Enhancing an English/Korean Dictionary

Kyonghee Paik[1] and Francis Bond[2]

[1]Spoken Language Translation Laboratories
ATR
kyonghee.paik@atr.co.jp

[2] NTT Communication Science Laboratories
Nippon Telegraph and Telephone Corporation
bond@cslab.kecl.ntt.co.jp

Abstract

In this paper, we introduce a machine-tractable Korean/English lexicon. We use `engdic`, an open source dictionary. `engdic` is an English-Korean dictionary for human use. The formatting is sometimes inconsistent, and there is missing or duplicated information, therefore it is not ready for machine use. We rearrange the disorganized format as well as improve the content. This makes it easier to use the dictionary bidirectionally. Our main purpose is to develop and document clear syntactic and semantic features useful for NLP applications such as machine translation. The original lexicon contains about 98,000 English lemmas and about 210,000 English-Korean pairs. Each entry consist of three parts: English lemma form, part of speech codes, and Korean translation/explanation. We transformed this to a more structured format consisting of eight fields.

1 Introduction

In this paper, we describe the process of formatting a machine-readable Korean-English lexicon to produce a machine-tractable Korean/English lexicon more suitable for use by both people and machines. The original lexicon is `engdic` (Park, 2000), an open-source machine-readable English-Korean lexicon. This resource is licensed under the GPL (Gnu Public License), so it can be modified and redistributed. Our reformatted dictionary will be available as part of the Papillon project (`http://www.papillon-dictionary.org/`).

In the following section (§ 2), we describe the original dictionary. In Section 3, we describe our new format, and the changes we made to produce it. Then we introduce some applications of the reformatted dictionary, such as building a Japanese-English dictionary and using it with standard dictionary software. Finally we discuss what further work needs to be done to integrate the dictionary fully into the Papillon multilingual database.

2 A Description of `engdic` (v0.2)

In this section, we describe the original structure of the `engdic`. It was compiled by KwangSuk Lee in 1996 and is currently distributed by several Linux distributions: e.g., Debian, SUSE and Redhat. We took our data from the Debian KR Project's `engdic`-0.2-6 (Park, 2000), and describe it here. The package contains the dictionary (encoded

in `euc-kr`), split into 26 files (`a-z.dic`), and a shell program (`edic`) for looking up words. There is no documentation apart from a brief README and changelog.

The dictionary has 98,000 entries. The default format is an English Lemma followed by a colon; one or more English Parts of Speech followed by commas; and one or more Korean Translation/Explanations, separated by commas. An example is given in (1):[1]

(1) `dictionary` : n, 사전, 사서
sajeon, *saseo*
dictionary, lexicon

About 5,000 lemmas are linked to an English synonym or spelling variation (i.e. *color/coulor*). 51% of the entries have only one translation. The translation is sometimes replaced with an explanation, in Korean or English. Expanding out the translations gives 210,000 pairs.

In the next sections we describe the three parts in more detail.

2.1 English Lemma

The English lemma can consist of one or more words. Almost 5,000 of the 98,000 entries (5%) contain no Korean. They are either expansions of abbreviations, or pointers to spelling variants (marked with =). We give examples below:

(2) `A.B.S.` : x, American Bible Society

(3) `abo-o, ab-o` : n, =ABORIGINAL

(4) `colour` : x, =COLOR

(5) `exp` : x, expense(s), expired; exportation, export(ed), exporter

Approximately 20,000 (20%) of the English lemmas are multiword expressions, such as *ballot paper* or *demilitarized zone*.

There are also around 2,000 English multiword expressions in the Korean Translation/Explanation section, such as (6).

(6) `abominable` : a, 싫은, the ∼ snowman (히말라야의) 설인
silun, *(himallaya-uy)* *seolin*
abominable, (Himalaya-GM) snowman

There are some entries (around 500) where the entries are not whole words, either prefixes or suffixes: *ambi-,-ly*.

2.2 English Part of Speech

Altogether there are 16 English parts of speeches used in `engdic` v0.2: `adjective, adverb, conjunction, interjection, noun, plural only noun, pronoun, predicative adjective, prefix, preposition, suffix, verb, intransitive verb, transitive verb`, and `other`, which includes abbreviations, multiword expressions, etc. The part of speech `v` includes explanations of past and past participle forms. Many entries have multiple parts of speech. In the original `engdic`, many entries had no, or misspelled parts of speech. Because both the parts of speech and the Korean translations are separated from each other by commas, it was impossible to tell them apart automatically.

[1] We have added a transliteration and gloss of the Korean, only the first line is actually in the `engdic` file. We transliterate the adnominal case marker as ADN.

2.3 Korean Translation/Explanation

Around 50% of the entries have only one translation, the rest have more than one. They are not divided into senses. The word with the most translations is *cut* with 111 translations.

There are also around 61,000 pieces of information in brackets. These are mainly free language elements, giving additional explanations, semantic classes, domains, dialects and so forth. The explanations were inconsistently formatted. For example, the explanation saying that an entry was a name (male or female) was generally given not in brackets, but rather directly, as though it were a translation.

As well as explanatory material, optional grammatical elements (the equivalent of *(s)* in *expense(s)*) is given in brackets. This is acceptable for human readers, but makes it impossible to do a simple search of the lexicon for the variant form.

Let us examine some different variations in `engdic`'s format with more examples.

Explanatory Material First is an example of explanatory material. In (7), an explanation is given in Korean. This is the most common kind of free language element.

(7) `abacus` : n, 수판, 주판, (둥근 기둥의) 대접받침
supan, jupan, (dunggun gidung-uy) daejeop-batchim
abacus, abacus, (column's) capital

Optional Grammatical Elements — Substitution In (8), there are two possible Korean translations of *excavator*: 굴착자 *gulchakja* "a person who excavates" and 굴착기 *gulchakgi* "a machine that excavates". These have been collapsed into a single string: 굴착자(기) *gulchakja(gi)* "excavator". To extract the Korean translations, we have to substitute the parenthesized material for the final character.

(8) `excavator` : n, 굴착자(기), 발굴자
gulchakja(gi), balgulja
excavator-person(machine), digger

Another example of substitution is (9), where *abbacy* can refer to "the office or term of an abbot".

(9) `abbacy` : n, 대수도원장의 직(임기)
dae-sudowon-jang-uy jik(imgi)
big-monastery-head-ADN position(term)

Optional Grammatical Elements — Insertion In (10), the Korean translations of the verb *feast* are the intransitive 즐기다 *jeulgida* "*X* feasts" and the causative 즐기게 하다 *jeulgige hada* "*Y* makes *X* feast". Here the parenthesized expression has to be inserted.

(10) `feast` : vi,vt,n 축제, 축연, 대접, 즐기 (게 하) 다
chukje, chukyeon, daejeob, jeulgi (ge ha) da
festival, banquet, treat, enjoy (make) do

Another example of insertion is (11). Here the difference is between relating to an abby (no insertion), or an abbot (insertion).

(11) `abbatial` : a, 대수도원(장)의
dae-sudowon-(jang)-uy
big-monastery-(head)-ADN

Other Free Language Elements There are many other kinds of free language elements. One common one (2,000 examples) is to include an English multiword expression. For example, in (12), the multiword expression "abnormal psychology" is given. Since multiword expressions are allowed as lemmas, it would be more consistent to add "abnormal psychology" as a separate lemma.

(12) `abnormal` : a, 비정상적인, 이상의, 변태의, 변칙의, (∼
bijeongsangjeokin, isanguy, beontaeuy, beonchikuy, (∼
unusual, strange, perverted, anomaly, (
psychology 이상 심리학)
psychology isang simrihak)
abnormal psychology)

2.4 Summary (v0.2)

The original `engdic` is a good dictionary for Korean speakers who wish to look up the meaning of English words. It has a lot of useful information and a wide coverage. However, it is not formatted consistently. For example, the part of speech field is missing or incorrect for many entries, brackets do not match and so on. There are some duplicate entries, although they are very rare. In addition, there were some characters that could not be encoded in `euc-kr`, which made it hard to edit the files. Further, as parts of speech are not separated from translations/explanations, and many are mistyped, it is impossible to parse 100% accurately. These inconsistencies make the lexicon less useful for reverse look-ups, and not suitable for natural language processing.

3 Redesign of `engdic` Lexicon (v0.3,0.4)

We have reformatted the original `engdic` to become a more machine-tractable lexicon. Our aim was to clean the dictionary to a state where we could parse it, and then parse it to make the information more accessible. We wanted to keep all the information in the original dictionary, so that it would remain useful for Korean speakers looking up English. We also want to make it more useful for Korean or English speakers looking up Korean, and for NLP applications. The tractable version is also suitable for conversion to XML, and thus is on the way to becoming usable for the Papillon project (Mangeot-Lerebours, 2002). Our aim was always to exploit the existing information as fully as possible.

We reformatted the dictionary in three steps:

1. A brute force clean up of the original format
2. Restructuring the dictionary, partially parsing the free elements
3. Adding Hanja (Chinese characters) using other resources

As the version we started with was v0.2, we call the cleaned up version v0.3, and the reformatted version v0.4.

We describe the brute force clean up in Section 3.1, the parsing in Section 3.2 and adding Hanja (Chinese characters) in Section 3.3.

3.1 Brute Force Cleanup (v0.3)

In the brute force cleanup, we made the format easier to parse, and hand corrected many errors. The new format is a tab separated list of: English Lemma; comma separated list of parts of speech; semicolon separated list of translations/explanations. We made sure that only commas, and no semicolons were used inside parenthesized free language elements.

Among the errors we corrected were: un-encodable characters, misspellings, missing parentheses, inconsistent explanations, duplicated translations, unmatched translations and so on. Most of the correction was done using regular expressions inside a text editor (EMACS).

We fixed any errors as we found them, but concentrated on errors which affected parsing, as we have limited time. As the corrected lexicon will be made freely available, other users can also make corrections. The resulting dictionary we call `engdic` v0.3.

3.2 Reformatting the Lexicon (v0.4)

Our goal in reformatting the lexicon is to make it (a) more tractable; and (b) more accessible, so that it could be searched for both Korean and English.

To make it more tractable, we extracted as much information as we could from the free language elements, and mapped this information to standard representations. To make it more reversible, we split each entry into pairs of Korean and English, and associated information with the pairs. We gave each pair an ID, so that we could link the information to them. We also kept a separate table of links, to show which links were originally the same entry.

In the new format each entry consists of a tab separated list:

Field 1 A Unique ID

Field 2 An English lemma (obligatory)

Field 3 One or more English parts of speech (obligatory)

Field 4 Unparsed free elements in pre-position (optional)

Field 5 Korean translation/explanation OR English Equivalent (obligatory)

Field 6 Korean word in Chinese Characters (optional)

Field 7 Korean unparsed free elements in post-position (optional)

Field 8 Meta information as a list of attribute value pairs (optional)

Simplified examples are given in Table 1 (with no Chinese Characters). A fuller description of the fields is given in the following sections.

ID	English	EPOS	(Pre)	Korean	(Post)	Meta
1	dead freight	x	-	*Kongha unim*	-	DOM=commerce
2	deaf-aid	n	-	*bocheonggi*	-	DIA=UK
3	debase	vt	*pumjil 'quality'*	*jeoha-sikida*		
4	feast	vi,vt,n	-	*jeulgida*	-	SYN=vi
5	feast	vi,vt,n	-	*jeulgige hada*	-	SYN=vt
6	hedonistic	a	-	*kwaerakjuuy(ja)uy*	-	SEM=person

Table 1: Some examples of `engdic` v0.4.

The format was designed to be able to easily access the translation equivalents (English ⇔ Korean) without losing any information. We do not suggest that it is a generally useful format. Instead, it is a useful transitional format.

We describe the process of converting the lexicon to this format in the following sections.

3.2.1 English Part of Speech

There are several problems with the English part of speech codes.[2] The first is that too many entries (around 40,000) are tagged as `other`, including most MWEs. The second is that the POS codes are not well structured. Noun countability is not marked and so on. However, because the choice of detailed part of speech codes depends on your theory of grammar, we decided to postpone any refinement until we needed more detailed information. Deciding the lexical types of multiword expressions is particularly tricky (Sag et al., 2002), we decided we would mark MWEs with the same high-level set of POS tags that we used for single words. The fact that an entry is a MWE should be recoverable by checking if the lemma includes spaces.

3.2.2 Meta Information

Meta-information includes information about register (META = colloquial, written, vulgar, archaic); dialects (DIA = UK, US, AU, ...); origin (OR = ja, de); Korean syntactic information (SYN = vi, vt, ...); Domain (DOM = maths, botany, law ...), semantic classes (SEM = person, animal, ...) and so on. A fuller listing is given in the documentation.

The information is extracted by matching against the parenthesized free language elements. Some examples are given in Table 2.

Type	String	Position	Value	Example
DIA	(미)	Pre	US	deputy sheriff 군 보안관 대리
DIA	(오스)	Pre	AU	drongo 얼간이
DOM	(전산)	Pre	computing	debugging 디버
DOM	(법)	Pre	legal	day in court 법정 출두일
META	(고)	Pre	arch[aism]	dexter 운이 좋은
SEM	(증)	Post	pathology	impetigo 농가진

Table 2: Some patterns used to extract Meta-Information

[2]We would like to thank an anonymous reviewer for their comments about this.

As much as possible we tried to use standard representations for the data, such as ISO codes for languages and countries. We hope that the Papillon project will settle on a standard for DOMAIN, SEMANTIC and META tags which we can adopt. For the time being, we are trying to remain compatible with JMDict (Breen, 1995).[3]

3.2.3 Korean Optional Grammatical Elements

We expand out Korean optional grammatical elements so that each entry can have its own semantic code and part-of-speech.

For example, (8) becomes three entries (tabs are shown as colons).

(13) `excavator` : n : 굴착자 : SEM=person, SYN=n
gulchakja
excavator

(14) `excavator` : n : 굴착기 : SEM=machine, SYN=n
gulchakgi
excavator

(15) `excavator` : n : 발굴자
balgulja
digger

Because the grammatical element also gives information about the semantics and Korean syntax, they are tagged at the same time. This information was coded by hand for each of the optional elements we recognize, currently 14 different types.

Similarly, (9) is split into two entries:

(16) `abbacy` : n : 대수도원장의 직 : SEM=position
dae-sudowon-jang-uy *jik*
big-monastery-head-ADN position

(17) `abbacy` : n : 대수도원장의 임기 : SEM=term
dae-sudowon-jang-uy *imgi*
big-monastery-head-ADN term

A similar process can be applied to insertions. The entry for *feast* (10) is expanded into (18–19):

(18) `feast` : vi,vt,n : 즐기 다 : SYN=vi
jeulgi da
enjoy

(19) `feast` : vi,vt,n : 즐기게 하다 : SYN=vt
jeulgige hada
enjoy make

3.2.4 Remaining Free Language Elements (FLEs)

Not all the free language elements (text in brackets) could be parsed. Many of them are grammatical usage notes, giving, for example, typical arguments of verbs. There are

[3] `http://www.csse.monash.edu.au/~jwb/jmdict_dtd_h.html`

also several encyclopedic comments and explanations. The format for these is extremely variable.

Boitet (2002) presents various strategies for translating free language elements (FLEs). For the time being, we leave any such elements untranslated, but separate them from the main Korean translation. The main Korean translation can then be more easily used to look up Korean words. We store the remaining free language elements in two fields, one for those that come before the main Korean entry (pre) and one for those that come after it (post). In this way we can reproduce the original layout of the dictionary.

3.2.5 Harmonizing POS codes

In v0.4, each English headword has only one Korean translation. In this case, we want to specialize the part of speech listed with the English word to the appropriate one for that particular Korean translation. We call this process harmonization. The process proceeds as follows. First, we estimate the POS according to the morphology of the Korean word. Then, we attempt to match it with the English POS. If there was a match, we use it, discarding any other POS candidates. Otherwise if the POS was unknown, we use the estimated POS. In all other cases we keep the original POS.

The estimated POS was the same as the original POS for slightly over half the entries. We specialized the POS for around 61,000 entries (30%). We were unable to specialize for the remaining 20%.

For example, *feast* in (10) has four translations *chukje, chukyeon, daejeob, jeulgi (ge ha) da*. One of the Korean translations, 즐기 (게 하) 다 *jeulgi (ge ha) da*, ends with *-da* and is thus a verb, the rest are nouns. One, 축연 *chukyeon*, ends in *-n* and is thus ambiguous between a noun and an adjective. The rest are nouns by default. Further, 즐기 (게 하) 다 *jeulgi (ge ha) da* matches one of the optional grammatical rules, so it is resolved to two entries, one transitive and one intransitive (shown in 18–19).

The English part of speech codes are `n,vi,vt`. The harmonization process matches the nouns and the verbs, but leaves the ambiguous case. The resulting entry is shown in Figure 2

3.3 Adding Hanja

Finally, we have a list of Korean (hangul) matched with Korean (Hanja: Chinese characters) and their English translations, made from `engdic` v0.2 while building a Korean-to-Japanese dictionary (Paik et al., 2001). We were able to add Chinese characters to roughly 6,000 translation pairs (3%).[4]

(20) `dictionary` : n : 사전 : HAN=辭典
sajeon *sajeon*
dictionary

(21) `dictionary` : n : 사서 : HAN=辭書
sajeo *saseo*
lexicon

[4] `engdic` v0.2 had Hanja for seven entries.

Chinese characters are not used so much in Korean these days, but are still useful for learners of Korean who know them; for finding the etymology of words; and for reading older Korean text.

3.4 Documentation

We created documentation (`engdic-doc.html`) to go with the lexicon, that describes the various meta tags.

3.5 Summary (v 0.4)

The final dictionary (`engdic` v0.4) consists of 217,620 entries. Of these, 20,841 have no Korean, leaving 196,779 Korean-English pairs. There are 92,982 unique English lemmas, of which 19,441 are MWEs. There are 130,228 unique Korean Translation/Explanations, of which 64,087 are MWEs. Not all the Korean entries are translation equivalents, many explanations remain in the dictionary.

5,823 of the entries have Hanja. 20,763 of the entries have pre-position FLEs. 20,673 of the entries have post-position FLEs. 27,587 of the entries have some Meta Information, including 2,624 with information about the semantics, and 1,910 with information about the domain 915 with information about dialect use.

Note, however, that this is still a work in process, and the dictionary released for the Papillon seminar is a somewhat arbitrary snapshot.

4 Applications of the Redesigned Lexicon

The particular application we had in mind when we reformatted the dictionary was creating a Korean/Japanese machine translation lexicon using English as a pivot (Paik et al., 2001). We did this using v0.2 and found many problems due to the incorrect formatting.

As we have shown in the previous sections, we have rearranged the translation/explanation parts so they are more useful from the viewpoint of users both for native speakers of Korean and for second language learners of Korean. Since we use semantic class which will be linked to a dictionary with semantic classes, users can obtain more precise and detailed information. At the same time, it is more useful for natural language processing.

We can also output the dictionary in a format suitable for use by the standard dictionary protocol `dictd` (Faith and Martin, 1997). We describe the process of building a Korean-to-Japanese Lexicon in the Section 4.1, and the conversion to `dictd` format in Section 4.2.

4.1 Generation of Korean-to-Japanese Lexicon using `engdic`

In this section, we introduce briefly how to build a Korean-Japanese dictionary using English and Chinese characters as dual pivots.

We need four things: a Korean-English dictionary with Chinese characters, a Japanese-English dictionary with Chinese characters, a way of matching the English and a way to match the Chinese characters.

First we reversed the `engdic`, using only the Korean translation field, to get Korean-English pairs. Then, we added candidate Hanja, from the Korean input tool **freewnnkserver** (`<www.freewnn.org>`). Most words for which Hanja were found, had multiple

candidates. Then we linked the English words to English words in EDICT (Breen, 1995), and looked up their Japanese translations. This matching was done using both simple string matching and using WordNet (Fellbaum, 1998) to match words.

This gave us Korean-English-Japanese triples. To be sure that the Japanese and Korean was a good match, we then matched the Korean Hanja with Japanese Kanji (which are all derived from Chinese Characters: Hanzi), as shown in Figure 1. The matching was done using the Han unification in Unicode, which maps similar characters in Chinese, Japanese and Korean onto identical code points, and a table of equivalent characters (such as 気⇔ 氣).

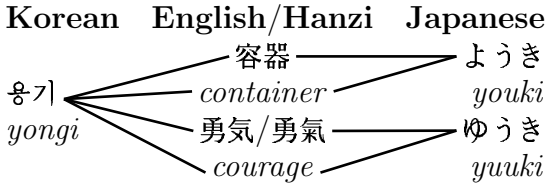

Figure 1: Matching through multiple criteria

This simultaneously disambiguates the link to the Japanese, and gives the correct Hanja for the Korean-English pairs. A fuller description of this process is given in Paik et al. (2001).

Now we have added more information to `engdic` v0.4, we can redo this, also exploiting the META tags.

4.2 Use with the dictd format.

`engdic` comes with a shell script to look up words on the command line. It is basically a matcher using grep with a little bit of prettifying of the output. However, having to maintain a custom lookup tool is inconvenient, and does not allow us to benefit from the enormous amount of existing work people have done building dictionary interfaces. Therefore we suggest two modes of use: with normal text manipulation tools or converted to a standard dictionary format.

Our base format for v0.4 is tab-separated, so it can easily be manipulated by standard Unix tools such as `cut`.

For a more user-friendly interface, we converted `engdic` v0.4 to a format that can be used by `dictd`. This makes it usable with a TCP transaction based query/response protocol that allows a client to access dictionary definitions. The resulting dictionary can then be accessed with a variety of existing tools (see `<http://www.dict.org/>` and `<http://www.freedict.de/index.html>` for more information).

We make two dictionaries: `eng-kor`, with English head words and `kor-eng` with Korean head words. The Korean-English dictionary reverses the English-Korean dictionary and discards any remaining Korean free elements. Both dictionaries are formatted in `utf-8`. We have tested them with the command-line interface `dict` and the gnome dictionary tool `gnome-dictionary`.

An example of the Entry for *feast* is given in Figure 2.

Note how the harmonization has specialized the part of speech for almost all the entries.

feast	
n:	축제 <祝祭>
vi,vt,n:	축연
n:	대접
n:	즐거움
vi,vt:	잔치를 베풀다
vi,vt:	대접을 받다
vt:	즐기게 하다 (SYN=vt)
vi:	즐기다 (SYN=vi)

Figure 2: Dictionary formatted for dictd

5 Further Work

There are several things that could be done to improve the dictionary:

- Correct the remaining spelling errors. Around 6,000 (3%) of pairs contain a word that is not recognized by `aspell` (Atkinson), although not all are errors.
- Improve our parsing of the formatting. Currently we are unable to parse parts of around 15,000 (7.5%) pairs. There are three main types, and a long tail of minor types and typos. The main remaining free language elements we would like to parse are:
  - Multiword English expressions contained in the translation.
  - English derivational suffixes contained in the translation.
  - Low frequency Korean optional grammatical elements.
- Determine the part of speech for entries where it is unknown (6,910 entries).
- Look at the `JIS X 4081` dictionary format: `http://member.nifty.ne.jp/~satomii/freepwing/`.

We would also like to use it to rebuild the Japanese-Korean lexicon discussed in Section 4.1, and then feed back the information on Hanja to `engdic` once more.

Finally, we would like to make this data available for use in the Papillon project. The main remaining obstacle is the fact that the entries are not grouped into senses. One approach would be to treat each English-Korean pair as a seperate sense, or maybe each English-POS-Korean triple as a seperate sense. This would make the dictionary less useful for mono-lingual word sense disambiguation, but still usable as a bilingual dictionary. Further processing could group entries further. For example, by grouping according to wordnet synonyms, or using conceptual vectors (Lafourcade, 2002; Schwab and Lafourcade, 2002)).

6 Conclusion

We introduced the medium sized machine readable English-Korean dictionary `engdic`, and described how we made it more tractable.

Version 0.4 has 92,982 different English head-words and 130,231 different Korean translation/explanations. They are joined in around 196,784 Korean-English pairs. Hanja have been added to around 5,823 entries and meta-information to around 27,587.

Acknowledgments

This work was built on the tremendous effort of KwanSuk Lee. We thank him and all the people who helped to package `engdic`. We also thank the members of the Papillon Project, especially the reviewers, for their illuminating discussions.

References

Kevin Atkinson. Gnu aspell. `<http://aspell.sourceforge.net/>`.

Christian Boitet. The translation of examples, citations, definitions and glosses in the Papillon project. In *Proceedings of Papillon 2002 Workshop (CDROM)*, NII, Tokyo, Japan, 2002.

Jim Breen. Building an electronic Japanese-English dictionary. Japanese Studies Association of Australia Conference (`http://www.csse.monash.edu.au/~jwb/jsaa_paper/hpaper.html`), 1995.

R. Faith and B. Martin. Request for comments: 2229 — a dictionary server protocol. `http://www.dict.org/rfc2229.txt`, 1997.

Christine Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

Mathieu Lafourcade. Automatically populating acception lexical database through bilingual dictionaries and conceptual vectors. In *Proceedings of Papillon 2002 Workshop (CDROM)*, NII, Tokyo, Japan, 2002.

Mathieu Mangeot-Lerebours. How to import an existing XML dictionary into the papillon plateform. In *Proceedings of Papillon 2002 Workshop (CDROM)*, NII, Tokyo, Japan, 2002.

Kyonghee Paik, Francis Bond, and Satoshi Shirai. Using multiple pivots to align Korean and Japanese lexical resources. In *Workshop on Language Resources in Asia*, NLPRS-2001, pages 63–70, Tokyo, 2001.

Chu-Yeon Park. Debian package `engdic 0.2-6: Korean<->English Dictionary`. `<http://www.debian.or.kr/Packages/unstable-kr/text/engdic.html>`, 2000.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. Multiword expressions: A pain in the neck for NLP. In Alexander Gelbuk, editor, *Computational Linguistics and Intelligent Text Processing: Third International Conference: CICLing-2002*, pages 1–15. Springer-Verlag, Hiedelberg/Berlin, 2002.

Didier Schwab and Mathieu Lafourcade. Hardening of acception links throught vectorized lexical functions. In *Proceedings of Papillon 2002 Workshop (CDROM)*, NII, Tokyo, Japan, 2002.