

Analyzing Interlanguage Links of Wikipedias

Yoshiaki Arai¹, Tomohiro Fukuhara², Hidetaka Masuda¹, Hiroshi Nakagawa³

¹ Tokyo Denki University, 2-2 Kanda-Nishiki-cho, Chiyoda-ku, Tokyo, JAPAN
{arai,masuda}@cdl.im.dendai.ac.jp

² The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba, JAPAN
fukuhara@race.u-tokyo.ac.jp

³ The University of Tokyo, 7-3-1, Hongo, Bunkyo, Tokyo, JAPAN
nakagawa@dl.itc.u-tokyo.ac.jp

Abstract. *Interlanguage-links (ILLs)* of Wikipedias is one of emerging multilingual resources. By using ILLs, we can create a multilingual dictionary that can be used for various applications such as cross-lingual information retrieval and the machine translation. Meanwhile, the quantity and quality of ILLs have not been investigated in detail. In this paper, we report analysis results of ILLs by using Chinese, Japanese, Korean, and English (CJKE) editions of Wikipedias. From this analysis, we found (1) quantity of ILLs in each edition, (2) connection state of ILLs in each edition, and (3) patterns of ILLs. We also analyzed the quality of ILLs by using existing dictionaries. We propose a cross-lingual keyword navigation system called *ILL visualizer* as an application of ILLs. An overview of analysis results and the system are described.

1 Introduction

Today, Wikipedia⁴ becomes one of the largest encyclopedias. Wikipedia is an online public encyclopedia maintained by many Internet users in the world. There are various language editions of Wikipedias: English, German, French, Polish, Japanese, Dutch, Italian, Portuguese, Spanish, Swedish, and so on. There are 254 language editions for Wikipedia⁵ at this moment⁶.

One of key features of Wikipedia is an *interlanguage-link (ILL) system*, that is a function of MediaWiki⁷. With an ILL system, Wikipedia editors can create a hyperlink between two entries appeared in different editions on the same entry. Because contents for a topic are quite different by editions, we can find another facts and information that are not described in an edition but described in another edition. Thus, we can find additional information by following ILLs.

ILLs of Wikipedia are important resources for multilingual information access (MLIA[1]). We are creating a cross-lingual concern analysis system using multilingual blog articles[2]. For analyzing concerns of people across languages,

⁴ <http://www.wikipedia.org/>

⁵ http://en.wikipedia.org/wiki/Wikipedia:Multilingual_statistics

⁶ On March 16, 2008

⁷ <http://www.mediawiki.org/>

multilingual textual resources are needed. By using ILLs of Wikipedia, we aim to create a multilingual dictionary that can be used for cross-lingual information retrieval (CLIR), MLIA, machine translation, and so on. Meanwhile, because the quality of Wikipedia contents is not controlled, investigation of the quality of ILLs is needed.

In this paper, we describe analysis results of ILLs of Chinese, Japanese, Korean, and English (CJKE) editions of Wikipedias. We propose a cross-lingual keyword navigation system called ILL Visualizer that utilizing ILLs.

This paper consists of following sections. Section 2 describes the previous work. Section 3 describes analysis results of ILLs. Section 4 describes evaluation results of ILLs by comparing with existing dictionaries. Section 5 describes an overview of the cross-lingual keyword navigation system. In section 6, we summarize arguments and describe future work.

2 Previous work

Ortega reported differences of contributions of Wikipedias across languages[3]. Their study revealed that a few authors contribute many articles. This study surveys several editions of Wikipedias, but does not treat ILLs directly. We analyze ILLs among several language editions of Wikipedias.

Geser analyzed the evolution pattern of ILLs across languages[4]. He used a tool called *Wikipedia Statistics*⁸, which is developed by Erik Zachte⁹, that analyzes all of Wikipedia data dumps, and shows statistics on various viewpoints such as contributors, new Wikipedians, article count, and so on. This tool also facilitates users to analyze ILLs. Meanwhile, detail analysis such as link patterns of ILLs is not provided. In this paper, we analyze the detail of ILLs.

Adler proposed an evaluation method of Wikipedia contents based on the revision history of an article[5]. They applied their method to Italian and French Wikipedias, but links across languages is not considered. In this paper, we focus on link patterns of ILLs.

Krizhanovsky proposed a system called *Synarcher*¹⁰ which is a system for browsing synonyms based on Wikipedia entries[6]. This system finds synonyms by applying adapted HITS algorithm to Wikipedia entries. Although this system focuses on synonyms in a specific language, we focus on cross-lingual keyword navigation using ILLs (see section 5).

3 Analysis results of interlanguage-links of Wikipedia

In this section, we describe (1) an overview of the data used in this study, and (2) link patterns of ILLs obtained from our analysis.

⁸ <http://stats.wikimedia.org/EN/Sitemap.htm>

⁹ <http://infodisiac.com/>

¹⁰ <http://sourceforge.net/projects/synarcher>

Table 1. Wikipedia dump data used in this analysis

Language	Dump date	Size (MByte)
Chinese	October 14, 2007	700
Japanese	October 13, 2007	2,399
Korean	October 11, 2007	176
English	October 18, 2007	13,049

Table 2. Number of articles and ILLs in CJKE Wikipedias

Language	(upper) # of standard articles (lower) # of articles containing ILLs	# of ILLs
	232,669	
Chinese	122,226 (52.5%)	1,536,757
	569,836	
Japanese	211,390 (37.1%)	2,050,491
	73,782	
Korean	54,797 (74.3%)	1,061,280
	3,552,823	
English	895,235(25.2%)	4,072,516

3.1 Data

As data source, we used Wikipedia data dumps that can be obtained from Wikimedia¹¹. We downloaded `pages-articles.xml.bz2` of CJKE editions. Table 1 shows the Wikipedia dump data used in this study. We extracted ILLs by using MediaWiki’s PHP functions such as `Title::newFromText`¹².

Table 2 shows the number of articles and ILLs found in the data. In this table, the number of standard articles shows the number of articles except for special pages. The number of ILLs shows the total unique number of ILLs in each edition.

3.2 Destination languages of ILLs

We analyzed the destination of ILLs in each edition. Table 3 to Table 6 shows the top 20 destination languages of ILLs in each edition. We can see that English is appeared at the top of CJK tables (Table 3 to Table 5). In English edition, German, French, Dutch, Italian, and Spanish can be seen at the top of the table.

In Chinese edition, there are 232,669 standard articles¹³, and 52.5% of these articles (122,226 articles) have ILLs. The number of ILLs is 1,536,757. Table 3 shows the top 20 destination languages from Chinese edition. We can see English, Japanese, German, French, and Spanish at the top five destination languages.

¹¹ <http://download.wikimedia.org/>

¹² <http://svn.wikimedia.org/doc/>

¹³ This does not contain redirect articles.

Table 3. Destination languages of ILLs in Chinese edition

	Destination language	# of ILLs
1	English	107,358
2	Japanese	70,274
3	German	66,524
4	French	66,378
5	Spanish	48,955
6	Polish	48,912
7	Dutch	46,903
8	Portuguese	44,499
9	Russian	43,282
10	Swedish	42,507
11	Italian	42,040
12	Norwegian	35,464
13	Finnish	32,959
14	Korean	32,959
15	Cesky	27,811
16	Slovak	31,534
17	Bahasa Indonesia	24,947
18	Danish	24,896
19	Esperanto	24,473
20	Hebrew	22,986

Table 5. Destination languages of ILLs in Korean edition

	Destination language	# of ILLs
1	English	51,171
2	Japanese	39,827
3	French	38,115
4	German	36,390
5	Chinese	31,391
6	Russian	28,839
7	Spanish	27,749
8	Polish	26,591
9	Dutch	24,841
10	Swedish	24,552
11	Portuguese	22,983
12	Italian	22,933
13	Norwegian	22,077
14	Finnish	20,772
15	Slovak	19,989
16	Cesky	19,980
17	Slovenian	17,959
18	Bahasa Indonesia	17,822
19	Turkish	17,775
20	Danish	17,104

Table 4. Destination languages of ILLs in Japanese edition

	Destination language	# of ILLs
1	English	195,154
2	French	113,410
3	German	113,295
4	Polish	78,697
5	Dutch	78,209
6	Italian	76,653
7	Spanish	74,809
8	Portuguese	69,859
9	Chinese	68,497
10	Swedish	63,270
11	Russian	55,766
12	Finnish	49,458
13	Norwegian	45,248
14	Korean	39,801
15	Cesky	35,257
16	Esperanto	35,178
17	Hebrew	31,584
18	Danish	31,364
19	Bahasa Indonesia	28,256
20	Slovak	28,234

Table 6. Destination languages of ILLs in English edition

	Destination language	# of ILLs
1	German	359,290
2	French	332,634
3	Dutch	235,502
4	Italian	199,809
5	Spanish	193,848
6	Japanese	191,587
7	Polish	190,102
8	Portuguese	185,492
9	Swedish	148,799
10	Russian	131,043
11	Finnish	102,445
12	Chinese	102,317
13	Norwegian	97,976
14	Cesky	63,569
15	Esperanto	63,044
16	Volapuk	59,137
17	Hebrew	56,637
18	Danish	54,143
19	Korean	51,563
20	Slovak	50,194

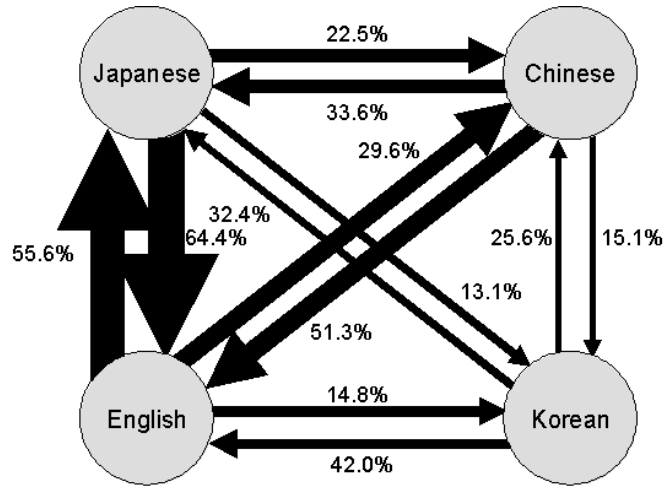


Fig. 1. Connection state of ILLs among Chinese, Japanese, Korean, and English.

In Japanese edition, there are 569,836 articles¹³, and 37.1% of these articles (211,390 articles) have ILLs. The number of ILLs is 2,050,491. Table 4 shows the top 20 destination languages from Japanese edition. We can see English, French, German, Polish, and Dutch in the top five of the destination language. We can find that Chinese (9th) and Korean (14th) are not appeared at the top of the table although these languages are geographically and historically related.

In Korean edition, there are 73,782 articles¹³, and 74.3% of these articles (54,797 articles) have ILLs. The number of ILLs is 1,061,280. Table 5 shows the top 20 languages connected from Korean edition. We can see English, Japanese, French, Germany, and Chinese in the top five of the destination languages.

In English edition, there are 3,552,823 articles¹³, and 25.2% of these articles (895,235 articles) have ILLs. Number of ILLs is 4,072,516. Table 6 shows the top 20 destination languages. We can see German, French, Dutch, Italian, and Spanish in the top five destination languages.

3.3 Connection state of ILLs among CJKE Wikipedias

We also analyzed the connection state of ILLs among CJKE Wikipedias. Figure 1 shows the connection rates of ILLs. We can see that English edition plays a hub role because Chinese, Japanese, and Korean editions have ILLs towards English edition. For example, 42.0% of ILLs in Korean edition are classified into ILLs towards English. In case of Japanese, 64.4% of ILLs are connected to English¹⁴.

¹⁴ Note that the percentage of ILLs in each edition is valid in CJKE editions, i.e., we did not count ILLs towards/from other languages except for CJKE.

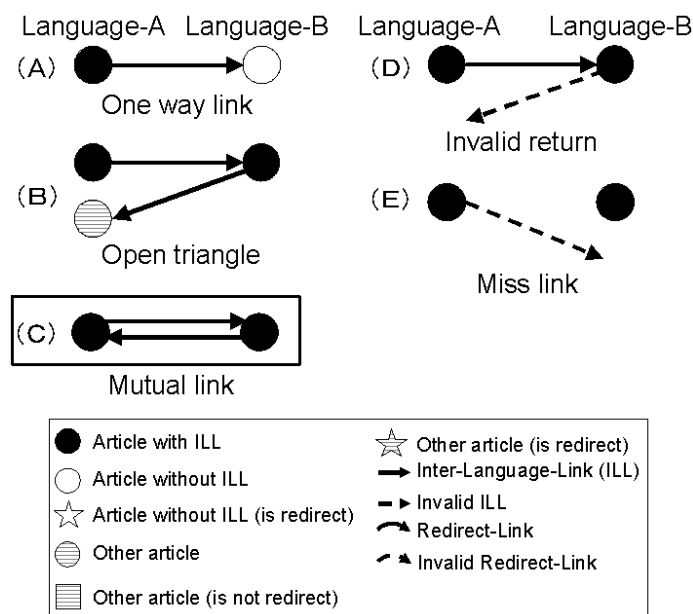


Fig. 2. List of ILL patterns.

3.4 Link patterns of ILLs

We classified link patterns of ILLs into following five patterns.

1. One-way link (*Pattern A*)
2. Open triangle (*Pattern B*)
3. Mutual link (*Pattern C*)
4. Invalid return (*Pattern D*)
5. Miss link (*Pattern E*)

Figure 2 shows the list of patterns, and Table 7 shows the number of ILLs. We explain each pattern in the following.

The first pattern is the *one-way link (Pattern A)*. In this pattern, there is an ILL from language-A to language-B, but there is no return link from language-B. The percentage values of this pattern are from 2.1% (English) to 7.9% (Chinese).

The second pattern is the *open triangle (Pattern B)*. In this pattern, there is an ILL from language-A to language-B, and there is a return link from language-B to language-A. Meanwhile, the return link is not connected to the source entry. This pattern appears rarely, i.e., the percentage values of this pattern are from 0.5% (Korean) to 2.7% (English).

The third pattern is the *mutual link (Pattern C)*. Most of ILLs is classified into this pattern (see Table 7). At the best case, 93.2% of ILLs in English edition are connected to other languages mutually. At the worst case, 88.7% of

Table 7. Number of ILLs in each pattern.

Language	Patterns				
	A	B	C	D	E
Chinese	16,356 (7.9%)	3,303 (1.6%)	183,958 (88.7%)	1,605 (0.8%)	2,218 (1.1%)
Japanese	14,508 (4.8%)	4,697 (1.6%)	278,281 (91.9%)	271 (0.1%)	5,208 (1.7%)
Korean	3,517 (2.9%)	661 (0.5%)	114,910 (93.8%)	435 (0.4%)	2,934 (2.4%)
English	7,099 (2.1%)	9,200 (2.7%)	317,971 (93.2%)	331 (0.1%)	6,446 (1.9%)

ILLs in Chinese edition are connected mutually. Therefore, ILLs can be used for constructing multilingual dictionaries.

The fourth pattern is the *invalid return (Pattern D)*. In this pattern, there is an ILL from language-A to language-B, and there is a return link from language-B. Meanwhile, the return link is connected to an invalid (removed) entry of language-A. This pattern is also a rare case, i.e., we can see 0.1% (Japanese, English) to 0.8% (Chinese).

The last pattern is the *miss link (Pattern E)*. In this pattern, there is an ILL from language-A to language-B, but this ILL is connected to an invalid entry of language-B. The percentage values are from 1.1% (Chinese) to 2.4% (Korean).

Detail classification of pattern A and B We further analyzed *Pattern A* and *Pattern B*. The translation rate can be improved by analyzing *Pattern A* and *Pattern B*.

Pattern A can be separated into two patterns according to the destination of a link (see Figure 3). *Pattern A-1* can connect to other articles by using the redirection. We can extract ILLs from destination articles, if those articles have ILLs. Meanwhile, ILLs can not be extracted from *Pattern A-2*. In our analysis, 26% of *Pattern A* was classified as *Pattern A-1*.

The *Pattern A-1* can be further classified into a *Pattern A-1-4* from a *Pattern A-1-0*. In our analysis, 44% of *Pattern A-1* is classified into *Pattern A-1-4*. 56% of *Pattern A-1* is classified into *Pattern A-1-3*. Thus, 44% of *Pattern A-1* was an indirect-mutual-link.

We also analyzed *Pattern B*. 31% of *Pattern B* was *Pattern B-1* (see Figure 4). An indirect mutual-link such as *Pattern B-1-2* was 89%. Thus, almost of *Pattern B-1* is an indirect-mutual-link. Therefore, about 31% of *Pattern B* can be repaired. By repairing these ILLs, translation rates can be improved.

4 Qualitative analysis of ILLs

We evaluated ILLs by comparing with dictionaries. For comparison of the translation-words between translation-dictionaries and ILLs, we randomly sampled 200 ILLs

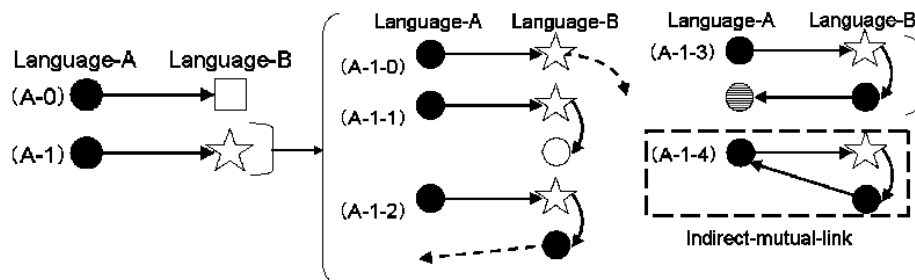


Fig. 3. Detail analysis of pattern A

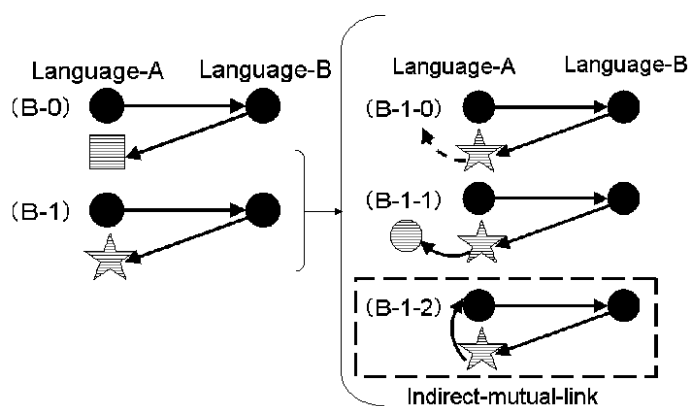


Fig. 4. Detail analysis of pattern B

that of *Pattern C* in each edition, and we manually classified 200 ILLs into following three categories.

- (1) **Unregistered** An entry is not registered in dictionaries.
- (2) **Complete match** Translation-words appear as ILLs is registered as same as an entry in a dictionary.
- (3) **Partial match** An entry is registered in dictionaries, but translation does not matched perfectly.

Table 8 shows the results of comparison for the category (1). In case of $J \Leftrightarrow E$, 149 out of 200 entries, In case of $J \Leftrightarrow C$, 170 of 200 entries, In case of $C \Leftrightarrow K$, 189 of 200 entries are unregistered in a translation-dictionary. These entries are translatable by using ILLs.

For the detail analysis, we classified unregistered-words into two subcategories.

- (1-1) **Partial translation is possible** The entry which consists of two or more words. It can be translated in a existing dictionary.

Table 8. Qualitative evaluation of ILLs using translation-dictionaries

	Unregistered	Complete match	Partial match
Japanese \Leftrightarrow English	149	42	9
Japanese \Leftrightarrow Chinese	170	21	9
Chinese \Leftrightarrow Korean	189	11	0

Table 9. Partial translation of the unregistered-word

	Possible	Impossible
Japanese \Leftrightarrow English	96	53
Japanese \Leftrightarrow Chinese	120	50
Chinese \Leftrightarrow Korean	129	60

(1-2) Partial translation is impossible The entry which consists of two or more words. It can not translate in a existing dictionary.

Table 9 shows the results of classification. In case of J \Leftrightarrow E, 96 of 149 entries, In case of J \Leftrightarrow C, 120 of 170 entries, In case of C \Leftrightarrow K, 129 of 189 entries can be translated partially. Their entries can be called *complex-terms*. And, the entry which cannot translate partially can be called a *new-word*. Thus, by using ILL a lot of complex-terms and new-words can be translatable.

5 Cross-lingual keyword navigation system using ILLs

As an application of ILLs, we propose a cross-lingual keyword navigation system called ILL Visualizer. Figure 5 shows a screen image of the system. By using ILL visualizer, one can browse and explore ILLs across languages on a web browser. Users can also specify a keyword, and find translations for the keyword. Users can find categories to which translations belong. The system is available on the Web¹⁵.

In Figure 5, an input word ‘Video game’ is provided to the system. Users can find translations for the input word in other languages. Furthermore, users can find synonyms for the input word because the system follows redirect of an entry. In the figure, users can find translations for the input word and ‘Personal_computer_games’ can be seen as a synonym.

6 Conclusion

In this paper, we described analysis results of ILLs of Chinese, Japanese, Korean, and English editions of Wikipedias. We found five patterns of ILLs, and the most of ILLs are classified into the mutual link. We also analyzed of pattern A and B,

¹⁵ <http://arai.cdl.im.dendai.ac.jp/>

and found that some of patterns can be corrected automatically. We described a cross-lingual keyword navigation system called ILL Visualizer as an application using ILLs. Our future work contains (1) to incorporate other languages, and (2) to analyze temporal evolution patterns of ILLs.

References

1. Gey, F.C., Kando, N., Lin, C.Y., Peters, C.: New directions in multilingual information access. *SIGIR Forum* **40**(2) (2006) 31–39
2. Fukuhara, T., Utsuro, T., Nakagawa, H.: Cross-lingual concern analysis from multilingual weblog articles. In Nijholt, A., Stock, O., Nishida, T., eds.: *Proceedings of the 6th Workshop on Social Intelligence Design*. (2007) 55–64 (ISSN: 1574-0846).
3. Ortega, F., Gonzalez-Barahona, J.M., Robles, G.: On the inequality of contributions to wikipedia. In: *HICSS '08: Proceedings of the Proceedings of the 41st Annual Hawaii International Conference on System Sciences*, Washington, DC, USA, IEEE Computer Society (2008) 304
4. Geser, H.: From printed to “wikified” encyclopedias: sociological aspects of an incipient cultural revolution. Technical report, Sociology at the University of Zurich (2007)
5. Adler, T.B., de Alfaro, L.: A content-driven reputation system for the wikipedia. In: *WWW '07: Proceedings of the 16th international conference on World Wide Web*, New York, NY, USA, ACM Press (2007) 261–270
6. Krizhanovsky, A.: Synonym search in wikipedia: Synarcher. In: *The eleventh International Conference “Speech and Computer” SPECOM'2006*. (2006) 474–477

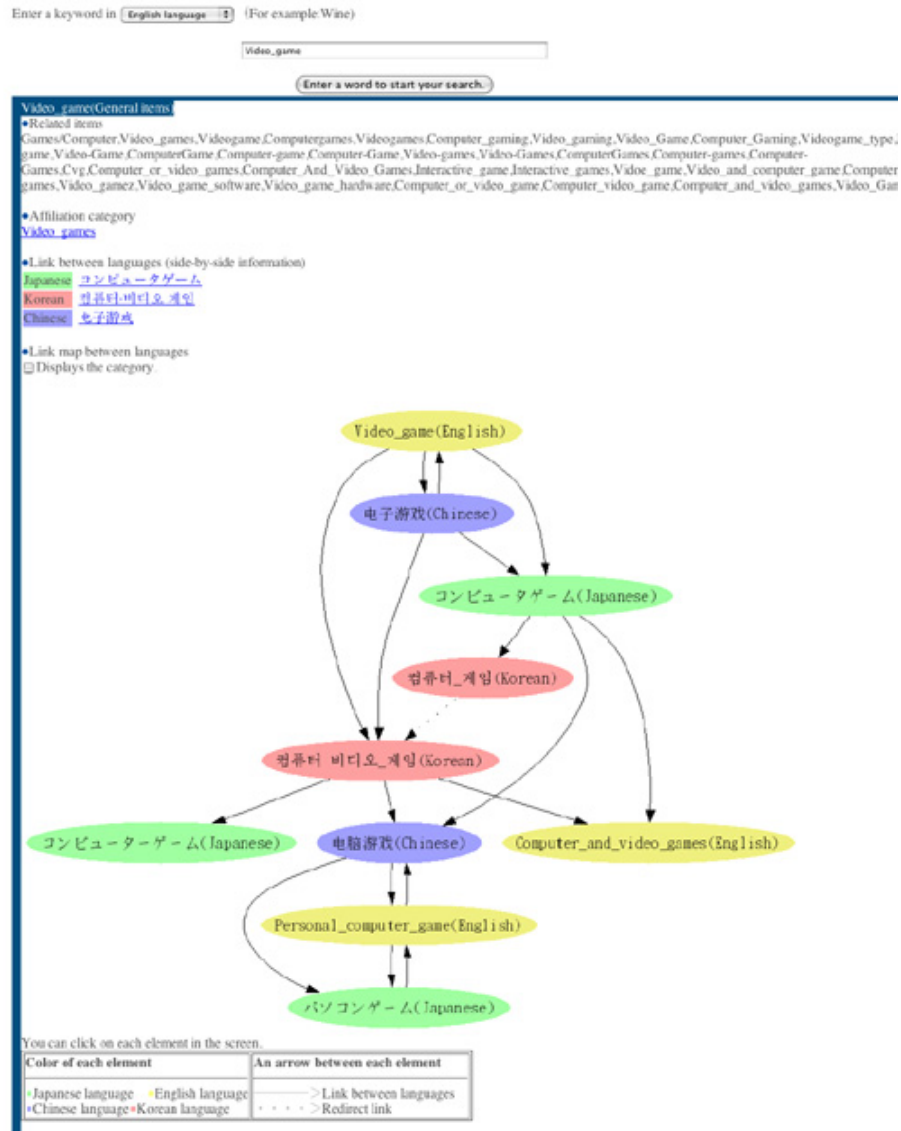


Fig. 5. Screen image of ILL visualizer. One can find the connection state of ILLs in a graph.