A Tool for a High-Carat Gold-Standard Word Alignment

Drayton C. Benner
Near Eastern Languages & Civilizations Department
University of Chicago
Chicago, IL USA
drayton@uchicago.edu

Abstract

In this paper, we describe a tool designed to produce a gold-standard word alignment between a text and its translation with a novel visualization. In addition, the tool is designed to aid the aligners in producing an alignment at a high level of quality and consistency. This tool is presently being used to align the Hebrew Bible with an English translation of it.

1 Introduction and Background

Gold-standard word alignments have been produced for a variety of purposes, but the machine translation community has been the most interested in aligned texts. For this community, aligning texts is not an end in and of itself. Rather, gold-standard aligned texts have served to train and also evaluate machine translation algorithms or their components, especially automatic alignment algorithms. However, there are other scholarly endeavors in which gold-standard word alignments are useful in and of themselves. Within linguistics they are certainly helpful to the subfields of contact linguistics, corpus linguistics, and historical linguistics, but they are also useful in humanistic inquiry more broadly, especially in in studies of translation technique, textual criticism, philology, and lexicography. In addition, presenting gold-standard aligned texts can make texts more accessible to a broader audience, especially to an audience that has limited skill in either the source or target language.

A gold-standard alignment that is designed to aid the humanist is likely to have different requirements with regard to quality, consistency, and visualization than a gold-standard alignment designed as an input to a machine translation algorithm. Results from research into the effect of the quality of alignments above a certain level on machine translation quality has been mixed (Fraser and Marcu, 2007; Fossum et al., 2008; Lambert et al., 2012). Thus, the extra cost of making a good alignment excellent might outweigh its benefits if its only purpose is to aid in machine translation. Put differently, a 14 carat gold-standard alignment may be sufficient for the purposes of machine translation. However, for the humanistic endeavors enumerated above, incremental improvements in quality continue to be useful to scholars; a 24 carat gold-standard alignment is highly desirable. Similarly, consistency is important for many of these humanistic endeavors. For example, a scholar researching the way in which a particular word or class of words is translated needs the alignment to be done consistently across the translated corpus. Finally, when the translation and alignment themselves are an object of study, the alignment needs to be presented visually in an appealing manner, and the researcher needs to be able to access additional information easily.

2 Alignment Project and Tool

Achieving a high level of quality and consistency requires a software tool designed to facilitate this, and the visualization techniques for this software tool can be similar to the visualization of the final alignment. In what follows, we present a manual alignment tool that has been built as a Java application for desktop operating systems in order to achieve these goals for an ongoing project to align the Hebrew Bible with an English translation of it. For the Hebrew Bible, we use the *Westminster Leningrad Codex* (*WLC*) and

Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH) @ EACL 2014, pages 80–85,
Gothenburg, Sweden, April 26 2014.

Westminster Hebrew Morphology (*WHM*), both version 4.18. *WLC* is a diplomatic edition of *Codex Leningradensis*, the oldest complete manuscript of the Hebrew Bible in the Tiberian tradition. *WHM* tokenizes the text and provides a lemma and morphology codes for each token. *WLC* and *WHM* are presently maintained by the *J. Alan Groves Center for Advanced Biblical Research*. For an English translation, we use the *English Standard Version*, 2011 text edition. Its tokenization is straightforward and was done at the word level.

While various groups have aligned the Hebrew Bible with various English translations, beginning with (Strong, 1890), and even to the Greek Septuagint translation (Tov, 1986), this project is unparalleled in its focus on quality and consistency in the alignment, and the alignment tool reflects that. The Alignment Panel provides the primary visualization of the alignment and allows for its manipulation while several other panels provide data to aid the aligner with regard to quality and consistency. The aligners follow a lengthy document outlining consistency standards.

2.1 Alignment Panel

Several types of visualizations have typically been used to display aligned texts. Most commonly, lines have been used to show links between aligned tokens (Melamed, 1998; Daume III; Smith and Jahr, 2000; Madnani and Hwa, 2004; Grimes et al., 2010; Hung-Ngo and Winiwarter, 2012). While this is helpful, the lines become difficult to follow when the word order differs significantly between the source text and its translation or even if one text requires significantly more tokens than the other. The second common approach uses an alignment matrix (Tiedemann, 2006; Germann, 2007; Germann, 2008). Again, this is a helpful visualization technique, but it takes time for the user to see which source tokens link to which target tokens at a glance, and it is easy to accidentally move over a row or column with one's eye. A third approach involves coloring linked words using distinguishable colors (Merkel, 2003; Ahrenberg et al., 2002; Ahrenberg et al., 2003). When used by itself, this is helpful but slow for the eye to find which source token links to which target token. A fourth approach requires the user to place the mouse over a particular token of interest to see links for just that token (Germann, 2007; Germann, 2008). This removes the clutter but is cumbersome for a user trying to see the entirety of the alignment.

The approach taken here, shown in Figure 1, combines the first and third of these visualization techniques but modifies them in order to make the alignment easier to read and to enable the aligner to align quickly while maintaining high quality. In addition, the Alignment Panel includes language helps to speed up the human aligner. Tokens are displayed vertically. While previous alignment tools have more conventionally displayed the tokens horizontally, whether as a flowing text or as separated tokens, Hebrew is written right-to-left, while English is written left-to-right, so a vertical display, as done by (Grimes et al., 2010) for an Arabic-English alignment, makes more sense: both languages can be read top to bottom. The Hebrew tokens are grouped by the human aligner into token sets, and these token sets form a partition over all the Hebrew tokens. The same is true for the English tokens. Hebrew token sets can then be aligned with English token sets. In addition, in token sets with two or more tokens, the human aligner can optionally declare precisely one token in the token set to have primary status if it is most basic to the token set on a semantic level. For example, in Figure 1, the Hebrew word רשׁעים ("wicked") is linked to an English token set consisting of two tokens: *the* and *wicked*. The aligner has correctly identified *wicked* as the primary token in this English token set. In token sets containing just one token, the one token always has primary status.

Alignment visualizations using lines can be difficult to process if the word order differs sharply between the source text and its translation. In order to combat this issue, a key innovation of this tool is that blank rows are inserted at times on both the source and target sides. The blank rows are inserted in such a way that the number of straight, horizontal lines linking source token sets to target token sets is maximized. That is, the maximum possible number of aligned token sets are aligned horizontally. Subject to this constraint, blank rows are inserted so as to minimize the sum of the length of the vertical components of the lines, including both the lines joining multiple tokens into a token set and the lines indicating links between source and target token sets. When the user changes the alignment, which is done primarily using drag-and-drop, the tool immediately recalculates the optimal blank rows and redraws if necessary, all the while remaining responsive. While multiple formats are supported for exporting the alignment data, all of the data is imported into memory during application startup. This requires more updating of complex internal

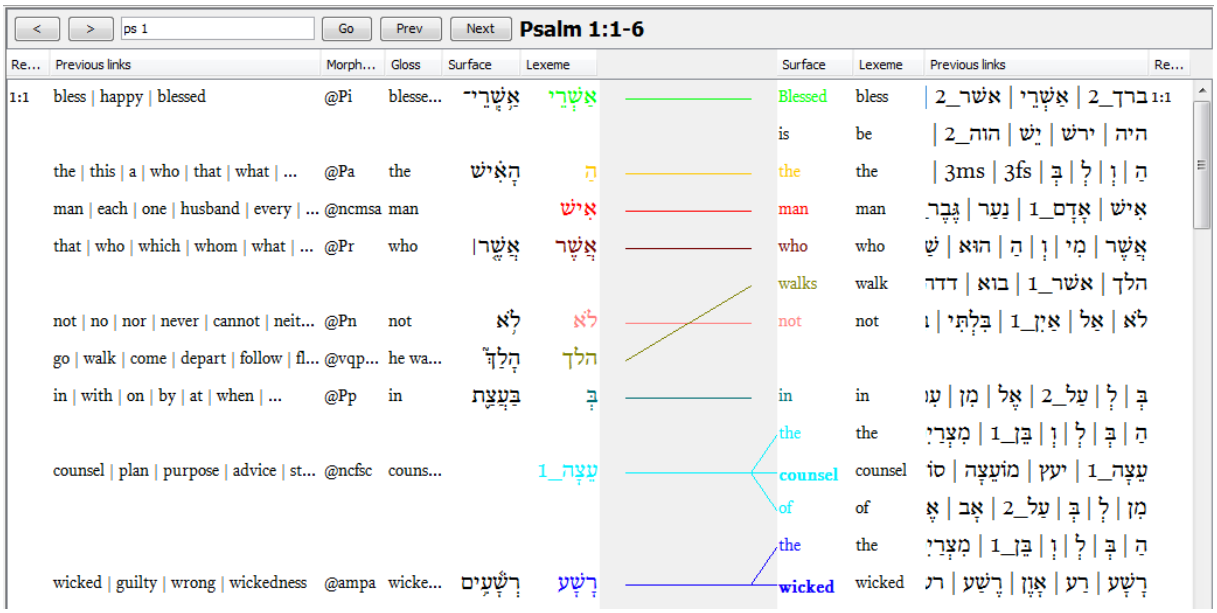

Figure 1. Alignment Panel

data structures during execution than if an external database were used, but the approach taken here supports responsiveness. Deciding where to put the blank rows is analogous to the more familiar problem of finding the weighted minimum edit distance between two strings with backtrace and thus can be done using the Wagner–Fischer algorithm, a dynamic programming algorithm that is *O(mn)* in both time and memory, where *m* and *n* are the number of source and target tokens (Wagner and Fischer, 1974). In addition, the tokens in token sets are connected via lines. For example, in Figure 1, the English tokens *the*, *counsel*, and *of* are connected together with lines. So as to avoid visual clutter, the line linking this English token set to the Hebrew meets at the primary token in the token set. If there is no primary token in the token set, a centrally located token is chosen instead.

The Alignment Panel uses fifteen different, easily distinguishable colors that still show up well on computer monitors for both tokens and lines to make it immediately clear which tokens are linked to one another. A few extremely common function words as well as pronominal suffixes in the source language always get a consistent color when they are linked. These are the tokens that cause Hebrew words often to contain multiple tokens. For the rest of the tokens, the colors are selected in such a way so as to avoid having similar colors near each other and to keep the colors as stable as possible as the user changes the alignment. In token sets containing multiple tokens, primary tokens are bolded.

When aligning modern languages, one might be able to assume that the aligners are fluent in both languages. However, when dealing with ancient languages or ancient dialects with relatively small corpora, language helps are a necessity in order to allow the aligner to work quickly. On the source language side, the Hebrew lemmas and morphology codes from *WHM* are presented to the aligner. The Hebrew lemmas are presented closest to the center rather than the surface forms simply because dividing multi-token Hebrew surface forms would look orthographically inappropriate and would be slower for the human aligner to process. For most languages the surface form should be presented closest to the center. A literal yet contextual gloss of the Hebrew token is also presented. These glosses were produced by Thom Blair using a separate software tool we wrote; they were designed for use in (*Hebrew-English Interlinear*, 2013). The English lemmas to which the Hebrew lemma has been linked elsewhere are also listed. To be listed, both the Hebrew lemma and the English lemma must have primary status. When there are multiple such English lemmas, they are listed in order of frequency of being linked. The target language side mirrors some of the source language side but is less extensive since we assume the aligner is fluent in English. The English lemmas were initially produced using *StanfordCoreNLP* (Toutanova et al., 2003, de Marneffe et al., 2006), with post-processing used to fix errors. The human aligner can edit them in case of errors.

2.2 Other panels aiding quality and consistency

Several other panels, shown in Figures 2-4, are designed to enable the aligner to check the alignment for quality and consistency.

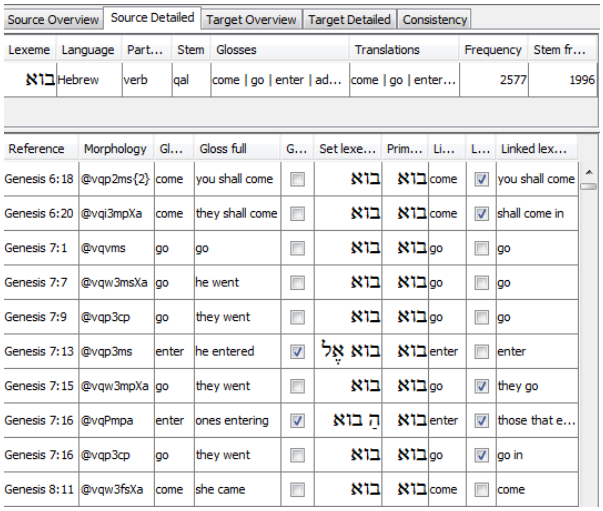

Figure 2. Source Detailed Panel

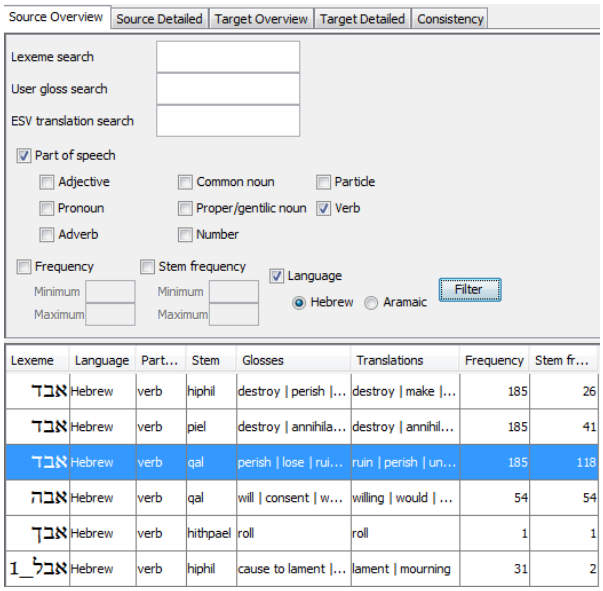

Figure 3. Source Overview Panel

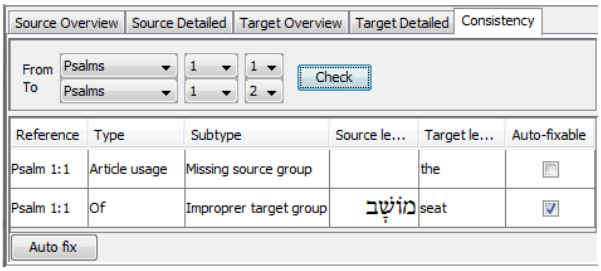

Figure 4. Consistency Panel

The Source Detailed Panel gives detailed information about the alignment for each occurrence of a lemma in the source text in a sortable table. In order to aid the aligner, the third column shows a form of the English gloss that has been shortened, usually to a single lemma, by making use of *WHM*'s morphology information and *WordNet*. The Target Detailed Panel is similar.

The Source Overview Panel briefly presents information concerning how all source tokens are aligned in a sortable, filterable table. The glosses shown are the short forms and are sorted based on frequency. Similarly, the translations are primary lemmas only and are sorted according to frequency. The Target Overview Panel is similar.

The Consistency Panel is oriented toward enforcing the consistency standards. It uses *WHM* as well as information from *StanfordCoreNLP*, including the syntactic dependency tree, to look for probable deviations from the project's consistency standards. It can fix some errors automatically if the human aligner allows it, but the human aligner is not required to follow its suggestions since it sometimes make mistakes, especially when the syntactic dependency tree from *StanfordCoreNLP* contains errors.

3 Conclusions and future work

The alignment tool is enabling a fast production of a high-quality, consistent gold-standard alignment between the Hebrew Bible and an English translation because of the way it provides an easy-to-process visualization of the alignment, provides options for aligners to dig deeper into the data and check their work, and makes changing the alignment easy. At present, the alignment tool is an in-house tool geared toward two specific texts, but with the exception of the consistency rules, which will be specific to particular languages and projects, it could be generalized to align other texts and languages. At that point, the generalized alignment tool could be licensed liberally to researchers.

Acknowledgments

The work presented in this paper has been funded by Crossway Books. The author would like to thank James Covington for his input.

References

Lars Ahrenberg, Mikael Andersson, Magnus Merkel. 2002. A System for Incremental and Interactive Word Linking. In *Proceedings of the 3nd Language Resources and Evaluation Conference (LREC 2002)*, Las Palmas, Spain. ELRA.

Lars Ahrenberg, Magnus Merkel, and Michael Petterstedt. 2003. Interactive Word Alignment for Language Engineering. In *Conference Companion of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003)*, pages 49-52, Budapest, Hungary. ACL.

Hal Daume III. HandAlign Documentation. `http://www.umiacs.umd.edu/~hal/HandAlign/`.

Victoria Fossum, Kevin Knight, and Steven Abney. 2008. Using Syntax to Improve Word Alignment Precision for Syntax-Based Machine Translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 44-52, Columbus, Ohio. ACL.

Alexander Fraser and Daniel Marcu. 2007. Measuring Word Alignment Quality for Statistical Machine Translation. *Computational Linguistics*, 33(3), 293-303.

Ulrich Germann. 2007. Two Tools for Creating and Visualizing Sub-sentential Alignments of Parallel Texts. In *Proceedings of the Linguistic Annotation Worship*, pages 121-124. Prague, Czech Republic. ACL.

Ulrich Germann. 2008 *Yawat*: Yet Another Word Alignment Tool. *Proceedings of the ACL-08: HLT Demo Session (Companion Volume)*, pages 20-23. Columbus, Ohio. ACL.

Stephen Grimes, Xuansong Li, Ann Bies, Seth Kulick, Xiaoyi Ma, and Stephanie Strassel. 2010. Creating Arabic-English Parallel Word-Aligned Treebank Corpora at LDC. In *Proceedings of the 7th International Language Resources and Evaluation Conference (LREC 2010)*, Valletta, Malta. ELRA.

Hebrew-English Interlinear ESV Old Testament: Biblia Hebraica Stuttgartensia (BHS) and English Standard Version (ESV). 2013. Wheaton, Il. Crossway.

Quoc Hung-Ngo and Werner Winiwarter. 2012. A Visualizing Annotation Tool for Semi-Automatically Building a Bilingual Corpus. In *Proceedings of the 8th International Language Resources and Evaluation Conference (LREC 2012)*, pages 67-74, Istanbul, Turkey. ELRA.

Patrik Lambert, Simon Petitrenaud, Yanjun Ma, and Andy Way. 2012. What types of word alignment improve statistical machine translation? *Mach Translat* 26, 289–323.

Nitin Madnani and Rebecca Hwa. 2004. The UMIACS Word Alignment Interface. `http://www.umiacs.umd.edu/~nmadnani/alignment/`.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. In *Proceedings of the 5th International Language Resources and Evaluation Conference (LREC 2006)*, pages 449-454, Genoa, Italy. ELRA.

I. Dan Melamed. 1998. Manual Annotation of Translational Equivalence: The Blinker Project. IRCS Technical Report #98-07. The University of Pennsylvania.

Magnus Merkel, Michael Petterstedt, and Lars Ahrenberg. 2003. Interactive Word Alignment for Corpus Linguistics. In *Proceedings of Corpus Linguistics 2003*, 533-542, Lancaster University, United Kingdom. UCREL technical paper 16.

Noah A. Smith and Michael E. Jahr. 2000. Cairo: An Alignment Visualization Tool. In *Proceedings of the 2nd Language Resources and Evaluation Conference (LREC 2000)*, Athens, Greece. ELRA.

James Strong. 1890. The exhaustive concordance of the Bible: showing every word of the text of the common English version of the canonical books, and every occurrence of each word in regular order: together with A comparative concordance of the Authorized and Revised versions, including the American variations: also brief dictionaries of the Hebrew and Greek words of the original, with references to the English words. Cincinnati: Jennings & Graham.

Jörg Tiedemann. 2006. ISA & ICA – Two Web Interfaces for Interactive Alignment of Bitexts. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 2154-2159, Genoa, Italy. ELRA.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 173-180, Edmonton, Canada. ACL.

Emmanuel Tov. 1986. A Computerized Data Base for Septuagint Studies: The Parallel Aligned Text of the Greek and Hebrew Bible. Computer Assisted Tools for Septuagint Studies (CATSS) Vol. 2. Journal of Northwest Semitic Languages Supplement Series 1. Stellenbosch.

Robert A. Wagner and Michael J. Fischer. 1974. The String-to-String Correction Problem. *Journal of the Association for Computing Machinery*, 21(1): 168-173.