, it is usually possible to express that meaning by changing just one or two words in the text.

For example, 1Sam.5.8-9 in ESV:

> So they sent and gathered together all the lords of the Philistines and said, `What shall we do with the ark of the God of Israel?` They answered, `Let the ark of the God of Israel be brought around to Gath.` So they brought the ark of the God of Israel there
> (9)But after they had brought it around, the hand of the LORD was against the city, causing a very great panic, and he afflicted the men of the city, both young and old, so that tumors broke out on them.

This is marked with different possible meaning at points underlined. When a user hovers over them, they will see:

[37] I do not know of a definitive count of English translations, but Tyndale House library in Cambridge houses 113 different English translations of the complete Bible or New Testament.

* "They answered" – Means: "The people of Gath
* "Gath" (evidence includes the Hebrew manuscripts)
  – Possibly: "us" (evidence includes some Greek manuscripts)
* "around" (evidence includes most Hebrew manuscripts)
  – Possibly: "to Gath" (evidence includes manuscripts from Qumran)
* "panic" – Or: "discomfort"
* "tumours" (evidence includes the Hebrew manuscripts)
  – Possibly: "haemorrhoids" (evidence includes Hebrew scribal notes)
  – KJV: "emerods"

The alternate Meanings that are recorded is determined by ten standard translations.[38] Occasionally Tyndale scholars highlighted possibilities that are not present in any of the standard translations, and these have been treated with extra caution before being accepted. It is often difficult to decide if a different translation conveys a different meaning, or merely states the same meaning with a difference nuance. The word "panic" and "discomfort" by no means covers all the translations available, but they represent the two disparate ideas of fright and confusion.

Occasionally an interpretive meaning is added as "Means:" when this is significant and it is indicated by the standard translations. Archaic words in the KJV are recorded because even if they have the same meaning, they can be linked to the equivalent modern English.

The Manuscript variants that are recorded is determined by the UBS Hebrew Old Testament Text Project chaired by D. Bathélemy and the variants noted in USB 3rd & 4th editions of the NT. In both of these works, the principle for selection was any variant that possibly changed the meaning of the text. The decisions of this committee also determine whether a change is marked as "possible" or "probable". The latter is rare because both this UBS committee and the ESV committee have similarly conservative commitments to the Hebrew text.

Manuscript evidence for the NT is exhaustive up to AD 500,[39] and for the OT it is recorded as evidence from "Greek" (i.e. LXX), "Latin" (i.e. Old Latin & Vulgate), "Samaritan" (i.e. the Pentateuch), "Aramaic" (i.e. Targums), "Scribes" (i.e. Masora and Talmudic traditions), "Qumran", "Geniza", "some Hebrew" (i.e. Hebrew MSS) and "most Hebrew" (i.e. the Massoretic tradition). This simplified language was deemed to be less difficult for non-scholars while scholars would have no problem understanding what it referred to. There is no attempt to list manuscripts, so scholars are expected to go to other sources to look up the details. Much more information is available online.[40]

At STEPBible.org these will be presented as translation options that can be clicked on and be inserted into the text so that users can experiment and see how the sense of the passage is affected. They will discover that these changes usually have minimal affect on the overall message and the Bible certainly can't be made to say 'whatever you want' as is popularly thought.

This data will also be useful for translators as an indication of problems that need deeper investigation. This data will be integrated into the Hebrew and Greek texts. For the NT, this will tie up with the variants recorded for all the MSS up to AD 500. For the OT this will be the first time that the Hebrew alternatives based on the ancient versions will be presented alongside the Hebrew text.

[38] ESV, KJV, NASB, NRSV, NIV, NET, NJB, NLT, JPST, and CEB. These span traditional, scholarly, protestant, Catholic, conservative, and what might be called 'mildly adventurous' styles.
[39] Codex D Bezae is included even though it may be a little more recent, because of its importance.
[40] https://stepweb.atlassian.net/wiki/spaces/SUG/pages/7766018/Manuscripts+and+Meanings.

Conclusion

Datasets define the modern world. Major and growing datasets include historical trending of news,[41] social and health statistics,[42] the historical use of words and phrases in literature,[43] publicly funded science,[44] health,[45] and even songs.[46] The Bible was among the first literature to be computerised and analysed, and now that this data can be used in more creative ways, it needs to be cleaned and structured.

It is hoped that these Tyndale STEPBible datasets will add to this trend and provide tools for more advanced studies.

[41] Google Trends at https://trends.google.com/trends/
[42] Nations have different levels of transparency. The USA has set a high standard at http://data.gov and http://www.census.gov/data.html. UK stats are good though with some unfortunate delays, at http://data.gov.uk.
[43] The Ngram dataset includes all English publications from 1800, at https://books.google.com/ngrams
[44] CERN stands out for quantity – more than a petabyte of data at http://opendata.cern.ch.
[45] US at https://www.healthdata.gov, UK at http://www.hscic.gov.uk/home, WHO at http://www.who.int/gho/en.
[46] Amazon Million songs at http://aws.amazon.com/datasets/6468931156960467.