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

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.

- 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 timeconsuming step of submitting an alternate entry.

- 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 overconfident 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-frequentlyused 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

O O Japanese Reading Practice - Sentences with 不
< ト 🍈 + ■http://www.manythings.org/Japanese/reading/tanaka/129.html 🗸 Q= sentences with 不 📀
www.manyhings.org/spanses Home I Ronaji I Kara I Karaji Nevepare Wood (GE) Hanzarasi I Sigri L API Corput Bending Daily I Videos I Pad I Links I EBUER, Japanese Reading Transition - Transite Corput Sentences • How to Practice • Name Hamming Transition - Transite Corput Sentences • Name Hamming Transition - Transite Corput Sentences • Name Hamming Transition -
Sentences with the Kanji 不
金不足になった。
電話は不通なの。
金に不足している。
I am short of money. English #18813- Juansee #178754
むだがなければ不足もない。
*不足のため おおしおわた

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

00		English-Japanese S	entences - Ind	ax Page (英和	(対訳)			
▶ 🙆 🕂 <mark>■</mark> http:/	(/www.manythings	i.org/ejs/		¢	Q* sentence	s with 不		
w.ManyThings.org -	Vocabulary -	Reading (with MP3) -	Grammar -	Listening 8	Speaking -	Other -	iPad / iPhone	•
	Engli	sh-Japanese	manythings.org/o Sentend 英和対訳	;⊧ ces - In	dex Pag	e		
Please n If you don't read Japanese	ead the warning ab a well, use Firefox	out using the Tanaka Con with the Rikalchan plugin	pus so you clea to mouse-over a	rly understand iny Japanese	that this data i word to see its	s not free of reading and	errors. meaning or use Fu	rigana
		Injector f	or Firefox or Chi	ome.				
	High	Frequency Wor	de with F	vamnle	Sontonce			
	nigiri	requency wor	us with E	xample	Sentence	55		
	1	Page 1 - Page 2 - Pag	ge 3 - Page 4	- Page 5 - I	Page 6			
а	exar	nple r	me		side			
able	expe	rience r	mean		similar			
about	eyes	e r	neans		since			
above	face	r	nembers		single			
according	fact	r	nen		situation			
act	fami	ly r	method		small			
action	far	r	might		50			
activity	fathe	ar r	mind		social			
after	few	r	model		society			
again	field	r	money		some			
against	fig	r	more		something			
age	figur	e r	most		sometimes			
air	find	r	mother		south			
all	first	r	movement		special			



00		English-Japanese S	entences with	'necessary' (英和対訳)			
◄ ► 🙆 + ■ htt	p://www.manythir	igs.org/ejs/necessary.html				C Q- Google	
www.ManyThings.org *	Vocabulary •	Reading (with MP3) -	Grammar •	Listening & Speaking -	Other -	iPad / iPhone +	ſ
Disc	on mod the	Englisi Sentences Found	With "ne 227 Show	tences BCESSARY" rn: 200	a in and from	of open	
lt is n それ I'll cor 必要	ecessary. には必要だ。 meifnecess ^{変なら私は来ま}	12130 12204887 ary. 1318267 중구 185446	orpus so you cie	earry understand that this dat	a is not mee	or errors.	
Food 食生 食物 He m	is necessar 活は生存に必 すは生命に必要 akes necess	y for life. 1268756 公要である。 1145804 見だ。 1145775 Sarv changes. 131	8272				
彼は Prom 早愈	k必要な変更を pt action is n Aな処置が必要	た加える。 ¹⁹⁵⁴⁴¹ ecessary. 1274062 見だ。 1140510					
it's ne 君は	cessary for に行かなけれに	you to go. 〒16160 ばならない。 〒177310					
Salt is 塩は	s necessary 料理人にとっ	for a cook. <u>*25746</u> って必要なものだ。	7188601				
It is n	ecessary for	you to go. 716156					

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

• F 😨 🕂	http://www.manythin	igs.org/bilingual/		Reader C	Q∗ Google			
ww.ManyThings.org	 Vocabulary - 	Reading (with MP3) -	Grammar -	Listening & S	peaking -	Other -	iPad / iPhone -	
		Bilingual	Sentenc	e Pairs				
		Diniguai	Contonio	orano				
Bilingu	al Sentence Pair	s - About the Tatoeba	Project - Wan	ning / Caution	- Japanes	e Readin	g Practice	
Bilingual S	Sentence Pairs							
 English 	- Arabic							
 English 	- Bulgarian							
 English 	- Dutch							
English English	- French							
English	- Hebrew							
English	- Hindi							
 English 	- Icelandic							
 English 	- Italian							
 English 	i - Japanese							
 English 	- Polish							
 English 	- Portuguese							
English	- Russian							
English	- Openish - Turkieb							
English	- Ukrainian							
 If you u 	se Anki or a simil	ar flashcard program, y	you can <mark>down</mark>	load this data				
About The	se Sentences							
This na	rt of our website	uses sentences from th	e Tatoeba Pr	niect (Read A	hout the T	atoeba P	miect)	
 Please 	read the warning	about the sentences f	rom the Tatoe	ba Project so	vou clearl	v understa	and that	
senten	ces from the Tato	eba Project are not free	e of errors.			,		
 I have d 	only used senten	ces by native speake	rs and Engli	sh sentences	that I hav	e persor	ally	
proofre	ead, so most ser	tences should be en	ror-free and s	ound natural.	However,	there is st	ill a chance	
that sor	ne of these sente	nces aren't correct.						
 I have a 	also attempted to	titter out sentences that	t are potentia	ly offensive a	ng not app	ropriate f	or all ages	

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.

- 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

000	English	-Japanese Sentences from	m the Tatoeba	Project Show	English - Pag	e 50		
	p://www.manythin	igs.org/bilingual/jpn/50b.h	itml	c	Q* Google			_
www.ManyThings.org -	Vocabulary -	Reading (with MP3) -	Grammar •	Listening &	Speaking -	Other -	iPad / iPhone -	
Englis Please n To lower the number of errors	sh-Japane ead the warning at you are likely to s	ese Sentences out the sentences from tato ee, only English sentences This is part of the Bilingual	from the beba.org so you I have proofrea Sentence Pairs	e Tatoe clearly unders d and only ser section of thi ext=>	ba Proje tand that this d itences by Japi s website.	ect - Pi ata might ha anese native	age 50 ave errors. • speakers have been inv	cluded
First	- 23 - 46 - 69 - 92	- 115 - 138 - 161 - 184 - 207 Show Both - Engli	- 230 - 253 - 27	6 - 299 - 322 - English Hid	345 - 368 - 39	1 - 414 - 437	- Last	
		Onlow Doin - Engi	ian onown - i	-ngilan nu	2011			
		Glick a senter	ice to see its	uansiauon.				
								-
		Check b	back next	week.				
		Childre	en need lo	ving.				
		Colds a	are contag	ious.				
		Come h	ome befo	re six.				
		6時にり Orange #237	帚ってきな 379 Blu	さい。 • •72433 cm				
		Could w	ve have a	fork?				

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.



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

	httr	. (/www.manythir	English Sentences w	ith Audio Usin	g the Word "Speech"			
www.ManyThing	s.org •	Vocabulary •	Reading (with MP3) •	Grammar •	Listening & Speaking -	Other •	iPad / iPhone •	n
		т	Read and Listen T this page is part of www Click [T] to go to th Click the blu Click the blu Click the blu Click the blu Click the blu Click the blu	o Sentences Speech" manythings nat sentence le play buttor ([742] N andom Page	Using the Word .org/audio/sentences. on tatoeba.org. to listen. ext =>			
с (П) с (П) с (П) с (П)	His s He d How She	s peech mo lelivered a did your s made a go	ved us. speech . peech go? od speech .					
	Tom Do y	's speech to	was excellent. make a speech ? pired all the bour	?				
о (1) о (1)	I was	s deeply ma	oved by his spee pressed us very n	ch . nuch.				
о П о П	Nobo You	ody was lis don't need	tening to the spe to prepare a form	ech. nal speec	h.) 4 4

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.

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."

000		English Sentences v	vith Audio Usir	g the Word "	swim"			
◄ ► 🙆 + ■ htt	p://www.manythir	ngs.org/sentences/search/q	s-audiotagged.	cgi C	Q+ sentence	is with 不		0
www.ManyThings.org •	Vocabulary •	Reading (with MP3) •	Grammar •	Listening 8	Speaking +	Other -	iPad / iPhone •	
		Sentenc	es Using the	Word				
			"swim"					
	If	Click [1] to go to th there is a blue play I	at sentence	on tatoeb can click it	a.org. to listen.			•
		Return to www.many	things.org/s	entences/	search/.			
								1
(T) 🖸	I can	swim.						
[T] Na	oko swim	s.						
[T] He	can swi	m.						
[T] Ca	n you sw	im?						
[т] Үо	u may sw	im.						
(T) Tr	y and sw	im!						
(T) TO	m can sw	im.						
(T) I	can't sw	im.						
(T) 🖸	Mary o	can swim.						
[T] Do	gs can s	wim.						
[т] Не	can't s	wim.						
[T] I	like to	swim.						
(T) 🖸	I saw	her swim.						×.

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

- Breen, Jim (1992-2012). KANJIDIC, http://www.csse.monash.edu.au/~jwb/kanjidic_d oc.html
- Breen, Jim (2012). Tanaka Corpus Wiki, Retrieved January 29, 2012 from http://www.edrdg.org/wiki/index.php/Tanaka_C orpus
- Breen, Jim (1997-2012). WWWJDIC, http://www.csse.monash.edu.au/~jwb/cgibin/wwwjdic.cgi?1C
- 4) Elmes, Damien (2009-2012). Anki, http://ankisrs.net/
- Ho, Trang (2009-2011). Tatoeba Project Blog, http://blog.tatoeba.org
- 6) Ho, Trang (2009-2012). Tatoeba Project, http://tatoeba.org
- Simon, Allan (2010). Tatoeba.org, base de données de phrases d'exemple, Retrieved January 29, 2012 from http://linuxfr.org/news/tatoebaorg-base-dedonnées-de-phrases-dexemple
- 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

15. Search Sentences

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

(平成 24 年 3 月 19 日)