愛知工業大学研究報告
第 47 号 平成 24 年

タトエバ・プロジェクト・コーパスを使った www.ManyThings.org の語学学習教材

Language Study Materials on www.Manythings.org That Use the Tatoeba Project Corpus

Charles Kelly †
チャールズ・ケリー

Abstract: This paper demonstrates and explains how I use the data from the Tatoeba Project for some of my web projects. The Tatoeba Project maintains a large multi-lingual database of example sentences. An explanation of how data is selected and a description of the various projects are included. Ensuring the accuracy of English examples is a critical job that I do personally. For further details on methodology, see Section 3 below. By selecting and correcting data from the Tatoeba Project I have developed several web-based language learning tools with the highest possible degree of accuracy in example sentences.

1. Introduction

The Tatoeba Project at tatoeba.org has a large database of example sentences translated into many languages by its members who volunteer their time. Sentences from this project may be used under the Creative Commons - Attribution 2.0 license. This project imported the large Tanaka Corpus of English-Japanese sentence pairs. (See below.)

This paper gives a short background of the Tatoeba Project, describes how I select sentences from this project for use in my own projects and briefly describes some of my projects that use these sentences.

2. Background of the Tatoeba Project

The history of the Tatoeba Project begins with the creation of a corpus that was compiled by Yasuhito Tanaka at Hyogo University in Japan. He compiled this corpus with more than 200,000 English-Japanese sentence pairs over several years by having his students each contribute 300 sentence pairs. This file was released as public domain at the Pacific Association for Computational Linguistics (PACLING) 2001 Conference where he presented a paper related to this corpus. Christian Boitet of Joseph Fourier University in France distributed the Tanaka Corpus on a CD-ROM at the 2002 Papillion Workshop in Tokyo, Japan. Jim Breen of Monash University in Australia integrated the Tanaka Corpus into his WWWJDIC Online Japanese-English dictionary in 2003 after removing duplicates, cleaning up a lot of the errors, removing garbled sentences and other editing. In March of 2006, Paul Blay took over the maintenance of the Tanaka Corpus from Jim Breen and continued with the cleanup of the data. This corpus with edits and additions was distributed as public domain until October 2008. The last public domain version contained about 150,350 sentence pairs.

In August of 2006, Ngoc Phuong Trang Ho, then a student at University of Technology of Compiègne in France, started the project at http://multilangdic.sourceforge.net under the name of Trang's Dictionary Project that she described as a "Wikipedia type of thing, except people add sentences, not articles." In the summer of 2007, the project moved to http://wwwcyg.utc.fr/tatoeba and officially became known as the Tatoeba Project. In September

†愛知工業大学　基礎教育センター（豊田市）

of 2007, the Tanaka Corpus, as maintained by Jim Breen and Paul Blay, was imported into the Tatoeba Project. The Tanaka Corpus was officially maintained on the WWWJDIC side for quite a while with Trang Ho and Paul Blay keeping the two synchronized. The tatoeba.org domain name was registered in June of 2008. In December of 2009, Jim Breen closed down his feedback system for correcting errors and started using the Tatoeba Project to maintain the sentence examples he uses with his WWWJDIC. Also in December of 2009, Trang made an announcement of a change from public domain to the Creative Commons Attribution license.

3. Selecting Data to Use

The Problem

One problem with using data from the Tatoeba Project was deciding how to select good sentences that sound natural and are error-free. Briefly, the kinds of problems seen in the data are the following.

(1) Some sentences contain grammar and vocabulary errors.
(2) Some sentences sound unnatural, even though they may be grammatically correct.
(3) There is archaic and old-fashioned language usage that is not suitable for people learning a foreign language.
(4) Some sentences are not appropriate for use in international educational projects because they are potentially offensive to some cultures and/or are inappropriate for young children.

To help solve this problem, I use the following methods.

Blacklisting Data

On tatoeba.org, there is a tagging system that allows advanced contributors to tag sentences with tags such as “needs native speaker check,” “archaic” and “XXX.” Using these tags, such sentences can easily be filtered out.

It is also possible for members of the Tatoeba Project to create lists on the site such as “perhaps not child safe” and “not for my projects.” Such lists can also be used to filter out unwanted sentences.

The usernames connected with sentences is also available. This allows easy filtering out of all sentences by members who have a certain number of sentences not appropriate for my projects.

Whitelisting Data

In addition to filtering out sentences, I whitelist sentences.

The steps I use in whitelisting are as follows.

(1) I use all sentences by a very limited number of highly trusted native English speakers.
(2) I proofread all other sentences personally. Since native speakers are more likely to contribute the kind of sentences I can use, I give priority to proofreading their sentences.
(3) Next, I proofread sentences by the non-native English speakers that I know make few errors.
(4) When I have additional time, I proofread other English sentences.

4. Trusted Pairs

For my projects, I only use bilingual sentences pairs made up of sentences that I feel can be trusted as very likely to be error-free. For English, I choose to trust only the sentences that I have proofread and whitelisted. For the other languages, I choose to trust only sentences by native speakers of those languages.

To increase the number of trusted pairs, I encourage native speakers of the other languages to translate from sentences already proofread by me. I have done this by creating lists of non-translated, proofread sentences for various users.

I have also contributed a large number of English sentences to the Tatoeba Project. Once translated by a native speaker of another language, I can use these sentence pairs.

Also, I have been maintaining a list of native speakers who are contributing to the Tatoeba Project in order to help people find trustworthy sentences to translate into their own native languages. This helps create trusted sentence pairs.

5. Advantages of the Tatoeba Project

There Are Many Sentences

There is a large amount of data that can be used under the Creative Commons Attribution license.

It is a Collaborative Effort

Since this is a collaborative effort, members help each other by catching errors and suggesting corrections.

The Website Has an Effective User Interface

The website is set up to allow members to easily contribute new material and to edit existing material. Since it is easy to use, many new sentences are being added every day.

6. Disadvantages of the Tatoeba Project

One Must Work Within the Tatoeba Project's Guidelines

In order to follow the set of guidelines, editing data requires a bit more time. The project does not want correct sentences to be changed even if they are archaic or unnatural-sounding.

Instead of being able to quickly edit sentences, changes such as the following require the more time-consuming step of submitting an alternate entry.

(1) Instead of changing archaic language in place, one is required to submit a modern equivalent as a separate entry.
(2) Instead of correcting unnatural-sounding, yet grammatically correct sentences in place, one needs to submit an alternate entry.
(3) In order to standardize spelling and number formats, it is necessary to submit an alternate entry and then blacklist the original.

All Items in the Database Are Not Sentences

Even though the stated aim of the project is to collect sentences, the project allows incomplete sentences and items such as movie titles. Therefore, the end user of the data must be careful to eliminate these items before using the data with a project focused on sentences.

Some Members Contribute Sentences in Languages Other Than Their Own Native Language.

Though it is possible for some people to create natural-sounding error-free sentences in a language other than their own, this is not always the case. Therefore, the end user of the data must spend time proofreading sentences.

There Can Be Malicious Members or Uncooperative Members

From time to time, there are members who do malicious things such as (1) tagging lots of "safe" sentences as "not for safe search," (2) changing correct sentences to something incorrect, or (3) provoking arguments when commenting on sentences which eventually causes productive members to leave the project.

There are also times when members who are over-confident in their own non-native language ability refuse to make corrections suggested by native speakers. This results in errors that could otherwise be eliminated remaining in the database.

7. A Problem I Have Not Yet Dealt With

There is an over-abundance, of old-fashioned language usage and usage of certain not-so-frequently-used English phrases. This is possibly due to the fact that students who produced many of the original English-Japanese sentence pairs used commercial entrance exam study books to help create their sentences. While English native speakers understand such sentences, we do not normally use them with the same frequency they occur in this corpus.

In the future, I plan to go through the sentences and eliminate a certain number of these in order to get a more balanced corpus of sentences to use with my projects.

8. Japanese-English Parallel Corpus

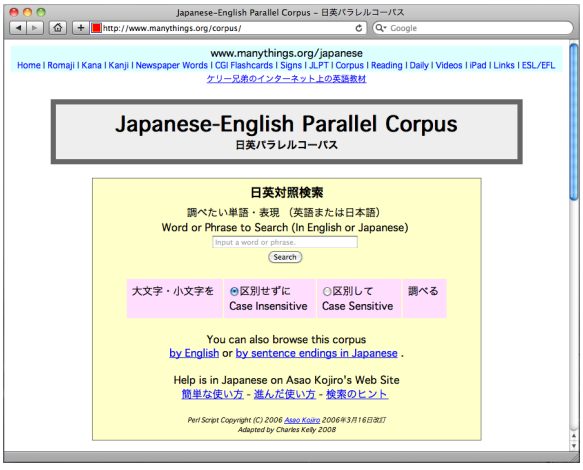

Figure 1. www.manythings.org/corpus

Early in 2008, I started using the Tanaka Corpus as maintained by Jim Breen using a Perl script by Kojiro Asao that I adapted for this project.

To make this search engine useful for English study, sentences are sorted in the following manner.

(1) All sentences are sorted by length, with the shortest sentences first.
(2) The sentences that have been personally proofread are displayed first.
(3) These are followed by other sentences by English native speakers.
(4) Finally all other non-blacklisted English sentences are displayed.

9. Japanese Reading Practice

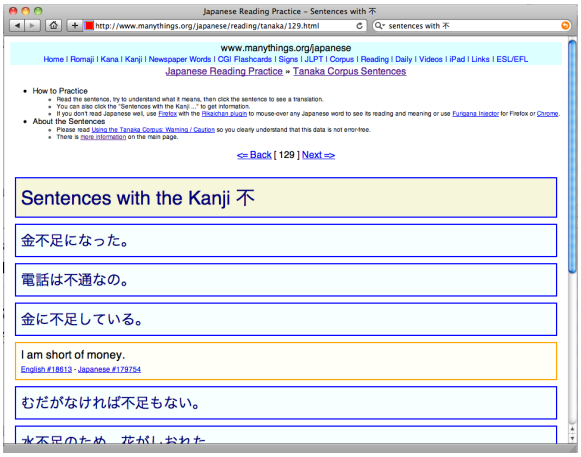

Figure 2. www.manythings.org/japanese/reading/tanaka

In 2010, using sentences from the Tatoeba Project, I created a set of pages to help people learn to read Japanese. I sorted the sentences according to what kanji is contained in them, starting with kanji found on the Japanese Language Proficiency Test (JLPT) Level 5 up through Level 2, which is similar to the *kyoiku kanji* learned in elementary schools in Japan. Next are the JLPT Level 1 kanji and beyond. The kanji are arranged in order of their frequency of use.

I limited the length of sentences to those with more than three and less than 63 characters. I included up to 50 items per page, starting with the shortest sentences.

Items are not included on a page if they contain kanji that have not yet been introduced. Additionally, each page has the explanation of the kanji from Jim Been's KANJIDIC file.

10. English High Frequency Words with Example Sentences

Figure 3. www.manythings.org/ejs

Figure 4. www.manythings.org/ejs/necessary.html

This project has about 3,000 pages, each with sentences focusing on a certain word. These sentences are grouped into six sets based on word frequency, with group one having the highest frequency words. In addition to the English, Japanese translations are included. The sentences are displayed in order of their lengths, with the shortest sentences being shown first.

11. Bilingual Sentence Pairs

Figure 5. www.manythings.org/bilingual

For this project, only sentences by native speakers of the non-English language and English sentences I have personally proofread have been included. I only included the 16 languages that had over 2,000 matches with trusted English sentences. At the time of publication of this paper (March 2012), there were 18,954 trusted bilingual English-Japanese sentence pairs. The highest number of trusted sentence pairs was 54,643 in the English-Turkish section.

For each set of language pairs, the sentences were first sorted by the length of the English sentence and then put into groups of 40 pairs. There are three ways to view each set of pairs.

(1) Sentences can be browsed while viewing both languages with the option to listen to the audio of the English sentence if it is one of the over 10,000 sentences that I recorded (Figure 6).

(2) English sentences can be seen with the other language being hidden (Figure 7).

(3) The other language can be seen with the English being hidden (Figure 8).

For the two options with hidden languages, the user only needs to click to reveal the hidden language, making it a flashcard-like activity.

Figure 6. www.manythings.org/bilingual/jpn/50.html

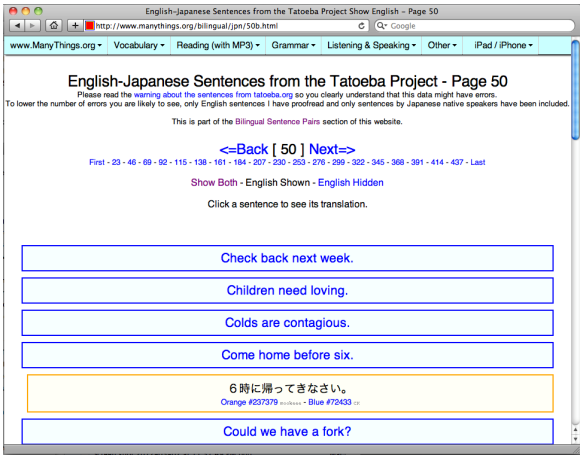

Figure 7. www.manythings.org/bilingual/jpn/50b.html

Figure 8. www.manythings.org/bilingual/jpn/50c.html

12. Tab-delimited Bilingual Sentences

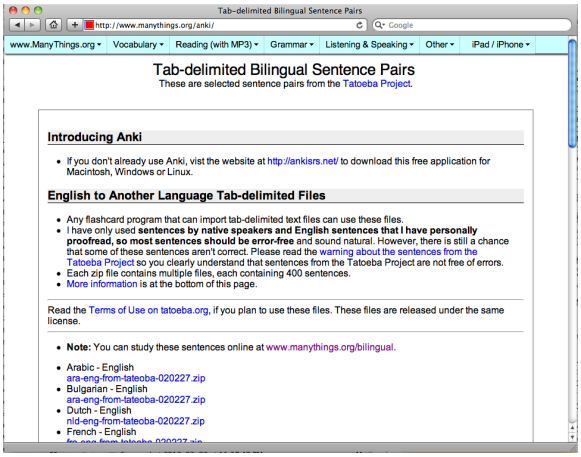

Figure 9. http://www.manythings.org/anki

This project uses the exact same data that I put together for the "Bilingual Sentence Pairs" project and makes it available for students to use in applications such as the well-known free Anki spaced-repetition flashcard application.

13. Daily Pronunciation Practice

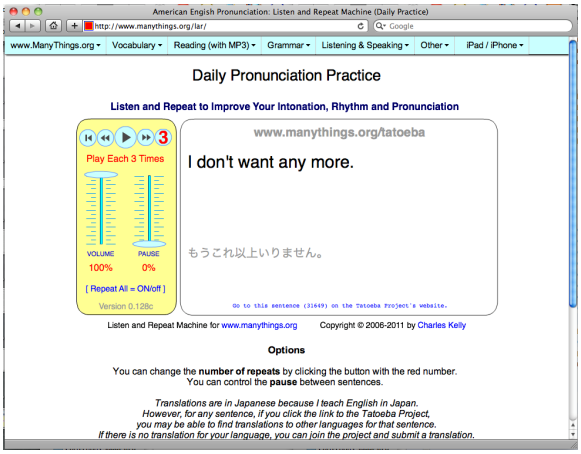

Figure 10. www.manythings.org/lar

In December of 2011, after I made thousands of the recordings for the Tatoeba Project, I started using those sentences in my "Daily Pronunciation Practice" project that had been online since 2006. I adapted my existing "Listen and Repeat" flash application to also include a direct link to tatoeba.org for each sentence.

14. English Sentences with Audio

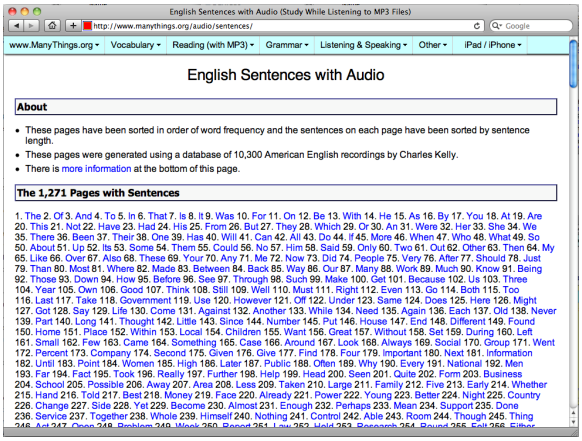

Figure 11. www.manythings.org/audio/sentences

Figure 12. www.manythings.org/audio/sentences/742.html

This section of www.manythings.org has over 2,000 pages with each page focusing on one word. This is somewhat similar to the "English High Frequency Words with Example Sentences." However, this set of pages is limited to only displaying sentences that have audio. Also, this set of pages allows the user to browse through the pages in word-frequency order. Students can also click the [T] link to jump directly to tatoeba.org to see a translation when they need one.

15. Search Sentences

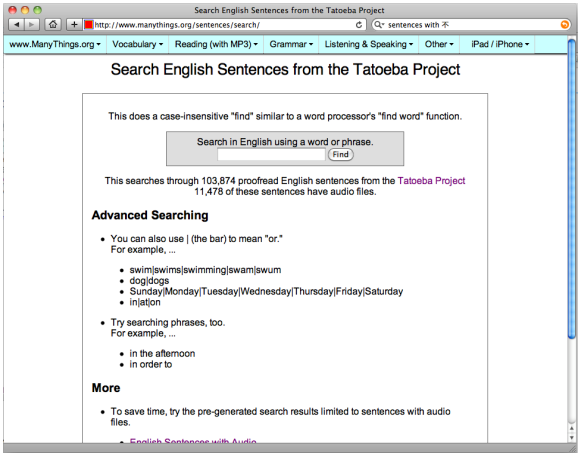

Figure 13. www.manythings.org/sentences/search

This search engine allows students and teachers to search through the trusted English sentences to find sentences that contain given words or phrases.

The search results are sorted by sentence length, include an audio player for sentences that have been recorded and include direct links to the sentences on tatoeba.org.

This search engine is also set up to allow advanced searching. The next figure shows an advanced search for “swim, swims, swimming, swam or swum.”

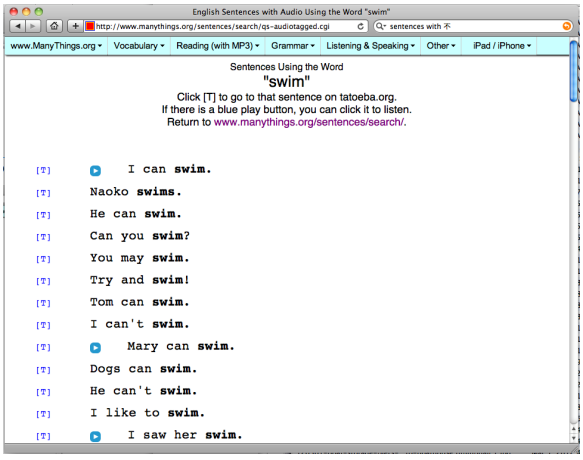

Figure 14. Search Results using swim|swims|swimming|swam|swum

16. Conclusion

Using sentences from the database maintained by the Tatoeba Project has allowed me to create several useful learning tools on the Web. These projects include a searchable corpus of English-Japanese sentence equivalents, a searchable corpus of English sentences, a set of pages to help people study Japanese reading, English sentences with audio files, and sets of bilingual sentence pairs in various languages which can be browsed or used in a flashcard-like manner. Rather than using the complete raw data from the Tatoeba Project, my projects were created after carefully selecting which sentences to use. This means that the resulting projects can be trusted and useful for language study. The Tatoeba Project has been, and will continue to be, a valuable source of material for language learners and teachers around the world.

References

1) Breen, Jim (1992-2012). KANJIDIC, http://www.csse.monash.edu.au/~jwb/kanjidic_doc.html
2) Breen, Jim (2012). Tanaka Corpus Wiki, Retrieved January 29, 2012 from http://www.edrdg.org/wiki/index.php/Tanaka_Corpus
3) Breen, Jim (1997-2012). WWWJDIC, http://www.csse.monash.edu.au/~jwb/cgi-bin/wwwjdic.cgi?1C
4) Elmes, Damien (2009-2012). Anki, http://ankisrs.net/
5) Ho, Trang (2009-2011). Tatoeba Project Blog, http://blog.tatoeba.org
6) Ho, Trang (2009-2012). Tatoeba Project, http://tatoeba.org
7) Simon, Allan (2010). Tatoeba.org, base de données de phrases d'exemple, Retrieved January 29, 2012 from http://linuxfr.org/news/tatoebaorg-base-de-données-de-phrases-dexemple
8) Tanaka, Yasuhito (2002). Compilation of a Multilingual Parallel Corpus, http://www.edrdg.org/projects/tanaka/tanaka.pdf

Appendix: Project URLs

• www.manythings.org/corpus
Japanese-English Parallel Corpus
• www.manythings.org/japanese/reading/tanaka
Japanese Reading Practice
• www.manythings.org/ejs
High Frequency Words with Example Sentences
• www.manythings.org/bilingual
Bilingual Sentence Pairs
• www.manythings.org/lar
Daily Pronunciation Practice
• www.manythings.org/audio/sentences
English Sentences with Audio

• www.manythings.org/sentences/search
Search English Sentences

（平成 24 年 3 月 19 日）