Construction of a Bilingual Dictionary
Intermediated by a Third Language

Kumiko TANAKA
Division of Engineering, University of Tokyo.
7-3-1 Hongo, Bunkyouku, Tokyo, 113, Japan.
kumiko@ipl.t.u-tokyo.ac.jp

Kyoji UMEMURA
NTT Basic Research Laboratories.
3-9-11 Midori, Musashino, Tokyo, 180, Japan.
umemura@nuesun.NTT.JP

and $Dic_{j-f}$. $Dic_{e-j}$ is called an inverse dictionary of $Dic_{j-e}$. Japanese words have information in the following format: pronunciation in romaji, and its equivalence in English. English words are written in this font and French words in this font.

2 Overview of the Method
2.1 Inverse Consultation
The most naive way to use English to obtain French words corresponding to a Japanese word is to look up the Japanese word in a $Dic_{j-e}$ and then look up the resultant English words in a $Dic_{e-f}$. The resulting French words can be regarded as equivalence candidates (ECs) of the original Japanese word. For example, in Fig. 1, ECs for a Japanese word “競争” (kyousou : competition) are compétition, concours, race etc. Among these, race and hâte are inadequate as equivalents of “競争”.

As for race, the English word race has several meanings with the same spelling: one is to compete and another is human race. It is human race which induces the inadequate EC race. As for hâte, the English word race has the wider meaning to hurry which the original Japanese word “競争” does not. Since hâte is a direct translation of to hurry, it is inappropriate as an equivalence.

The following three cases generate irrelevant ECs.
1. An English word with the same spelling but with different meanings is intermediated. (race in the above example)
2. An English word with a wider meaning than that of original Japanese word is intermediated. (hâte in the above example)
3. There are mistakes in dictionaries.
The first two cases are due to the ambiguity in English. An English word with a narrower meaning than the Japanese source may miss some French equivalents. We think that if the original word has ambiguity and several meanings, the dictionary gives the corresponding English words.

We could handle the problem of choosing equivalencies from among ECs by semantic processing, but expressing meaning of vocabulary in dictionaries is a great problem. A simpler way is to look the ECs up in the inverse dictionary. For example, one can consult $Dic_{e-j}$ for compétition, concours, and race and thereby get “競争”, “競
competition consists of two “競争”s and three “競技”s (kyougi : game). For race, the SA has “競争”, “人種”, “先祖”(senzo : ancestry). Since “競争” appears only once for race, we discard the EC race.

There can be a infinite number of inverse consultations according to the number of consulted inverse dictionaries. If the inverse dictionaries are consulted n times, we call the method n times inverse consultation (IC_n). Which inverse dictionary to use does not always have a unique answer. For IC_2 for example, we may consult Dic_e to after consulting Dic_f with the SA in French.

2.2 Selection Procedure

Once the SA for a given word is obtained, equivalencies are selected by handling two collections of words. We call this process the selection procedure.

One way to do this is to count the number of specific elements in the SA. For example, if the SA is in Japanese, the number of the element “競争” itself is counted. Another way is to count how many parts of words (PWs) are contained in the SA. For example, the number of “競” and “争” contained in the SA is considered (thus “競” in “競技” is also counted).

If we handle the meaning of words, a third way is to look up “競争” in a Japanese thesaurus and count how many times the synonyms appear in the SA. For example, if “競技” is a synonym of “競争”, then the number of appearances of “競技” is added to that of “競争”. If we go further to handle the meaning, we might as well process words by semantic processing.

Since a kanji is an ideogram, the second method also handles the meaning. When the selection area is in English or French, the corresponding method is to count morphemes such as “inter” and “national” for the word “international”. We are interested in to what extent the method that does not explicitly concern the meaning may be used to handle the meaning of words. In the following, we focus on the former two methods.
3 Fundamental Concepts
3.1 Harmonized Dictionary
A bilingual dictionary forms a graph whose nodes are words and whose branches are correspondences between the words. Branches have directions, which make the graph asymmetric. $\text{Dic}_{x\rightarrow y}$ is a graph with all the branches in $\text{Dic}_{x\rightarrow y}$ in reverse direction.

Since the purpose of bilingual dictionaries is to denote the correspondences of words that have the same meaning [Harst], it is natural that branches are bidirectional. We therefore design symmetrical dictionaries and denote a dictionary from language $x$ to $y$ as $\text{Dic}_{x\rightarrow y}$, calling it a harmonized dictionary. When $\text{Dic}_{x\rightarrow y}$ and $\text{Dic}_{y\rightarrow x}$ are constructed from the same dictionaries, $\text{Dic}_{y\rightarrow x} = \text{Dic}_{x\rightarrow y}$ holds. We remove the overlaps of branches.

3.2 Syntactic Selection Procedure
A multiset here is a set in which each element has a weight that is a natural number. The weight of an element is defined as the number of times it appears when looking up words in dictionaries. In the example shown in Fig. 3, the multiset SA for competition consists of “競争” with weight 2 and “競技” with weight 3. We denote the weight of element $x$ in multiset $X$ as $\delta_a(X, x)$; for instance, $\delta_a(SA, 競争) = 2$. Using the same notation, $\delta_a(X, Y)$ is defined as follows when $X$ and $Y$ are multisets:

$$\delta_a(X, Y) = \sum_{y \in Y} \delta_a(X, y)$$

When a multiset $Z$ consists of “競争” and “競技”, then $\delta_a(SA, Z) = 5$.

The notation $\delta_b(X, x)$ represents the sum of the weights of the elements that contain PWs of $x$ in multiset $X$. For instance, if PWs are defined as kanji, $\delta_b(SA, 競争)$ is 7 by adding 5 (sum of weights of elements in the SA that have kanji “競”) and 2 (sum of weights of elements in the SA that have kanji “争”). Using the same notation, $\delta_b(X, Y)$ is defined as follows when $X$ and $Y$ are multisets:

$$\delta_b(X, Y) = \sum_{y \in Y} \delta_b(X, y)$$

For instance, $\delta_b(SA, Z) = 12$; that is $7 + 5$ (= $\delta_b(SA, 競技)$). We use the notation $\delta$ as a parameter for $\delta_a$ or $\delta_b$.

3.3 Properties of Inverse Consultation
In the following, we use $\text{Dic}_{x\rightarrow e}$ and $\text{Dic}_{e\rightarrow j}$ as inverse dictionaries when starting from Japanese words and we use $\text{Dic}_{j\rightarrow e}$ and $\text{Dic}_{e\rightarrow f}$ as inverse dictionaries when starting from French words. French ECs for a Japanese word $j$ form a multiset $F$ expressed as $F = \text{Dic}_{j\rightarrow e} \cup \text{Dic}_{e\rightarrow f}$.

French equivalences selected by $IC_1$ form a multiset whose element $f$ satisfies $f \in F$ and $\delta(D_{j\rightarrow e}f, D_{j\rightarrow e}j) > 1$.

French equivalencies selected by $IC_2$ form a multiset whose element $f$ satisfies $f \in F$ and $\delta(D_{j\rightarrow e}j, D_{j\rightarrow e}f) > 1$.

In the following, we focus on the $IC_1$ and $IC_2$ described above and examine their properties.

Property 1

$$\delta_a(D_{e\rightarrow j}D_{j\rightarrow e}f, j) = \delta_a(D_{j\rightarrow e}j, D_{j\rightarrow e}f) = \delta_a(D_{e\rightarrow f}D_{j\rightarrow e}j, f)$$

This property indicates that when using selection procedure $\delta_a$, equivalences selected by $IC_1$ and $IC_2$ are exactly the same. Moreover, it is sufficient to choose ECs whose weights are greater than 1 in F. The proof of Property 1 is shown in Appendix.

Property 2

If $\delta$ is $\delta_b$, the resulting Japanese-French dictionary is a harmonized dictionary. This is not always true for $\delta_b$.

This property is clear with symmetrically structured dictionaries. When $\delta_b$ is used, the resulting dictionary depends on how PW is defined and does not always become a harmonized dictionary.

Property 3

If there is a structure such as shown in Fig. 4, $\delta_b$ must be used as $\delta$ to exclude irrelevant ECs.

When $\delta_a$ is used, the SA for two ECs will be exactly the same, which makes it impossible to discard inapplicable ECs. Although this kind of structure seems to be rare, it can exist because of the historical transition of words. When a single English word is intermediated, inapplicable ECs cannot be discarded by using $\delta_a$ for $IC_1$.

4 Experiment
4.1 Dictionary Data
The dictionaries used in the experiment are $\text{Dic}_{j\rightarrow e}$ [Ish99], $\text{Dic}_{e\rightarrow j}$ [Kog99], $\text{Dic}_{e\rightarrow f}$ [For82], and $\text{Dic}_{j\rightarrow e}$ [Lec82]. The whole experimental procedure is shown in Fig. 5. Word-to-word dictionaries are first extracted from each dictionary. All words are nouns; in particular, they are one word nouns in English and French. Since the dictionary syntax was not always consistent, word-to-word dictionaries contain some mistakes (inadequate correspondences).

Harmonized dictionaries are then constructed from the word-to-word dictionaries as follows:

$$D_{e\rightarrow j} = \text{Dic}_{j\rightarrow e}^{-1} \cup \text{Dic}_{e\rightarrow j}$$
\[
\begin{align*}
D_{j \rightarrow e} &= D_{e \rightarrow j}^{-1} \\
D_{e \rightarrow f} &= \text{Dic}_{f \rightarrow e}^{-1} \cup \text{Dic}_{e \rightarrow f} \\
D_{f \rightarrow e} &= D_{e \rightarrow f}^{-1}
\end{align*}
\]

Although there are other ways to symmetrize dictionaries, (for example, by removing all branches that are not bidirectional), we chose the above procedures for the lexicographical reason described below [Har83].

There are two kinds of bilingual dictionaries, one is from a foreign language \(f\) to the mother language \(\text{Dic}_{f \rightarrow m}\), and the other is from a mother language \(m\) to a foreign language \(\text{Dic}_{m \rightarrow f}\). In \(\text{Dic}_{m \rightarrow f}\), when there are no equivalencies in \(m\) for a foreign word, the dictionary gives its definition or explanation of the word in \(m\). Therefore, all the foreign words can be contained in the dictionary. In \(\text{Dic}_{m \rightarrow f}\) on the other hand, if there are no equivalencies in \(f\) for a word of \(m\), the word itself is often dropped from the dictionary. The words contained in the dictionary are therefore a part of \(m\), and \(\text{Dic}_{m \rightarrow f}\) lacks many entries. A harmonized dictionary is a solution to this problem because it contains equivalences of \(\text{Dic}_{f \rightarrow m}\) as entries of \(\text{Dic}_{m \rightarrow f}\).

4.2 Procedure of Inverse Consultation

From Property 1, we use \(\delta_a\) with \(IC_1\) and we use \(\delta_b\) with \(IC_2\). The PW for \(\delta_a\) are defined as follows:

- Japanese 6353 kanjis.
- French morphemes [Mar85].
- 1151 prefix and 710 suffix.

From Property 2, inverse consultation is applied to both Japanese and French entries, and then the results are put together to construct a harmonized dictionary. We denote it as \(D_{j \rightarrow f}\) or \(D_{f \rightarrow j}\).

Each entry within \(D_{j \rightarrow e}\) and \(D_{f \rightarrow e}\) is classified into one of five types according to the procedure used to select its equivalencies from ECs.

- Type A A single EC exists and is selected unconditionally as the equivalence.
- Type B Equivalencies by \(IC_1\) exist; all are selected and the rest of the ECs are discarded.
- Type C There are no equivalencies by \(IC_1\) but there are by \(IC_2\). One third (empirically decided) of ECs by \(IC_2\) are selected according to the value of the following fraction (larger ones are selected):
  \[
  \text{Value of function } \delta_b \\
  \frac{SA \text{ (byte)}}{3}
  \]
- Type D No equivalencies by \(IC_1\) nor by \(IC_2\) appeared but there are several ECs. For this entry, it is impossible to select the relevant equivalencies.
- Type E There are no ECs.

Entries of Type A acquire more appropriate equivalencies than do entries of Type B, and entries of Type B acquire more appropriate equivalencies than do entries of Type C.

5 Evaluation of Experiment

5.1 Result of the Example

Equivalencies obtained for the Japanese word “競争” are

- concours, rivalité, compétition,
- course, concurrence, émulation

and the intermediated English words are

- competition, contest, rival, rivalry, race.

The entry “競争” is of Type B. The number of ECs is 41 (including overlaps), and 13 of them are selected by \(IC_1\). The number of ECs in each category of irrelevant words described in Section 2.1 is listed in Table 1, which indicates that inverse consultation can detect the relevant words even when there are mistakes in the dictionaries. The word which should not be dropped was joute.

Equivalencies for “競争” in [Mar85] are
Table 1  Details of ECs of “競争”.

| Selected equivalences | 13 |
|-----------------------|----|
| Different meaning in English | 9 |
| Meaning extended in English | 9 |
| Mistakes of dictionaries | 9 |
| Words which must not be dropped | 1 |

Table 2  Classification of entries.

|       | $D_{j-f}$ | $D_{f-j}$ |
|-------|-----------|-----------|
| Total | 42190 (100.0%) | 23710 (100.0%) |
| Type A | 1600 (3.8%) | 3852 (16.2%) |
| Type B | 8179 (19.4%) | 7397 (31.3%) |
| Type C | 24047 (57.0%) | 6452 (27.2%) |
| Type D | 1514 (3.6%) | 2988 (12.6%) |
| Type E | 6850 (16.2%) | 3021 (12.7%) |

Table 3  Entries with no EC.

| category | example | English equivalence |
|----------|---------|---------------------|
| Cultural words | 3素堂 (otoshidama : Our tradition to give money to children on New Year) | handsel |
| Technical terms or proper nouns | ガス (gais : gauss) | cedilla |
| Borrowed words | Ciceron | Cicero |
| | アペリティフ (aperitif : apertizer) | apertif |
| | tee-shirt | teeshter |

concours, rivalité, compétition, course, concurrence.

Our result contains emulation (which means rivalry) in addition to the entries in the published dictionary.

5.2 Evaluation of Entries

Table 2 lists the each number of entries belonging to the Types A-D (defined in Section 4.2). Type D consists of the following entries.

- One English word is intermediated and several ECs appear.
- The entry contains no PW.

The Type D percentage difference between $D_{j-f}$ and $D_{f-j}$ shows that the number of entries of Type D depends on the number of registered PWs.

No EC appears when an English word to be intermediated does not exist as the entry of $D_{e-f}$ or $D_{e-j}$. Such entries can be categorized as in Table 3.

Original words are apt to be translated into uncommon English words, so they normally do not appear as entries if the same kind of words do not exist in the objective language. Technical terms and proper nouns depend very much on culture.

Many French placenames, for instance, are Type E. Borrowed words are expressed in inconsistent spellings, and which of them are to be found in the dictionary also depends on the culture (apertif is the equivalent in $D_{ic-j-e}$).

Since a harmonizing dictionary augments the entries, the resulting dictionary contains entries that are not in the published dictionary. As explained in Section 4.1, this phenomenon is conspicuous when we compare $D_{j-f}$ with the published dictionary [Su70]. These entries can be categorized as follows:

1. Colloquial words.
   - Ex. わんちゃん (wan-chan : puppy)
2. Technical terms or proper nouns.
   - Ex. アスベスト (asbestos : asbestos)
3. Compound nouns.
   - Ex. 水化作用 (shuukayou : disintegrate)

Harmonizing dictionaries help to gather correspondences between the mother language and a foreign language and are useful in revising published dictionaries.

5.3 Evaluation of Equivalencies

We evaluated the equivalencies of resulting dictionaries by comparing them with those of published dictionaries [Tama85] [Su70]. For random 100 entries in both dictionaries, the following two percentages (calculated manually) are listed in Table 4:

- R1  Fraction of equivalencies in the published dictionary which were also found by this method.
- R2  Fraction of equivalencies found by this method which were judged appropriate.

Note that entries with greater R2 contain appropriate equivalencies on higher rate. This is not true for R1, since R1 indicates the discrepancy of equivalencies between the resulting dictionary and published dictionaries.

Entries of R2=100% can be classified as follows:

1. Entries of Type A or B. (See Section 4.2.)
2. Entries that have less than three ECs. Entries with less than 3 ECs amount to 29.0% for $D_{j-f}$ and 34.2% for $D_{f-j}$. This indicates that words with specific meanings are apt to acquire appropriate equivalencies.
Table 5 French entries whose Ri are 0 %

| entry | Dř-j | Published dictionary |
|-------|------|----------------------|
| distique | 行 達 (gyou-|ri : coinage for the term for European poems) |
| pull | セーター (secta : pull-over written with Japanese letters. This word is common) |
| boulier | 算盤 (sanban : abacus in general) |

Table 6 Japanese entries whose Ri are 0 %

| entry | Dř-j | Published dictionary |
|-------|------|----------------------|
| つくみ (tsunomi : thrush) | grive | merle |
| 酒蔵 (sake : vit- | astuce | calemboir |
| ティック, joke, jest, pun) | badinage | | |
| | drôle | face |
| | face | plaisanterie |

The entries tend to have either Ri=80%~100% or Ri=0%~20%. Of the latter, some examples are listed in Table 5 and 6. For distique in Table 5, which is a term in French literature, Dř-j translates it into the term for the corresponding kind of Japanese literature. Although it resembles to the direct translation of distique, it is only an analogy. On the other hand, since the direct translation of distique in the published dictionary adopts the concept of French poems, it is uncommon and cannot be understood by most Japanese readers. The same is true for boulier except that the common Japanese is indicated in a published dictionary. Pull is borrowed from the English word pull-over whose direct translation is contained in a published dictionary, and it is not a common word in Japanese.

In the first example in Table 6, grive is the generic name equivalent to thrush, whereas merle is a kind of thrush. The second example shows that more equivalences are found in Dř-j.

To sum up, the resulting dictionary can be utilized in conjunction with the published dictionaries as follows:
- To revise the equivalences.
- To supplement the equivalences.

Sections of the resulting dictionaries is listed in Table 7. List 1 is Dř-j and List 2 is Dř-j.

For each list, entries are in the first row and their equivalencies are in the third row. Symbols in the second row indicate how appropriate each equivalence is. (Refer to the notes beside.)

6 Related Work

The use of a third language English as an intermediary in the construction of a bilingual dictionary was tested manually on a large scale on editing the Spanish-Japanese dictionary [Kuw90]. It is now a representative middle-sized dictionary having a large quantity of information.

Tokunaga and Tanaka [Tok90] tried to extract a conceptual dictionary from Japanese-English and English-Japanese dictionaries. Although they used a concept similar to ours that is the graph structure of a dictionary related to meaning of words, their frameworks and final product differ from ours.

7 Conclusion and Future Works

The proposed method for using an intermediate language to construct a bilingual dictionary utilizes the structure of dictionaries and morphemes and can choose appropriate equivalencies for most entries. Comparing the resulting dictionary with published dictionaries showed that data obtained are useful for revising and supplementing the vocabulary of existing dictionaries.

To increase the accuracy with which equivalences can be selected, mistakes in word-to-word dictionaries must be corrected even if our method may detect appropriate equivalencies. One way to do this would be to use thesaurus to check whether the extracted correspondences are relevant.

Nouns were taken into consideration in this research, and the next step will be to apply the proposed method to other parts of speech. We also need to establish a way to handle compound words in European languages.

Acknowledgements

We express our gratitude to Prof. Hideya Iwasaki of University of Tokyo for his guidance on theoretical aspects and essential remarks on experiments. We are also thankful to Prof. Nigel Ward of University of Tokyo for his lexicographical advice and for his invaluable aid in shaping our idea.

References

[Bog89] Boguraev, B. et al. (1989). Computational Lexicography for Natural Language Processing. Longman.
[Fill85] Fillmore, C. (1985). Frames and the semantics of understanding. Quaderni di Semantica, Vol. 6, No. 2.
[Fill92] Fillmore, C. and Atkins, B. (1992). Toward a Frame-Based Lexicon. The semantics of RISK and its Neighbors. Academic Press.
Table 7  A part of resulting Japanese—French dictionary.

| List1                  | List2                  |
|------------------------|------------------------|
| おり廢 | 0* | Zambie  | 0* |
| 2 | chiame | 2 | criniere |
| 2 | tifs | 1+ | gaminerie |
| 0* | makes | 0 | chien-chien |
| 0 | chichien | 0 | chichien |
| 0 | tonton | 0 | chichien |

| !アーチ | 0 | arc | 0* | abandonement |
| !アート | 0 | arche | 0* | abandon |
| !アービュクス | 2 | japonaisant | 1 | compte |
| !アームロック | 0 | arbonniser | 0 | centime |
| アーモンド | 1+ | amande | 1 | dragee |
| 鴨進元 | 0+ | promoteur | 0* | espace |
| 鴨語 | 1 | discours | 1 | esthetique |
| 鴨詩 | 1 | pantage | 1 | fantome |
| 巻尺 | 1 | ruban-cache | 1 | centimetre |
| !巻き添え | 2 | existence | 2 | magnetophone |
| !巻き物 | 0 | rouleau | 0 | voile |

| 0: Judged appropriate (counted for R2). |
| 1: Significance slightly different. |
| 2: Significance completely different. |

* Found common in two dictionaries (counted for R1). Mark "*" attached to entries means that the entries were not found in the published dictionaries.

---

[For82] Forbes, P. et al. (1982). Shorter English-French Dictionary. Harrap Limited.

[Harc83] Hartmann, R. (1983). Lexicography: principles and practice. Academic Press.

[Ich90] Ichikawa, S. (1990). New Japanese-English Dictionary. Kenkyusha.

[Ide93] Ide, N. and Veronis, J. (1993). Extracting knowledge basis from machine readable dictionaries. KB&KS, JIPDEC.

[Ko90] Koine, Y. (1990). New English-Japanese Dictionary. Kenkyusha.

[Kuw90] Kuwana, K. (1990). Spanish-Japanese Dictionary. Shougakukan.

[Led82] Ledesert, M. et al. (1982). Shorter French-English Dictionary. Harrap Limited.

[Mau85] Maubourguet, P. et al. (1985). Grand Dictionnaire Encyclopedique LAROUSSE. Larousse.

[Mii90] Miller, G. et al. (1990). Five Papers on WordNet. CSL Report 13. Cognitive Science Laboratory, Princeton University.

[Oht78] Ohtsuki, T. et al. (1978). Crown French-Japanese Dictionary. Sanseido.

[Suz70] Suzuki, S. (1970). Standard Japanese-French Dictionary. Taishukan.

[Tam85] Tamura, T. (1985). Dictionnaire Francais-Japonais ROYAL. Hakusensha.

[Tok90] Tokunaga, T. and Tanaka, H. (1990). The Automatic Extraction of Conceptual Items from Bilingual Dictionaries. PRICAL.

---

### Appendix

The following two lemmas are needed to prove Property 1.

**Lemma 1-1** \( x \in D_{y-x} \iff y \in D_{y-x} \)

This is clear from the symmetric structure of the harmonized dictionary.

**Lemma 1-2** If \( X \) is a set (every element has a weight of 1) then \( \delta_a(D_{x-y} X, y) = \delta_a(X, D_{y-x} y) \)

**Proof**: \( |X| \) denotes the number of elements in \( X \)

\[
\delta_a(D_{x-y} X, y) = |\{x | x \in X \land y \in D_{x-y} x\}| = |\{x | x \in X \land x \in D_{y-x} y\}| = \delta_a(X, D_{y-x} y) \quad \Box
\]

As branches do not overlap, \( D_{x-y} f \) is a set. From Lemma 1-2, the proof for Property 1 is given as follows:

\[
\delta_a(D_{x-y} f, y) = \delta_a(D_{y-x} f, D_{x-y} f) = \delta_a(D_{y-x} f, D_{y-x} f) = \delta_a(D_{y-x} f) \quad \Box
\]

The proof for the second equation is the same.