# Transferring Structural Markup Across Translations Using Multilingual Alignment and Projection

David Bamman
The Perseus Project
Tufts University
Medford, MA
david.bamman@tufts.edu

Alison Babeu
The Perseus Project
Tufts University
Medford, MA
alison.jones@tufts.edu

Gregory Crane
The Perseus Project
Tufts University
Medford, MA
gregory.crane@tufts.edu

## ABSTRACT

We present here a method for automatically projecting structural information across translations, including canonical citation structure (such as chapters and sections), speaker information, quotations, markup for people and places, and any other element in TEI-compliant XML that delimits spans of text that are linguistically symmetrical in two languages. We evaluate this technique on two datasets, one containing perfectly transcribed texts and one containing errorful OCR, and achieve an accuracy rate of 88.2% projecting 13,023 XML tags from source documents to their transcribed translations, with an 83.6% accuracy rate when projecting to texts containing uncorrected OCR. This approach has the potential to allow a highly granular multilingual digital library to be bootstrapped by applying the knowledge contained in a small, heavily curated collection to a much larger but unstructured one.

## Categories and Subject Descriptors

H.3.7 [**Information Systems: Information Storage and Retrieval**]: digital libraries

## General Terms

Design, Documentation, Performance

## Keywords

Annotation projection, multilingual alignment, knowledge transfer

## 1. INTRODUCTION

One method of enhancing intellectual access to primary source texts contained within digital libraries is to provide physical access to translations of those texts as well. For contemporary readers of classical texts such as those written in Greek, Latin, Sanskrit or Classical Chinese, translations provide a window into languages no longer spoken, and translations of modern English texts into any other contemporary language such as Spanish or Arabic greatly broadens the accessibility of information from a strictly English-speaking audience to a global one.

A number of digital libraries already include translations of the source materials in their collections, including the World Digital Library[1] (which provides parallel translations in several languages for many of its texts), the Perseus Digital Library[2] (which provides English translations of Greek and Latin texts), the World of Dante[3] (English translations of *The Divine Comedy*), the Cervantes project[4] (*Don Quixote* in its original Spanish and several translations), and the Decameron Web[5] (in Italian and English).

One barrier to easily providing access to translations as part of a digital library is simply the cost-benefit analysis involved: one could, for example, acquire and digitize 10 works of Charles Dickens at the same cost as acquiring and digitizing *Bleak House* in 10 different languages. One further barrier is the level of labor-intensive markup that goes into creating a sophisticated digital document.

Figure 1 shows one example of a richly marked-up Latin text that is part of the Perseus Digital Library. Texts encoded in TEI, the de facto standard for literary document encoding, generally include citation-level structure specifying the division of a text into books, chapters, sections, acts, stanzas, and so on, but can also include much more sophisticated information, such as markup for speakers, quotations, people and places. Figure 1 includes structural citation information along with tags marking the names of ethnic groups (*Gallos*, "Gauls") and linking place names to their geographical coordinates (e.g., [-0.6,43.33]) and registry in the Getty Thesaurus of Geographical Names (e.g., "tgn,1124123").

High-quality markup such as this is generally expensive and time consuming to create: citation structure usually needs to be manually annotated by comparing the original print document with the digitized text, and while other forms of annotation such as named entity disambiguation can be done semi-automatically with the use of machine learning techniques [14, 21, 4, 13, 43], the manual correction of automated output can add to the total cost involved in creating such highly marked-up text (and the accuracy of that hand-annotation is generally what makes it valuable).


*JCDL'10,* June 21–25, 2010, Gold Coast, Queensland, Australia.

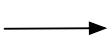

[1] http://www.wdl.org/
[2] http://www.perseus.tufts.edu/
[3] http://www.worldofdante.org/
[4] http://cervantes.tamu.edu
[5] http://www.brown.edu/Departments/Italian_Studies/dweb/

```
<name type="ethnic">Gallos</name>
ab
<name type="ethnic">Aquitanis</name>
<name key="tgn,1124123" type="place" reg="Garonne [-0.6,45.33] (river), Europe">
  <placeName key="tgn,1124123">Garumna</placeName>
</name>
flumen, a
<name type="ethnic">Belgis</name>
<name key="tgn,7009327" type="place" reg="Marne [4.183,48.916] (river), Champagne-Ardenne, France, Europe">
  <placeName key="tgn,7009327">Matrona</placeName>
</name>
et
<name key="tgn,7009707" type="place" reg="Seine [0.433,49.433] (river), France, Europe">
  <placeName key="tgn,7009707">Sequana</placeName>
</name>
dividit.
<milestone n="3" unit="section"/>
Horum omnium fortissimi sunt
<name type="ethnic">Belgae</name>
```

**Figure 1: Example of a richly marked-up XML fragment (Caesar, B.G. 1.1) including citation structure (`milestone n="3" unit="section"`) and named entity markers (`<name>` and `<placeName>`).**

The river Garonne separates the Gauls from the Aquitani; the Marne and Seine separate them from the Belgae.

Of all these, the Belgae are the bravest ...



```
The river
<name key="tgn,1124123" type="place" reg="Garonne
[-0.6,45.33] (river), Europe">
  <placeName key="tgn,1124123">Garonne</placeName>
</name>
separates the
<name type="ethnic">Gauls</name>
from the
<name type="ethnic">Aquitani</name>
; the
<name key="tgn,7009327" type="place" reg="Marne
[4.183,48.916] (river), Champagne-Ardenne, France, Europe">
  <placeName key="tgn,7009327">Marne</placeName>
</name>
and the
<name key="tgn,7009707" type="place" reg="Seine
[0.433,49.433] (river), France, Europe">
  <placeName key="tgn,7009707">Seine</placeName>
</name>
separate them from the
<name>Belgae</name>
.
<milestone n="3" unit="section"/>
Of all these, the
<name>Belgae</name>
are the bravest ...
```

**Figure 2: Projecting the XML tags from the Latin source document shown in figure 1 to an unstructured English translation of that same text.**

Citation-level TEI XML structure in both a source text and a translation is necessary as a fundamental first step for displaying both properly – if a digital library is to present "Canto 2" of Dante's *Inferno* in both the Italian original and an English translation, the canto needs to be delimited in both texts in the same way. Encoding other information (such as quotations) in both texts is, however, subject to the law of diminishing returns: if a source text has been marked up at the cost of considerable human labor, how much is it really worth to duplicate that labor for $n$ other translations?

The work described here aims to reduce the barriers to creating a multilingual reading environment for digital libraries by automating the XML markup of translations. Using techniques borrowed from computational linguistics and statistical natural language processing (which has produced an abundance of resources over the past twenty years for parallel text analysis), we present here a novel approach for automatically projecting structural information across translations, including canonical citation structure (such as chapters and sections), speaker information, and markup for quotations, people, places, and any other element in TEI-compliant XML that delimits spans of texts that are linguistically symmetrical. Figure 2 illustrates one such example of this: given the richly marked up source text shown in figure 1, our task is to project the XML tags present in that source document onto an unstructured translation (shown on the left) to produce the richly marked up version of the translated text at the right.

Our approach involves two stages: first, aligning the source document and the translation (hereafter the "the target document") at the level of individual sentences and then at the level of individual words, and, second, projecting the rele-

vant XML tags contained in the source document onto the target document in a way that exploits the linguistic similarity of the text pair. While the work described here has been developed using a collection of Greek and Latin texts along with their English translations, the methods themselves are language independent.

## 1.1 Background

Our work here builds on prior work aligning different editions of a text at the word and character level both within a single language (for the purpose of automatically evaluating OCR accuracy) [18] and across languages [38]. While our work here is focussed on bootstrapping a multilingual digital library and hence includes transferring structural information from a source document in one language to a document in another, aligning two texts in the same language (such as different editions) can still provide a valid initial alignment for subsequently projecting structural information as well.

The rise of parallel corpora in several languages such as the Canadian Hansards [40], the Europarl corpus [32], the JRC Acquis [46], News Commentary [5], and UN proceedings [49] has also driven an interest in computational linguistics in transferring linguistic information across parallel sentences. This has included syntactic information [52, 53, 30], morphological information [44, 17, 54], frame semantic information [48, 2], semantic roles [39] and temporal annotations [45]. These corpora have also been useful in using aligned second languages to improve NLP techniques in the first (e.g., using aligned Chinese-English data to help resolve English prepositional phrase ambiguity [19]). While this work generally focusses on transferring linguistic information between specific words, our task here involves the transfer of information that describes entire spans of text.

This work is also situated within the general landscape of multilingual digital libraries. Much of the research conducted in this area has focused on supporting more effective cross-language information retrieval (CLIR). An overview of the technical issues involved in supporting CLIR within the European Library with a specific focus on user query translation can be found in Agosti[1]. Clinchant[8] expands the standard language modeling approach by representing more than one language in the document model and then using a meta-dictionary in order to build a matching multi-language query model. A variety of research has also examined the multilingual mapping of different knowledge organization systems such as thesauri or subject headings in order to support CLIR in multilingual library collections. Wang[51] details the use of automatic methods to align multilingual subject headings in French, English and German, while Larson [33] explores a multilingual conceptual mapping resource that utilizes the online library catalog of the University of California as a translingual vocabulary resource. Rather than seeking to map multilingual query terms, Wang [50] studies the use of a web-based term translation approach to find translations for unknown cross-language queries in digital libraries. The issue of CLIR has also been explored in the cultural heritage domain. Szpektor[47] investigates aligning Hebrew and English queries in a museum collection through the use of a domain specific search engine combined with both semantic and cross-lingual expansion of user queries. The MultiMatch project supports cross-lingual user access to cultural heritage content across different media types by combining a domain-specific translation lexicon with a standard machine translation system [31]. While the parallel text analysis that underlies our work can provide an equal foundation for cross-language information retrieval, our goal is to first bootstrap a multilingual digital library to which CLIR techniques can later be applied.

## 1.2 Linguistic Symmetry

The techniques presented here are enabled by the simplifying assumption that the projection of structural information annotating spans of text (in the form of `<tag>` *text* `</tag>`) across translations is only theoretically sound if the two passages of text are linguistically symmetrical – i.e., a span of text $a_{\{1..n\}}$ in document $A$ is linguistically symmetrical with span $b_{\{1..n\}}$ from document $B$ if the two contain only equivalent expressions: $a_{\{1..n\}}$ cannot contain information not found in $b_{\{1..n\}}$ but found elsewhere in $B$, and $b_{\{1..n\}}$ cannot contain information not found in $a_{\{1..n\}}$ but found elsewhere in $A$.

What information can be said to be "equivalent" across translations is of course subject to debate, and the tradeoff between the fluidity of a translation and its fidelity to the source is a constant tension. Put simply, however, chapters of a book are generally linguistically symmetrical: a section of text from Book Two of Milton's *Paradise Lost* will usually not be found in Book Three of a Spanish translation.[6] In both the original and the translation, the spans of text between `<div type="book" n="2">` and their matching `</div>` tags will be equivalent. Names are also generally linguistically symmetrical: even if the word order changes from language to language, `<placeName>`The United States of America`</placeName>` and `<placeName>`États-Unis d' Amérique`</placeName>` both form contiguous units in both languages without outside elements intervening.

One important instance of asymmetry is the division of line breaks in poetry. Consider, for example, the original version of Vergil's Latin *Aeneid* along with John Dryden's English translation:

`<l>`Arma virumque cano, Troiae qui primus ab oris`</l>`
`<l>`Italiam, fato profugus, Laviniaque venit`</l>`
`<l>`litora, multum ille et terris iactatus et alto`</l>`
`<l>`vi superum saevae memorem Iunonis ob iram;`</l>`
`<l>`multa quoque et bello passus, dum conderet urbem`</l>`
(Vergil)

`<l>`Arms, and the man I sing, who, forc'd by fate,`</l>`
`<l>`And haughty Juno's unrelenting hate,`</l>`
`<l>`Expell'd and exil'd, left the Trojan shore.`</l>`
`<l>`Long labors, both by sea and land, he bore,`</l>`
`<l>`And in the doubtful war, before he won`</l>`
`<l>`The Latian realm, and built the destin'd town;`</l>`
(John Dryden)

As a whole, these spans of text are indeed linguistically symmetrical, but individually the lines are not: Vergil's line 1 does not correspond with any single line of Dryden's – while *arma virumque cano* is the equivalent of "Arms, and the man I sing," the second half of Vergil's line (*Troiae qui primus ab oris*) is distributed between line 1 ("who") and line 3 ("left the Trojan shore") of Dryden's edition.

[6]Though of course different editions of a text can create exceptions to this trend with the use of alternate citation schemes.

In the task of projecting structural information across translations, we take care to project only those tags that encode divisions of a text that result in linguistically symmetrical spans between the two languages. Small-scale logical divisions such as line numbers or arbitrary divisions such as physical page numbers will not usually result in symmetry, but markup for personal names (`<persName>`), place names (`<placeName>`), speakers `<speaker>`), quotations (`<quote>`), and large-scale logical text divisions (such as book, chapter, act, scene, poem, and canto) generally will.

# 2. METHODOLOGY

Transferring structural information from an edition of a text in one language to the same edition in another requires two subsequent stages: alignment at the level of individual words, and projection of information from its source location to its target one.

## 2.1 Multilingual Alignment

Textual alignment is the process of establishing a link between words in two different texts. In the alignment of two editions in the same language, the basis for this link may be simple identity (e.g., the first word – "arma" – of one version of the Latin *Aeneid* would correspond to that same "arma" in another) or string similarity (as measured by edit distance or Dice coefficient), but in two editions in different languages, the basis for the link is semantic – two words or sets of words are aligned to each other if they are translation equivalents. In the Latin phrase *omnia vincit amor* ("love conquers all"), *amor* would be linked to "love" in the English translation as its most direct equivalent.

Our work in projecting structural information from a text in one language to a text in another requires that the two texts be aligned at the level of individual words. Doing so requires a cascade of finer and finer alignments beginning at the document level, as shown in Figure 3.

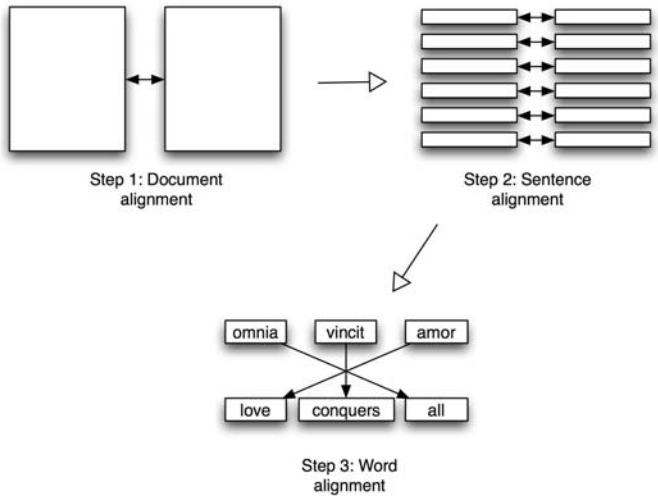



**Figure 3: Textual alignment workflow**

Once we manually establish that two documents are translations of each other (step 1), we align them on a sentence level (step 2) using Moore's Bilingual Sentence Aligner [36], which aligns sentences that are 1-1 translations of each other with a very high precision (98.5% for a corpus of 10,000 English-Hindi sentence pairs [42]). In our experiments to date, this process aligns approximately 30% of the sentences (since most are not 1-1 translations), but we use those high-precision alignments as anchors for the 1-2, 2-1 and many-to-many alignments that fall in between.

We then align all of these sentences – both the 1-1 aligned ones and the sentences in between – at the level of individual words (step 3) using MGIZA++ [20], a multi-threaded version of GIZA++ [38], an open-source implementation of the IBM alignment models 1-5 [3]. These alignment models are statistical approaches that attempt to capture the probability of a word $e$ at position $i$ in a sentence in a source language being translated by word $f$ at position $j$ in a target language sentence (and differ in respect to word order assumptions and fertility parameters – i.e., the probability that a target word is aligned to $\phi$ words in the source sentence). Our particular implementation makes use of IBM Model 4, in which a set of parameters $\theta$ for word equivalences, alignment and fertility is trained via the EM algorithm from the collection of parallel sentences to be aligned.

The parallel sentences are word aligned in both translation directions (from source to target and from target to source); we then take the intersection of those two alignments (yielding only highly precise alignments) and then extend the final alignment using the refined method described in Och and Ney [37].

Prior to alignment, all of the tokens in the source text and translation are stemmed to account for Greek and Latin's rich inflection and are then restored back to their original forms after the alignment is complete. This produces the raw data shown in figure 4, in which each word in the source text on the bottom (the beginning of Homer's *Odyssey* in Greek) is indexed to the corresponding words or words in the English translation at the top.

## 2.2 Projection

Step one provides us with an index between words in a source document and their corresponding translations in the target document. Since few translations are literal word-by-word reproductions of the original, not all words in the source document have an alignment in the target document (and vice versa): in prior experiments aligning 4.9 million words of Greek with 6.7 million words of English, a manual evaluation of 5,300 words yielded an F-measure of 67.9% in terms of overall accuracy.[7]

In order to project the XML tags from a source document to a target document, we use the alignments found in step one to locate the appropriate position in the target document into which each source tag should be projected. In this we exploit the principle of linguistic symmetry: for any pair of opening and closing tags delimiting a span of text in the source document, we locate the equivalent span of text in the target document using the word alignments and wrap that span of text with those start and end tags. Even if the word order changes within each of those spans, the boundaries will remain the same.

Figure 5 illustrates one simple example. The opening and closing division (`<div>`) tags on the Latin side enclose

[7]We measure word-level accuracy based on the evaluation standards used by the HLT/NAACL 2003 shared task on word alignment [35]: the overall F-measure of 67.9% reflects a 65.7% precision/70.4% recall; the non-null F-measure is 69.3% (63.7% precision/76.0% recall).

```
tell[1] me[2] ,[3] o[4] muse[5] ,[6] of[7] the[8] man[9] of[10] many[1] devices , who wandered full
ἄνδρα ({ 9 }) μοι ({ 2 }) ἔννεπε ({ 1 }) , ({ 3 }) μοῦσα ({ 4 5 }) , ({ 6 }) πολύτροπον ({ 12 }) , (

but[1] he[2] took[3] from[4] them[5] the[6] day[7] of[8] their[9] returning .
αὐτὰρ ({ 1 }) ὁ ({ }) τοῖσιν ({ }) ἀφείλετο ({ 2 3 4 5 }) νόστιμον ({ 9 10 }) ἦμαρ ({ 7 }) . ({ 11 })

of[1] these[2] things[3] ,[4] goddess[5] ,[6] daughter[7] of[8] zeus[9] , beginning where thou wilt
τῶν ({ 1 }) ἀμόθεν ({ 2 3 }) γε ({ }) , ({ 4 }) θεά ({ 5 }) , ({ 6 }) θύγατερ ({ 7 }) Διός ({ 9 }) ,
```

**Figure 4: Automatic alignment data for Homer's *Odyssey*.**

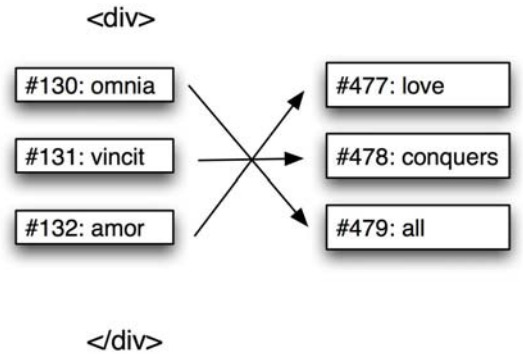



**Figure 5: Projecting XML tags around two spans of text that are linguistically symmetrical**

three words: *omnia* at source position #130, *vincit* at position #131 and *amor* at position #132. These words align, respectively, to "all" (at target position #479), "conquers" (#478) and "love" (#477). Since the opening and closing `<div>` tags wrap a continuous section of text in the source document that aligns to positions #477-479 in the target document, we can insert the start `<div>` tag immediately before position #477 in the target document and the end `</div>` tag immediately after position #479.

More formally, we can identify the positions in the target document for the start tags and the end tags with the following algorithms. Both presume an alignment $A\{a_0..a_n\}$ that consists of pairs of indices of words in the source document and the target document (e.g., source word #130 = target word #157, source word #143 = target word #171, etc.).

### *Start tags.*

For start tags such as `<div>`, the projected target position is immediately to the left of the leftmost alignment in the elements contained.

**Algorithm 1** Target position for start tags (`<example>`)

**Require:** alignment $A\{a_0..a_n\}$
  $minval \leftarrow \infty$
  **for all** words $i$ between start tag and end tag **do**
    $j \leftarrow alignment(i)$
    **if** $j < minval$ **then**
      $minval \leftarrow j$
    **end if**
  **end for**
  $position \leftarrow minval - 1$

### *End tags.*

For ending tags such as `</div>`, the projected target position is immediately to the right of the rightmost alignment in the elements contained.

**Algorithm 2** Target position for end tags (`</example>`)

**Require:** alignment $A\{a_0..a_n\}$
  $maxval \leftarrow 0$
  **for all** words $i$ between start tag and end tag **do**
    $j \leftarrow alignment(i)$
    **if** $j > maxval$ **then**
      $maxval \leftarrow j$
    **end if**
  **end for**
  $position \leftarrow maxval + 1$

### *Empty element tags.*

Even though empty elements (such as `<milestone />` tags) do not by definition "contain" anything, we can still treat them as start tags containing the rest of the document or as end tags containing the beginning, making use of either algorithm 1 or 2 as appropriate. We found a much higher accuracy in the experiments that follow by treating them as end tags, most likely due to the extremely high precision of aligning terminal punctuation across sentence pairs (which generally immediately precede structural tags like milestones). Algorithm 3 therefore reflects this property of our data and is a modification of algorithm 2.

**Algorithm 3** Target position for empty element tags (`<example />`)

**Require:** alignment $A\{a_0..a_n\}$
  $maxval \leftarrow 0$
  **for all** words $i$ between 0 and tag **do**
    $j \leftarrow alignment(i)$
    **if** $j > maxval$ **then**
      $maxval \leftarrow j$
    **end if**
  **end for**
  $position \leftarrow maxval + 1$

## 3. EVALUATION

To judge the accuracy of this approach, we evaluated it against two different datasets: first, a large collection of carefully transcribed English editions of several Classical texts, either manually double-keyboarded or scanned and OCR'd,

**Table 1: Accuracy by text (transcribed collection)**

| Author | Text | Translator | Date | Accuracy | # Tags Projected | Avg. Distance from Correct |
|---|---|---|---|---|---|---|
| Herodotus | Histories | Godley | 1920 | 85.3% | 5934 | 7.7 |
| Homer | Iliad | Murray | 1924 | 94.9% | 474 | 6.5 |
| Homer | Odyssey | Murray | 1919 | 95.8% | 337 | 5.7 |
| Pausanias | Description of Greece | Jones | 1918 | 89.0% | 3547 | 7.4 |
| Xenophon | Anabasis | Brownson | 1922 | 91.2% | 1533 | 5.7 |
| Xenophon | Hellenica | Brownson | 1921 | 91.9% | 1198 | 6.6 |
| Total | | | | 88.2% | 13,023 | 7.3 |

both with over 99.94% character-level transcription accuracy; and second, a smaller collection of post-1850 translations of Homer's *Odyssey* drawn from the scanned and OCR'd collections of the Internet Archive and Google Books, both publicly available online.

## 3.1 Transcribed collection

For the carefully transcribed collection, we draw from the texts that have been digitized and manually corrected as part of the Perseus Digital Library [10, 12, 11]. Established in 1987 in order to construct a large, heterogeneous collection of textual and visual materials on the archaic and classical Greek world, Perseus has accumulated a number of source texts along with translations, commentaries and lexica over the past twenty years to create an open reading environment for the study of Classical texts. All of these texts are marked up in TEI-compliant XML, and since the field of Classical Studies has long adopted a canonical citation scheme for referring to segments of texts (such as "Thuc. 1.1" to refer to "Book 1, Chapter 1" of Thucydides' *History of the Peloponnesian War*), a source text in Greek or Latin is often marked up with exactly the same book, chapter and section numbers as its English translation.

The presence of these identical citation schemes between a source text and its translation allows us to automatically evaluate on a large scale the accuracy of projecting structural information. After identifying a set of pairs of texts (the Greek original along with its English translation) with exactly the same textual divisions, we stripped the English translation of all XML markup, divided the source text and translation into sentences, aligned those sentences using Moore's Bilingual Sentence Aligner, aligned all of the words within each sentence pair using MGIZA++, and then projected the structural tags from the source text to the translation using the algorithms defined in section 2.2. For the purpose of this experiment, we projected only `<milestone>` and `<div>` tags associated with books, sections, chapters, and cards. The results of this evaluation are shown in table 1.

The overall accuracy in attempting to project 13,023 structural tags from the original Greek source text marked up in TEI-compliant XML to a plain-text English translation is 88.2%, but this varies relatively widely by author, from 85.3% for Herodotus up to 95.8% for Homer's *Odyssey*. To judge how effective this projection would be for the task of manual error-correction (i.e. to attain 100% accuracy), we also evaluated the average distance to the correct position in the case of errors. The average distance across all texts of approximately 7.3 words means that a human corrector would in most cases only have to search a window of 15 words (7.3 words before the projected tag and 7.3 after) to find the correct position.

## 3.2 Automatically OCR'd collection

The rise of large-scale digitization efforts such as those by Google Books and the Internet Archive is beginning to make huge volumes of data available for public use, often including several translations in different languages for any given text. One opportunity that these large digital collections present is bootstrapping a multilingual digital library from a monolingual one by simply locating translations for a source text within them and then projecting the structural information from that source text to the plain-text translations. Since the automatically scanned and OCR'd texts within these large collections often include significant OCR errors and formatting noise, we evaluated the performance of our approach on a set of uncorrected texts drawn from these collections in order to gauge the degradation in performance that might result as a consequence of transcription errors.

To evaluate this, we took one Greek text from the Perseus Digital Library, Homer's *Odyssey*, with a baseline transcribed accuracy of 95.8%, and located seven translations (six from the Internet Archive and one from Google Books): Avia (1880) [24], Barnard (1876) [23], Cotterill (1911) [28], Mackail (1903) [27], Morris (1887) [25], Norgate (1863) [22], and Palmer (1891) [26]. The Internet Archive provided OCR'd versions for all of its texts, and we OCR'd the Google text (Avia) in-house using Abbyy FineReader.

We then subjected the document pairs (each translation along with a copy of the Greek original) to the same process described for the transcribed texts above: each pair was divided into sentences, aligned using Moore's Bilingual Sentence Aligner, and each aligned sentence was word aligned using MGIZA++. The resulting alignment guided the projection of every `<milestone>` and `<div>` tag denoting books, chapters, sections and cards from the source Greek document to each target English document.

Since the texts from the Internet Archive and Google Books do not have gold standard versions (as our transcribed texts do), we evaluated a subset of them by hand, including all and only those tags from the first one-third of the *Odyssey* (books 1-8 of 24), for a total of 111 for each document. The results of this evaluation are shown in table 2.

Compared to the baseline of 95.8% for a transcribed and corrected English translation of Homer's *Odyssey* (Murray [29]), the uncorrected OCR versions achieve on average a degradation of 12.7%, for an average overall accuracy of 83.6%. Again, however, the performance by individual text varies, from 70.3% (Norgate) to 91.9% (Palmer). And in

**Table 2: Accuracy by text (automatically OCR'd collection)**

| Translator | Date | Accuracy | # Tags Projected | Avg. Distance from Correct |
|---|---|---|---|---|
| Avia | 1880 | 76.6% | 111 | 14.9 |
| Barnard | 1876 | 89.2% | 111 | 13.4 |
| Cotterill | 1911 | 82.8% | 111 | 21.2 |
| Mackail | 1903 | 90.1% | 111 | 9.3 |
| Morris | 1887 | 83.8% | 111 | 12.7 |
| Norgate | 1863 | 70.3% | 111 | 16.1 |
| Palmer | 1891 | 91.9% | 111 | 10.5 |
| Total | | 83.6% | 777 | 13.9 |

cases of error, while the distance to the correct location is almost doubled in comparison to perfectly transcribed texts (13.9), the correct position can, on average, often be found within a window of 28 words.

An analysis of the textual differences between these seven translations in comparison with Murray's 95.8% baseline reveals a strong relationship between the "poeticism" of the translation and the accuracy of the overall projection. Murray's translation (part of an *en face* edition with the Greek on one side of the page and the English translation on the other) is highly faithful to the original Greek. The translations with the highest projected accuracy also maintain a similarly high level of fidelity to the text:

- Tell me, O Muse, of the man of many devices, who wandered full many ways after he had sacked the sacred citadel of Troy (Murray, 95.8%)
- Speak to me, Muse, of the adventurous man who wandered long after he sacked the sacred citadel of Troy. (Palmer 91.9%)
- O Muse instruct me of the man who drew His changeful course through wanderings not a few After he sacked the holy town of Troy ... (Mackail, 90.1%)

The lowest performing texts, in contrast, show a marked freedom of translation, including strong deviations from the word order of the original (e.g., the postponement of "Tell me, O Muse" to the second half of the clause in Avia's translation) and the use of infrequent vocabulary (as in "craft-renown," "song-goddess" and "Troy-town" in Norgate's edition):

- The travelled Man of many a turn – driven far, Far wandering, when he sacked Troy's sacred Town; Tell me, O Muse, his tale (Avia, 76.6%)
- The Hero of craft-renown, O song-goddess, chant me his fame, who, when low he laid Troy-town, unto many a far land came ... (Norgate, 70.3%)

We can explain this phenomenon by the heterogenous nature of the corpus used to build our language models – the sentence and word alignment models are trained on a collection of Greek texts and translations that includes a much larger selection of prose than poetry, so translations that significantly deviate from a prose-like word order and vocabulary will naturally attain lower accuracy scores. In the future we may wish to explore training poetry-only alignment models, but for large-scale textual collections that include a mixture of both, we expect the fidelity of a translation to the original text to generally lead to higher alignment and projection scores.

# 4. BOOTSTRAPPING A MULTILINGUAL DIGITAL LIBRARY

Projecting structural information from a well-curated source text to a noisy set of translations has the potential to allow us to bootstrap a multilingual digital library from a monolingual collection. As more and more translations become publicly available as part of open digitization efforts, automated methods that can deal well with scale become increasingly valuable. There are two directions in which this work in particular can help expand the scope of a library's collection: expanding the breadth of translations in which a given work is available, and expanding their depth as well.

## 4.1 Expanding breadth

The most immediate impact of making a source text available in a number of different languages is surely the wider reach that such a multilingual library has in a global environment – every additional language added provides basic access to a corresponding group of native speakers. Beyond this broad impact, however, expanding the breadth of translations also fundamentally enables cross-linguistic scholarship.

Humanities collections and historical collections in digital libraries both tend to be be multilingual in the extreme. Works of literature that have taken their place within a language's canon are often translated multiple times in several different languages. The Internet Archive alone contains editions of Horace's *Odes* in at least eight different languages – not only in the Latin original, but in English, Spanish, Italian, French, Early Modern French, Portuguese and German, often in several different editions even in the same language.

- Latin: carpe diem quam minimum credula postero (Horace, Ode 1.11)
- English: Seize the present; trust tomorrow e'en as little as you may (Conington 1872) [9]
- French: Cueille le jour, et ne crois pas au lendemain (De Lisle 1887) [34]
- Early Modern French: Jouissez donc en repos du jour present, & ne vous attendez point au lendemain (Dacier 1681) [15]
- Italian: tu l'oggi goditi: e gli stolti al domani s'affidino (Chiarini 1916) [7]
- Spanish: Coge este dia, dando muy poco credito al siguiente (Campos and Minguez 1783) [6]
- Portuguese: colhe o dia, do de amanhá mui pouco confiando (Duriense 1807) [16]


**Figure 6: A screenshot of Homer's *Odyssey* from the Perseus Digital Library, along with William Morris' 1887 translation of it.**

- German: Pflücke des Tag's Blüten, und nie traue dem morgenden (Schmidt 1820) [41]

Presenting all of these different translations for any given literary work not only has the effect of appealing to a much broader global audience, but also of enabling fundamental research on the source text itself, including its evolution and reception across languages and different historical eras.

## 4.2 Expanding depth

Projecting structural information across translations also enables a digital library to present multiple translations within a single language. Presenting a number of different translations by different authors helps contextualize a source text by enumerating the different ways that it has been historically understood. This is especially important for digital libraries that double as pedagogical environments, as each translation is in effect a commentary on the source text as well.

Providing depth of translation also allows us to present literary translations in their historical context. Figure 6 presents one such example of this. The Morris (1887) translation we used from the Internet Archive is that by William Morris, an English artist, writer and pioneer of the Arts and Craft Movement in the late 19th century – as a translation, it is of interest as a literary object in its own right. By using the original Greek version to project the citation structure onto this translation, we are able to easily include it in our existing environment, making the resulting library of use not only to Classicists, but to English scholars as well.

# 5. CONCLUSION

Translations by their very nature tend to require the same kind of structural markup as the source texts they translate but the costs of manually annotating such markup for a number of different languages is often prohibitively high. The ability to automatically project structural information from one document to another has the potential to lower the costs involved in incorporating existing translations into a digital library, lowering one barrier to exposing collections to a much wider global audience.

Beyond this, however, the work described here stands at the intersection between small, carefully curated digital collections (such as the Perseus Digital Library) and much larger but unstructured ones (such the Internet Archive and Google Books). By projecting the knowledge contained in the heavily annotated texts in our collection, we can enrich a much larger collection of texts that have simply been scanned and OCR'd. And, in turn, by beginning to exploit the vast range and depth of those collections, we can approach a scale of texts and languages necessary to make a digital library truly multilingual.

# 6. ACKNOWLEDGMENTS

Grants from the Andrew W. Mellon Foundation ("The CyberEdition Project: Workflow for Textual Data in Cyberinfrastructure") and the National Endowment for the Humanities (PR-50013-08, "The Dynamic Lexicon: Cyberinfrastructure and the Automated Analysis of Historical Languages") provided support for this work. This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. This paper is made available under a Creative Commons Attribution license.

# 7. REFERENCES

[1] M. Agosti, M. Braschler, N. Ferro, C. Peters, and S. Siebinga. Roadmap for multilingual information

access in the European Library. *Research and Advanced Technology for Digital Libraries*, 2007.

[2] R. Basili, D. De Cao, D. Croce, B. Coppola, and A. Moschitti. Cross-language frame semantics transfer in bilingual corpora. In *Computational Linguistics and Intelligent Text Processing*, Lecture Notes in Computer Science, chapter 27, pages 332–345. 2009.

[3] P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer. The mathematics of statistical machine translation: parameter estimation. *Comput. Linguist.*, 19(2):263–311, 1993.

[4] R. C. Bunescu and M. Pasca. Using encyclopedic knowledge for named entity disambiguation. In *EACL*. The Association for Computer Linguistics, 2006.

[5] C. Callison-Burch, C. Fordyce, P. Koehn, C. Monz, and J. Schroeder. (meta-) evaluation of machine translation. In *StatMT '07: Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Morristown, NJ, USA, 2007. Association for Computational Linguistics.

[6] U. Campos and L. Minguez, editors. *Horacio, Español o Poesias Lyricas Q. Horacio Flacco*. De Sancha, Madrid, 1783.

[7] G. Chiarini, editor. *Odi, Epodi e Sermoni di Orazio*. Perrella, Napoli, 1916.

[8] S. Clinchant and J.-M. Renders. Multi-language models and meta-dictionary adaptation for accessing multilingual digital libraries, 2009.

[9] J. Conington, editor. *The Odes and Carmen Saeculare of Horace*. Bell and Daldy, London, 1872.

[10] G. Crane. From the old to the new: Integrating hypertext into traditional scholarship. In *Hypertext '87: Proceedings of the 1st ACM conference on Hypertext*, pages 51–56. ACM Press, 1987.

[11] G. Crane. New technologies for reading: The lexicon and the digital library. *Classical World*, pages 471–501, 1998.

[12] G. Crane, D. Bamman, L. Cerrato, A. Jones, D. M. Mimno, A. Packel, D. Sculley, and G. Weaver. Beyond digital incunabula: Modeling the next generation of digital libraries. In J. Gonzalo, C. Thanos, M. F. Verdejo, and R. C. Carrasco, editors, *ECDL*, volume 4172 of *Lecture Notes in Computer Science*, pages 353–366. Springer, 2006.

[13] G. Crane and A. Jones. The challenge of Virginia Banks: an evaluation of named entity analysis in a 19th-century newspaper collection. In *JCDL '06: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, pages 31–40, New York, NY, USA, 2006. ACM.

[14] S. Cucerzan. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 708–716, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[15] A. Dacier, editor. *Remarques Critiques sur les Oeuvres d'Horace*. Thierry & Barbin, Paris, 1681.

[16] E. Duriense, editor. *A Lyrica de Q. Horacio Flacco*. Impr. Regia, Lisboa, 1807.

[17] A. Feldman, J. Hana, and C. Brew. Experiments in cross-language morphological annotation transfer. In *Computational Linguistics and Intelligent Text Processing*, Lecture Notes in Computer Science, chapter 4, pages 41–50. 2006.

[18] S. Feng and R. Manmatha. A hierarchical, HMM-based automatic evaluation of OCR accuracy for a digital library of books. In *JCDL '06: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, pages 109–118, New York, NY, USA, 2006. ACM.

[19] V. Fossum and K. Knight. Using bilingual Chinese-English word alignments to resolve PP-attachment ambiguity in English. In *Proceedings of the 8th AMTA*, 2008.

[20] Q. Gao and S. Vogel. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57, Columbus, Ohio, June 2008. Association for Computational Linguistics.

[21] X. Han and J. Zhao. Named entity disambiguation by leveraging wikipedia semantic knowledge. In *CIKM '09: Proceeding of the 18th ACM conference on Information and knowledge management*, pages 215–224, New York, NY, USA, 2009. ACM.

[22] Homer. *The Odyssey or, The ten years' wandering of Odusseus, after the ten years' siege of Troy. Reproduced in dramatic blank verse (trans. T. S. Norgate)*. Williams and Norgate, London, 1863.

[23] Homer. *The Odyssey of Homer : rendered into English blank verse (trans. Barnard)*. Williams and Norgate, London, 1876.

[24] Homer. *The Odyssey of Homer (trans. Avia)*. C. Kegan Paul & Co., London, 1880.

[25] Homer. *The Odyssey of Homer done into English verse (trans. William Morris)*. Reeves & Turner, London, 1887.

[26] Homer. *The Odyssey of Homer (trans. George Herbert Palmer)*. Houghton Mifflin and Co., Boston, New York, 1891.

[27] Homer. *The Odyssey, in 3 vols. (trans. J. W. Mackail)*. John Murray, Albemarle Street, London, 1903.

[28] Homer. *Homer's Odyssey. A Line-For-Line Translation in the Metre of the Original by H. B. Cotterill M. A.* G.G. Harrap & company, London, 1911.

[29] Homer. *The Odyssey with an English Translation by A.T. Murray, PH.D. in two volumes*. Harvard University Press; William Heinemann, Ltd., 1919.

[30] R. Hwa, P. Resnik, A. Weinberg, C. Cabezas, and O. Kolak. Bootstrapping parsers via syntactic projection across parallel texts. *Nat. Lang. Eng.*, 11(3):311–325, 2005.

[31] G. J. F. Jones, Y. Zhang, E. Newman, F. Fantino, and F. Debole. Multilingual search for cultural heritage archives via combining multiple translation resources. In *Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007).*, pages 81–88. Association for Computational Linguistics, 2007.

[32] P. Koehn. Europarl: A parallel corpus for statistical

machine translation. In *Proceedings of the MT Summit X*, Phuket, Thailand, 2005.

[33] R. R. Larson, F. Gey, and A. Chen. Harvesting translingual vocabulary mappings for multilingual digital libraries. In *JCDL '02: Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*, pages 185–190. ACM Press, 2002.

[34] L. D. Lisle, editor. *Oeuvres de Horace*. Alphonse Lemerre, Paris, 1887.

[35] R. Mihalcea and T. Pedersen. An evaluation exercise for word alignment. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and using parallel texts*, pages 1–10, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[36] R. C. Moore. Fast and accurate sentence alignment of bilingual corpora. In *AMTA '02: Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users*, pages 135–144, London, UK, 2002. Springer-Verlag.

[37] F. J. Och and H. Ney. Improved statistical alignment models. In *ACL '00: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 440–447, Morristown, NJ, USA, 2000. Association for Computational Linguistics.

[38] F. J. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

[39] S. Padó and M. Lapata. Cross-lingual annotation projection of semantic roles. *Journal of Artificial Intelligence Research*, 36:307–340, 2009.

[40] S. Roukos, D. Graff, and D. Melamed. Hansard French/English.

[41] K. Schmidt, editor. *Des Horatius Flaccus Sämmtliche Lyrische Dichtungen*. Vogler, Halberstadt, 1820.

[42] A. K. Singh and S. Husain. Comparison, selection and use of sentence alignment algorithms for new language pairs. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 99–106, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[43] D. A. Smith and G. Crane. Disambiguating geographic names in a historical digital library. In *ECDL '01: Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries*, pages 127–136, London, UK, 2001. Springer-Verlag.

[44] B. Snyder and R. Barzilay. Crosslingual propagation for morphological analysis. In *Proceedings of the Twenty Third National Conference on Artificial Intelligence*, 2008.

[45] K. Spreyer and A. Frank. Projection-based acquisition of a temporal labeller. In *Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP-2008)*, pages 489–496, Hyderabad, India, January 2008.

[46] R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufis, and D. Varga. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. *CoRR*, abs/cs/0609058, 2006.

[47] I. Szpektor, I. Dagan, A. Lavie, D. Shacham, and S. Wintner. Cross lingual and semantic retrieval for cultural heritage appreciation. In *Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007).*, pages 65–72. Association for Computational Linguistics, 2007.

[48] S. Tonelli and E. Pianta. Frame information transfer from english to italian. In *Proceedings of LREC*, 2008.

[49] UN. ODS UN parallel corpus, 2006.

[50] J.-H. Wang, J.-W. Teng, P.-J. Cheng, W.-H. Lu, and L.-F. Chien. Translating unknown cross-lingual queries in digital libraries using a web-based approach. In *Digital Libraries, 2004. Proceedings of the 2004 Joint ACM/IEEE Conference on*, pages 108–116, 2004.

[51] S. Wang, A. Isaac, B. Schopman, S. Schlobach, and L. van der Meij. Matching multi-lingual subject vocabularies. *Research and Advanced Technology for Digital Libraries*, pages 125–137, 2009.

[52] A. Wróblewska and A. Frank. Cross-lingual projection of LFG F-structures: Building an F-structure bank for Polish. In *Eighth International Workshop on Treebanks and Linguistic Theories*, page 209, 2009.

[53] F. Xia and W. D. Lewis. Multilingual structural projection across interlinear text. In *Proceedings of NAACL HLT*, pages 452–459, 2007.

[54] D. Yarowsky and G. Ngai. Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In *NAACL '01: Second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies 2001*, pages 1–8, Morristown, NJ, USA, 2001. Association for Computational Linguistics.