Word Usage Examples in an Electronic Dictionary

J.W. Breen
School of Computer Science & Software Engineering
Monash University
jwb@csse.monash.edu.au

[Note: This paper was written for the Papillon workshop in Sapporo in 2003. It has been slightly updated since then to reflect modifications in the operation of the examples in the WWWJDIC server, and to correct some out-of-date URLs.]

Abstract

This paper describes a project in which the Tanaka corpus of matched Japanese-English sentence pairs has been linked to the WWWJDIC online Japanese-English dictionary. The process of linking the corpus is described in detail, as well as an analysis of the word coverage, and the editing of the corpus to remove some of the errors it contains. The paper concludes that the Tanaka corpus can successfully provide a source of example sentences for a Japanese-English dictionary

1. Dictionary Examples and the Electronic Corpus

The practice of incorporating sentences or sentence fragments as part of a dictionary entry appears to have originated with Latin and Greek dictionaries compiled in the 16th and 17th centuries, where such passages served as citations from classical texts establishing the provenance of the words. The incorporation of such citations was adopted in later major English dictionaries by lexicographers such as Johnson and Webster, and is now regarded as an essential feature of authoritative mono-lingual dictionaries.

The development of comprehensive bilingual dictionaries from the mid-19th century, and more recently mono-lingual "learner's dictionaries", extended this practice to include selected or composed examples illustrating the usage of the words. Such examples are considered to be an essential component of such dictionaries. In one English-Japanese dictionary [1] in the author's possession, the body of each entry consists entirely of parallel English and Japanese sentences utilizing the headwords.

The development of extensive electronic corpora such as COBUILD [2] and BNC [3] has brought corpus linguistics to a prominent position in lexicography. In the context of learners' or bilingual dictionaries, such corpora tend to be used as an aid to the construction of examples, rather than as a direct source. Landau [4] comments that "What a corpus can do above all else - even when it cannot provide verbatim examples that can be used in a dictionary - is to give examples at the right level of complexity and in a framework that is typical so that the lexicographer can devise examples that are not silly, stilted, or clearly artificial." One of the editors of *Taishukan's Unabridged Genius English-Japanese Dictionary*, Kosei Minamide [5], writing about corpora and the examples used in that dictionary, states "Such corpora is (sic) liable to drown us in data", and adds "Because of the complicated problems concerning copyright and the extreme difficulty of finding entirely suitable examples in the corpus, we had most of the illustrative examples invented by native speakers."

There are no reported cases of electronic corpora being used directly for the provision of dictionary examples. The difficulty of using such corpora for this purpose can be seen from examination of some of the text samples from the online COBUILD collection for the word *swimming:*

```
against Douglas Stern, Doug Stern's Swimming Clinic Inc., the United States
no-touch sex with clothes on [p] swimming - especially nude [p] smiling [p]
induced cloud or magical blackness swimming in the air; it was simply
likely to keep busy playing games, swimming, jeeping, or making crafts such as
historic feat of winning Olympic swimming medals 12 years apart. Janet
```

```
Silk Cup Derby (Hickstead) 1435b Swimming: National Champs & Euro Trials (
End. The quieter spots and the best swimming on one-mile Long Bay beach are at
suitable physical exercise such as swimming or cycling. He will find that any
in such a way that you feel you are swimming outdoors in an open-air pavilion.
sun-splashed conservatory - even a swimming pool. An unforgettably exotic or
and telephone. There's an indoor swimming-pool, sauna, solarium and
```

Clearly only one or two extracts in this sample contain useful material for example sentences, and in both cases some rewriting would be appropriate. It is only one sample, but it supports the views of Landau and Minamide.

2. Project Background

When the author began compiling a Japanese-English dictionary file as part of the EDICT [6] project in 1991, there were immediate calls from users of the file and software for example sentences to be associated with the dictionary entries. The initial dictionary format file did not readily allow for the inclusion of such examples, so a structure for such examples was implemented, involving a simple marker in the text of the English translation which indicated the availability of further explanatory information and examples in a linked adjunct file. As the early stages of the EDICT project benefited from considerable voluntary effort, a call was made for the preparation and submission of examples and other explanatory material. None was forthcoming; it appeared that while the user community had sufficient interest and enthusiasm to submit lexical material, preparation of examples was not such a high priority.

In 1999 the JMdict project, which involved an expanded dictionary structure, was launched. From the beginning of the project it was intended to incorporate example sentences within entries, with elements reserved in the DTD for this purpose.

```
<!ELEMENT sense (stagk*, stagr*, xref*, ant*, gram*, field*, misc*, gloss*, example*, s_inf*)>
.....
<!ELEMENT example (#PCDATA)>
```

Although the structure allowed for examples, there was no ready source which could be employed, and no voluntary contributions were forthcoming.

3. The Tanaka Corpus

As reported at the PACLING2001 conference in a paper on the compilation of multilingual corpora [7], Professor Yasuhito Tanaka at Hyogo University had assembled over several years a collection of over 200,000 Japanese-English sentence pairs. The technique he employed was to encourage a number of students each to enter approximately 300 items, drawn from instructional texts and other available sources. The resulting corpus, which he stated was in need of considerable editing, was placed in the Public Domain. At the 2002 Papillon Workshop, Professor Christian Boitet provided a copy of the corpus to participants, with a view to it possibly being used as the foundation for a set of examples within the Papillon dictionary project.

The author examined the corpus and concluded that it did indeed have excellent potential for providing such examples, but that it also had a large number of errors which would need eventual correction. It was decided to conduct a trial in which the corpus would be used to provide usage examples for entries in the author's WWW Japanese-English dictionary server (WWWJDIC). [8]

The broad purpose of the trial was:

a. to determine if such a sentence collection could effectively be used to provide example sentences in an electronic dictionary application. This of course extends into such matters as:
   i. experimentation with techniques for achieving the integration of a dictionary and a corpus

of example sentences;
   ii. detection and resolution of related problems, such as the capability to handle issues of polysemy and homonymy.
b. to determine if the Tanaka corpus could be edited to an adequate standard in a timely and cost-effective manner.

4. Initial Processing

As provided, the corpus was a text file with alternating Japanese and English sentences. After code conversion, the sentence pairs were aggregated into tab-delimited single lines to aid sorting and inspection. It was immediately apparent that there were a large number of duplicate or near-duplicate pairs, differing only by such things as punctuation, or spelling errors in the English portion.

After some simple harmonization of the punctuation, mainly consisting of ensuring that the punctuation in the Japanese sentences used "JIS" characters, and in the English sentences used ASCII characters, occurrences of examples which duplicated another example with regard to the Japanese sentence were removed. Whilst this may on occasions have removed an example with a "correct" English sentence in favour of an incorrect sentence, it was considered that this could eventually be corrected at a later stage.

The removal of this type of duplicated example reduced the file from an initial 203,000 sentence pairs to approximately 183,000. Further inspection at this stage revealed that a considerable number of errors and near-duplicates remained, however it was considered that the file was in a state that permitted at least a trial of its application to the role of providing example sentences for a dictionary. Further editing could, and did, take place in parallel with the implementation of the dictionary association.

5. Linking Examples to Dictionary Entries

The process of associating example sentences with dictionary entries, had it followed the same approach as with printed dictionaries (which was also the approach allowed for in the JMdict data structure), would have meant selecting one or two sentence pairs for each of approximately 20,000 words, and embedding them in the appropriate part of the dictionary database. This approach clearly has a number of problems:

a. it would inevitably limit the number of examples available for each word, when the corpus often contained a much larger number;
b. it would lead to the breaking-up of the corpus;
c. it would significantly increase the size of the dictionary file. Not all applications of the file can, or would, use the examples;
d. the process of selecting, editing and moving the example pairs would be very large.

Instead, an approach was adopted that achieved the same effect, i.e. the association of examples with dictionary entries, but which avoided the problems outlined above. The approach involved:

a. leaving the corpus intact, thus enabling continued editing and revision;
b. establishing dynamic links as required from dictionary entries to the sentence(s) that contained the entries' head-words.

Given the size of the file, it was not considered efficient to search it each time a link was required. Also the fact that many of the words involved were verbs, adjectives, etc., which often appeared in the sentences in inflected forms, would greatly complicate such a search. In order to expedite the linking process, it was decided to pre-process the sentences to identify the target words within the examples that could be used to attract links from the dictionary entries. Thus each example in the corpus would be extended so that it consisted of the triplet: (Japanese-sentence, English-sentence, word-list).

The extraction of the words in each sentence was carried out initially using the Chasen [9] morphological analyzer from Nara Institute of Science and Technology (NAIST). Each sentence was passed through the Chasen program, and the extracted words which contained at least one kanji were retained. Using a package such as Chasen had the advantage of bringing most of the inflected forms of words back to the plain (dictionary) form, and also of accurately segmenting the text so that trailing *okurigana,* etc. were retained. Extracted kana-only words were not retained initially, as for the most part they consisted of particles, conjunctions, etc. which have little relevance to the dictionary entries. It was recognized that a number of words which are always or often written with kana could end up being overlooked, but it was considered that they could be revisited at a later stage. *(See section 7 below.)*

An additional analysis was carried out to extract all sequences of katakana from the sentences, on the assumption that these would typically be loan-words.

The examples in the extended corpus were thus converted into the following format:

A: 後ろのドアを閉めてください。[TAB]Please shut the door behind you.

B: 後ろ 閉める ドア

The process described above identified approximately 660,000 word occurrences in the 177,500 sentences in the corpus at the time of writing, i.e. a mean of 3.7 words per sentence. In total approximately 23,000 unique words were identified. Of the unique words, approximately 3,500 do not occur as head-words in the JMdict/EDICT dictionary files. On inspection these words are for the most part proper names or verbs in the potential form (see below).

The frequency distribution of words is given in the following table.

No. of occurrences	**No. of words**	**Examples**
1	8,025	アーカイブ, グラム, 愛憎, 学外
2	3,132	アンコール, レシート, 塩梅, 信託
3	1,803	エラー, ブーム, 区役所, 標本
4	1,243	キロメートル, バスタオル, 色白, 抜歯
5	828	カトリック, テレビ局, 加工, 行く手
6-10	2,393	ヒーター, 無用, チキン, 馬車
11-20	1,884	パンフレット, 亡くす, 国籍, 服従
21-30	748	トマト, レッスン, 乗り換える, 火災
31-100	1,653	アドバイス, 稼ぐ, 協力, 前もって
101-500	943	ゲーム, 案内, 殺人, 事業
501-1000	109	ドア, 美しい, 息子, 降る
1000+	67	手紙, 電話, 問題, 彼女

A significant number of the infrequently-used words are proper names.

To enable the association of the example sentences with dictionary software, and the subsequent display of examples for a given word, an ancillary word-sentence index file was created and inverted. For example for the word 加工 it contains:

加工 4508470 4592547 4592637 5636146 6947087

where the integers are the byte-offsets in the corpus of the sentences containing 加工. (The WWWJDIC server has the file mounted as a read-only text file and "seeks" to the selected sentences.)

The integration of the corpus into the WWWJDIC server was carried out as follows:

a. as each entry in the main (EDICT) dictionary is displayed, the headword is checked against the corpus index. If the headword occurs in the index, a hyperlink is added to the display of the entry indicating there are example sentences available (the [Ex] at the end of the 加工 entry in the following example.) The URL of the hyperlink carries the headword as a parameter.

b. If a user selects the example hyperlink, the server:
    i. displays all the example sentence pairs, if there are 10 or fewer available;
    ii. otherwise displays a random selection of 10 example sentence pairs, and allows the user the options of viewing another selection, or viewing the complete set in batches of 100.

6. Corpus Problems

It is apparent that there are a number of problems with the Tanaka corpus that need to be considered when using it as a source of dictionary-related examples.

The first is that a number of the sentences are such things as short interjections, proverbs, quotations, aphorisms, etc. which while they are of interest, are not necessarily useful in the context of showing typical usage of words. Some examples of these are:

よ、ポール。 Hey, Paul.
「転ばぬ先の杖」はことわざである。"A stitch in time saves nine" is a proverb.
きょうの一針あすの十針。 A stitch in time saves nine.
なんだ、またか？ Oh, Jesus, another one.
ナザレの人で、ヨセフの子イエスです。 Jesus of Nazareth, the son of Joseph.
己の欲せざる所は人に施す勿れ。Do to others as you would have others do to you.

Fortunately the nature of most of these is obvious, but at some stage it may be useful to tag them as quotations, etc. lest the incautious learner be misled by them.

A more serious problem is presented by the presence in the sentence pairs of:

a. errors (spelling, grammar, etc.) in the English and Japanese sentences;

    Errors in the Japanese sentences often comprise:

    i. incorrect selection of a *jukugo* (kanji word) when writing the sentence:

        私は信心深い男で、死後の生命の存在を信じています。 I'm a religious man and believe in life after death.
        私は信心深い男で、私語の生命の存在を信じています。 I'm a religious man and believe in life after death.
        (死後 and 私語 are both pronounced しご, but the latter means "whispering; secret talk")

ＡＩは人口知能の略です。 AI stands for artificial intelligence.
(人口 is incorrect. The word should be 人工 - also pronounced じんこう)

ＤＮＡのサンプルを畜えることは許されるべきではない。 Storing DNA samples should not be permitted.
ＤＮＡのサンプルを蓄えることは許されるべきではない。 Storing DNA samples should not be permitted.
(畜える is clearly a typing error for 蓄える)

ii. errors in the kana:

ＤＮＡのサンプルを蓄えることは許されるべきでわない。
(でわない should be ではない.)

Many of these are being detected in near-duplicate sentences, and others are being found as well. The file will obviously benefit from being thoroughly proof-read by Japanese native speakers.

Errors in the English sentences, apart from the mistranslations discussed below, largely consist of spelling errors, faulty capitalization, and incorrect punctuation. Many of the spelling errors were removed by carrying out a spell-check on the file, however cases are still being found where incorrect words have been used.

首に湿疹ができました。 I have a rush on my neck.

コップが地面に落ちて砕けた。 The glass clashed to the ground.

Again, a thorough proof-reading would be required to detect and remove all these errors.

b. actual or near-duplicate sentences;

While a large number of duplications have been removed, a considerable number remain. These are due to such things as:

i. residual punctuation differences:

近ごろはいかがお暮らしですか。 How are you getting along these days.
近ごろはいかがお暮らしですか? How are you getting along these days?

ii. orthographical variations, typically resulting from words being written using both kana and kanji, or using equivalent kanji:

近頃彼にほとんどあわない。 I have seen little of him of late.
近頃彼にほとんど会わない。 I have seen little of him of late.

部屋は兎小屋みたいだけど。 But my place is like a rabbit hutch.
きつねの尾はウサギのより長い。 The tail of a fox is longer than that of a rabbit.

iii. differences of register, e.g. use of plain or polite verb forms:

金持ちが必ずしも幸福であるとは限らない。 The rich are not always happy.
金持ちが必ずしも幸福であるとは限りません。 The rich are not always happy.

iv. presence or absence of emphases, such as よ/わ, or of gender-specific forms, e.g. の.

幸運を祈る。 Good luck!
幸運を祈るよ。 Good luck!

すごいよ。 It's incredible.
すごいわ。 That's wonderful.

なぜいけないか。 Why not?
なぜいけないの。 Why not?

v. other small textual variations, such as the use of へ and に particles, which do not affect the meaning of the sentence or its translation.

In extreme cases sets of up to 130 such near-duplicate sentences have been detected.

Many of these cases can be detected by scanning the file with the sentences sorted by either the Japanese or English sentence. The approach being adopted is to eliminate the punctuation variations, retain the sentence form which makes the greatest use of kanji (as this will lead to more examples being available) and aim for a mix of register types, emphases, etc. across the sentence collection.

Another approach is to use a measure of similarity between sentences, based on the list of Japanese words in each sentence. Sentences with identical word-sets would be candidates for examination and possible reduction.

Editing of the sentence collection is still being carried out, with approximately 6,000 near-duplicates removed so far.

c. mismatched and mistranslated sentences.

A number of sentences have obviously had the English component derive from a literal translation of the Japanese with little regard to the validity of the result:

紅に染まった俺のこの傷を癒す奴はいない。 My heart has been gonna dye deep red with all op pain.

おまえを失いかけた時、俺は自分の汚れた心を見た。 When I was gonna be losing you on my mind found my heart in soil.

In other cases, the English, while correct grammatically, does not mean the same as the Japanese, perhaps because there has been some truncation:

彼は昨日アリスに合ったといったがそんな訳はない。 He said he met Alice yesterday but it cannot be true because she left for London a week ago.

7. Parsing Problems

As mentioned above, the Chasen morphological analysis package was used to extract target words from the Japanese sentences. In general this process was carried out very successfully, however in a number of cases either incorrect or inappropriate segmentation of the text occurred.

a. incorrect segmentation. In some cases incorrect choices of word-boundary were made:

娘達は父親の死のショックから元気を取り戻した。 The daughters recuperated from the shock of death of their father.

娘 達 父親 死 から元気 取り戻す ショック
(から元気 [空元気] is incorrect in this context)

b. inappropriate segmentation. The analysis software tended to break up compounds which for dictionary purposes would be better retained. This particularly applied to suffixes such as 者 and 的

君は正直者のようだ。 You seem an honest man.
君 正直 者

私は経済的に両親からひとり立ちしている。 I am economically independent of my parents.
私 経済 的 両親 ひとり立ち

In general this is not a major problem as the components themselves are usually dictionary entries, however it is appropriate to aggregate at least some of these when they are detected in order to increase the number of usable examples. In practice a relatively simple solution was available. An examination was made of adjacent pairs of words in the indices to determine if they were both continuous in the associated sentence and present as a headword in the dictionary file. If both conditions were met, the pair of words was joined. This process resulted 10,600 word pairs being joined, giving approximately 2,800 new unique index words.

c. generation of non-dictionary forms. In a number of cases Chasen does not generate the dictionary form of verbs. In particular this can be seen with verbs used in the potential or causative forms.

私の息子は時計が読めます。 My son can read the clock.
私 息子 時計 読める

These will need to be identified and corrected, as the JMdict/EDICT dictionary files, in common with most printed dictionaries, do not usually carry these inflected forms as separate entries.

d. absence of kana-only words.

As discussed above, kana-only words were not included from the original analysis as it would have been difficult to separate useful words from conjunctions, particles, etc. However a number of common words are usually written in kana alone, and it would be useful to be able to associate examples with them.

As the linkage employed in WWWJDIC uses the initial field in an entry, which usually has the kanji form of the word, the approach that has been followed is to add that form to the word list, leaving the kana form in the sentence. Several hundred such words such as 一寸 (ちょっと), 迚も (とても), etc. as well as many words usually only written in kana such as ずっと, けど, どうぞ, どうやって, etc. were added by hand.

私はいつもテレビを見て時間を過ごす。 I always pass the time by watching TV.
私 見る 時間 過ごす テレビ 何時も

It became clear that identifying and adding kana-only words to the index lines by hand was quite time-consuming and inefficient. As foreshadowed in section 5, the issue was revisited with a programmed extension of the indices to include such words. The process applied was:

i. the example sentences were reprocessed through Chasen and the hiragana-only words identified and collected;

ii. the words were matched against the dictionary file and where they could be uniquely identified with a headword, flagged as an acceptable index word;

iii. a second processing by Chasen was carried out, this time adding the word (in its kanji form if this was available) to the index line if the word was in the "accept" file.

As a result of this process, the number of unique words covered by the indices was increased by about 1,300, and a total of approximately 70,000 were added to the indices.

8. Polysemy and Homonymy

Direct association of dictionary head-words with example sentences containing those words does not immediately cater for situations where, for example, a single *gairaigo* (loanword) has multiple meanings, or where a word has more than one sense. For example, チップ can mean both "chip" and "tip", resulting in the following example sentences being selected:

> この企業はコンピュータ・チップを製造している。 This company manufactures computer chips.
> 彼は感謝のしるしにチップを与えた。 He gave a tip as a sign of gratitude.

Similarly, for お嬢さん, which can mean both "(your) daughter" and "young lady", we see:

> お嬢さんは試験に合格なさったそうですね。 Your daughter passed the examination, I hear.
> 彼女はいささかとりすました良家のお嬢さんだった。 She was rather prim and proper young lady.

The current approach to this problem is to append the sense-number to the index words in the examples file in the cases where more than one sense exists. Thus in the cases mentioned above, the full set of example sentences will be identified and displayed, but the senses will be stated. A more ideal solution, reserved for future implementation, is to allow the dictionary user to choose the sense for which example sentences are sought. (At the time of writing the file is being edited to add the sense numbers.)

The following example for the word 汚す, which has the senses of "to disgrace" and "to dirty", illustrate the approach taken.

There is also a potential problem with homonyms. For example, the kanji: 略 can be used to write two different words: ほぼ (almost; roughly; approximately), and りゃく (abbreviation; omission). There are different dictionary entries for these words, but without some special treatment, both would link to the same set of example sentences, as the link is normally based on the kanji headword. E.g.

> ユネスコが何の略か知っていますか。 Do you know what UNESCO stands for?

Although such cases of true homonymy are relatively rare in Japanese, a solution, such as the extension of the indices to include reading, is required to avoid confusion. The approach that has been adopted is to allow for the appending of the reading to the index word, so that the linked example sentences only apply to the correct word. Thus for the sentence above, the index list comprises: 何 略(りゃく) 知る ユネスコ

A number of example sentences which include homonymous words have been manually identified and marked in this way to prevent misleading linkages occurring.

9. Corpus Subset

The file size of the corpus, including indices, is at present over 18Mb. While this is not a problem for server systems, there has been interest in having a smaller version of the corpus for use with PDA-based dictionary software.

While it would be possible to make a manual selection which included representative examples of words, this would be a major task, and would also have the disadvantage of breaking up the corpus at a time when it is still being edited. Ideally there should be a technique available which can automatically extract a suitable subset from the full corpus at any time.

One approach is to use the measure of similarity mentioned above to cull near-duplicates. As an interim step, a simple heuristic was trialled to determine the potential effectiveness of such an automatic subset generation.

The steps in the heuristic are:

- add to the front of each sentence group in the corpus the count of index words and a random number, then sort the corpus on these keys. (The random number is to force the dispersion of near-duplicates.)
- starting with sentences with 4 or fewer index words, place a copy of the sentence group in the subset if any of the index words have not yet occurred at least 5 times in sentences previously added to the subset file. (The reason for starting with sentences with 4 or fewer index words is because it is a reasonable assumption that short-medium length sentences make better examples than long sentences.)
- repeat the step above for the remaining sentences.

This process resulted in a collection of 44,800 sentence pairs, i.e. a little over 25% the size of the full file. From inspection it appears to have a reasonable coverage of the more common words, and as intended has a full coverage of the less common words. It is interesting to note that changing some of the parameters of the heuristic does not significantly alter the outcomes. For example raising the occurrence threshold from 5 to 10 increases the size of the subset file by about 20%.

Removing proper names from the lists of index words is likely to reduce the size of the subset file.

10. Current Status and Assessment

The initial integration of the Tanaka corpus into the WWWJDIC server took place in August/September 2002, and has been operational since then. Minor revisions have been made since then, e.g. the introduction of a random selection of sentences as the initial display. An option which allows users to submit comments and corrections via a feedback form is about to be released.

Editing of the corpus has continued since its integration into the server, and main server and its mirror sites have their files updated approximately weekly. The complete corpus with associated index words is available for download from the Monash site. The subset file is under consideration for inclusion in the dictionary module of the popular JWPce package.

Feedback from the WWWJDIC user community has been very positive, with many responses that the example sentences are very useful in the study of Japanese. The errors in the file do not appear to be causing undue difficulty, in fact they seem to be more often a source of amusement.

Examination of the examples displayed for a selection of words was compared with those in several printed Japanese-English dictionaries. In general the results were comparable. The WWWJDIC/Tanaka case did not usually provide examples for as many words, however for some words it provided a much wider choice of examples. The following sets from WWWJDIC and the recent Sanseido "Grand Concise" (a Japanese-English dictionary designed for the domestic Japanese market) for the word 裏側/うらがわ - the reverse; other side; lining, illustrates a typical comparison.

> **WWWJDIC**
> 裏側の部屋に替えてください。 I'd like a room in the back.
> その男はコートの裏側に何か持っていた。 The man had something under his coat.
> それに、考えてもごらんなさいよ。あなたは地球の裏側にいるのにね。」 And just think, you're on the other side of the world."
> 月の裏側は見えません。 We cannot see the other side of the moon.
> えりはまず裏側にアイロンをかけ次に表側をかけなさい。 Iron the inside of collars first, and then the outside.
>
> **Grand Concise**
> 月の裏側 the back [hidden] side of the moon
> 封筒の裏側 the reverse side of an envelope
> 人生の裏側をのぞく get a peep of life on the seamy side
> 家の裏側へ回ってください Please come around to the back of the house

There have been a number of requests for the inclusion of examples using words which are not currently in the corpus. A small number of sentences have been added, and consideration is being given a subsequent project to identify missing common words and extract suitable sentences from available corpora.

11. Conclusion

This paper describes a project in which the Tanaka corpus of matched Japanese-English sentence pairs has been linked to an online Japanese-English dictionary. It has demonstrated that the corpus is capable of serving very well as the basis for example sentences in an electronic dictionary, and has indicated several avenues for improving and expanding the corpus.

The project has also demonstrated the viability and advantages of the approach of maintaining the example corpus as a separate entity from the lexicon, and only linking the two at the time of displaying an extended entry.

References

1. 全英連 (全国英語教育団体連合会編), *高校基本英単語話用集*, Kenkyusha, 1967
2. http://titania.cobuild.collins.co.uk/form.html
3. http://www.hcu.ox.ac.uk/BNC
4. Landau, Sidney I., *Dictionaries: The Art and Craft of Lexicography, 2nd Edition,* Cambridge University Press, 2001
5. Minamide, Kosei, *English-Japanese Lexicography and the Unabridged Genius,* Kernerman Dictionary News, Number 10, July 2002. (See http://kdictionaries.com/newsletter/kdn10-5.html)
6. http://www.edrdg.org/jmdict/edict.html
7. Tanaka, Yasuhito, *Compilation of A Multilingual Parallel Corpus,* PACLING 2001, Sep 2001, Japan. (See http://www.afnlp.org/archives/pacling2001/pdf/tanaka.pdf)
8. http://www.edrdg.org/cgi-bin/wwwjdic/wwwjdic?1C
9. http://chasen-legacy.osdn.jp/