



```
011010101011010101001
010101010111110101010
000010111010101001001
101010101010110101001
101010101010101010010
101010101110101010101
010101010101001010100
1010101010010101001
01010101001001010010
```

Analyzing Interlanguage Links of Wikipedias

Yoshiaki Arai,¹ Tomohiro Fukuhara²

Hidetaka Masuda¹, Hiroshi Nakagawa²

¹ Tokyo Denki University

² The University of Tokyo

Outline

- Introduction
- Interlanguage Links (ILLs)
- Analysis results of ILLs
- Cross-lingual keyword navigation system using ILLs
- Conclusion and Future work

Introduction

Multilingual Resources on the Web

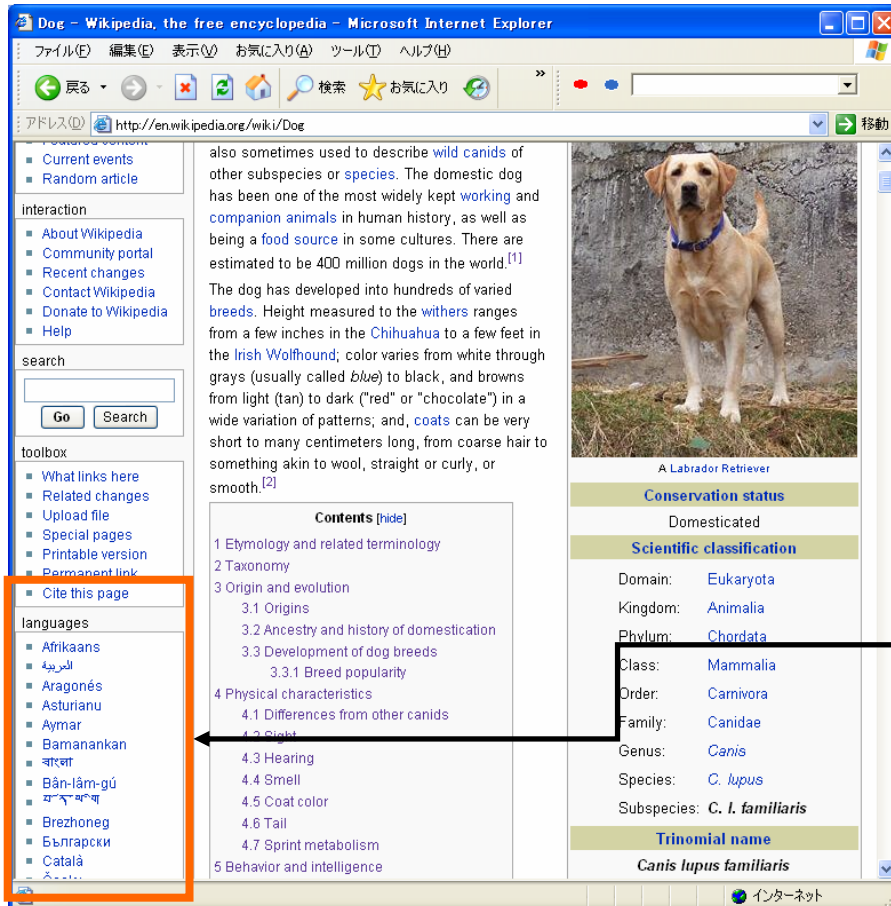
- There is a large number of documents on the Web.
 - It is said that there are 30 to 60 billion of documents on the Web*
- There are many languages on the Web.
 - Content was in English: 56.4%; next were pages in German (7.7%), French (5.6%), and Japanese (4.9%).
 - A more recent study, which used Web searches in 75 different languages are used on the Web.
- Therefore, monolingual search is not enough for utilizing the full knowledge in the world.

Multilingual search and translation are needed.

*`WWW' entry of English Wikipedia

Interlanguage-links of Wikipedias

- Wikipedia is one of huge encyclopedia available on the Web.
 - Over 7 million articles in over 250 languages, and still growing.



Interlanguage links (ILLs) are links from any page in one Wikipedia language to the same subject in another Wikipedia language. ILLs are applicable to multilingual translation.



- languages
- Afrikaans
 - العربية
 - Aragonés
 - Asturianu
 - Aymar
 - Bamanankan
 - বাংলা
 - Bân-lâm-gú

The aim of this research

- Questions
 - How many ILLs exist?
 - How about the quality of ILLs?
 - Who create and repair ILLs?
 - Which topics are rich in ILLs?
 - How complicated are ILLs?
 - Can we use ILLs for keyword translation in Web search?

- To understand the volume and quality of ILLs.

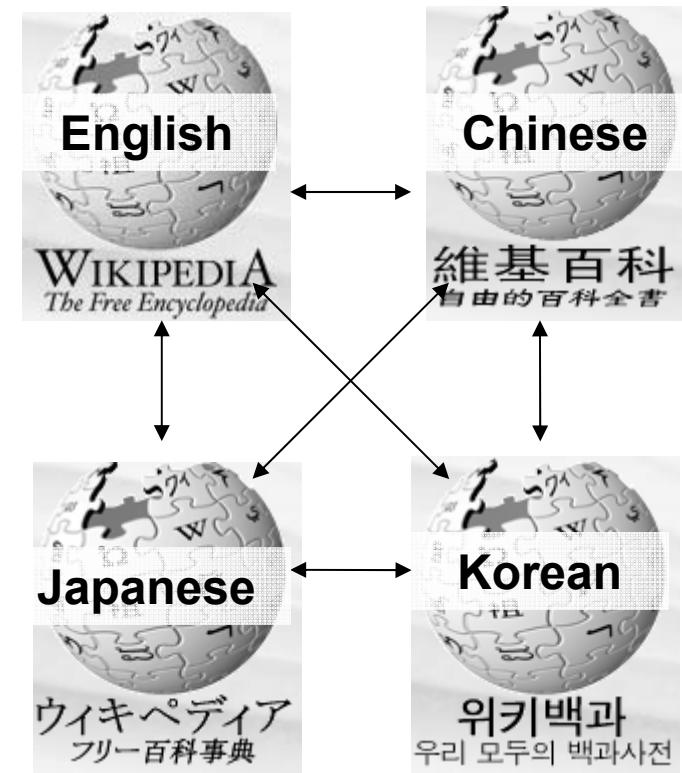
Analysis results of ILLs

1. Link patterns of ILLs
2. Number of ILLs
3. Evaluation of ILLs using dictionaries.

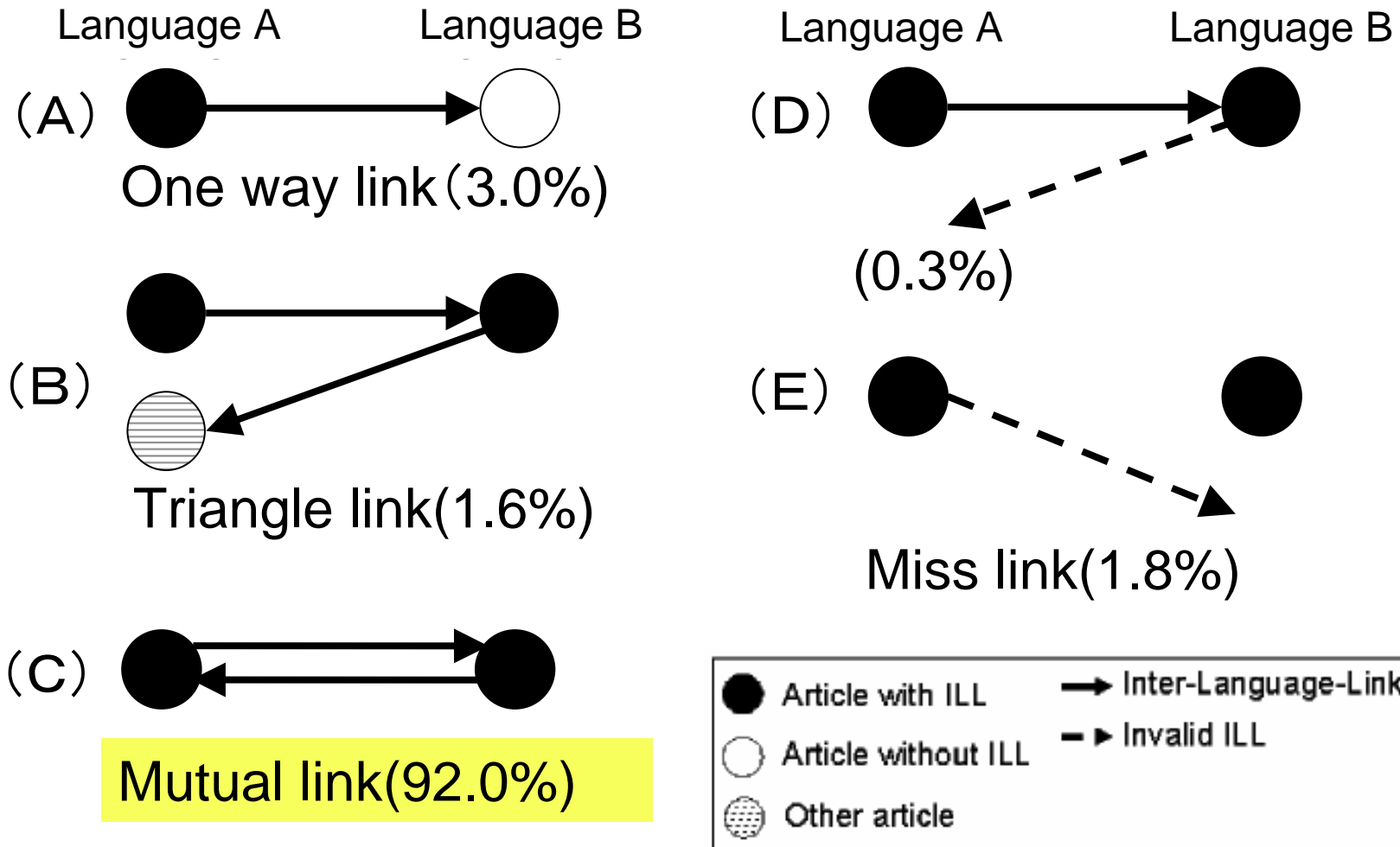
Download from www.mediawiki.org

Editions of Wikipedias	Date	File Size (MByte)
Chinese	October 14, 2007	700
Japanese	October 13, 2007	2,399
Korean	October 11, 2007	176
English	October 18, 2007	13,049

*Chinese data contains both of simplified and traditional Chinese.



1. Link patterns of ILLs



A numerical value is what averaged the value analyzed in 4 languages.

2.Number of ILLs

	A Number of Artciles* (with ILL)	B Number of Artciles* (All)	D Number of ILLs (All)	E (A/B × 100)
English Wikipedia	895,235	5,836,167	4,072,516	15%
Japanese Wikipedia	211,390	808,514	2,050,491	26%
Chinese Wikipedia	122,226	352,533	1,536,757	35%
Korean Wikipedia	54,797	93,850	1,061,280	58%

*Articles in all namespaces are used.

2.Number of ILLs

Detail analysis of Interlanguage-links by languages (English)

Details of English ILLs.

	Language	Number of ILLs	%
1	Germany (de)	359,290	8.8
2	French (fr)	332,634	8.2
3	Dutch (nl)	235,502	5.8
4	Italian (it)	199,809	4.9
5	Esperanto (es)	193,848	4.8
6	Japanese (ja)	191,587	4.7
7	Polish (pl)	190,102	4.7
8	Portuguese (pt)	185,492	4.6
9	Swedish (sv)	148,799	3.7
10	Russian (ru)	131,043	3.2



2. Number of ILL

Details analysis of Interlanguage-links by languages (CJK)

Details of ILLs in Chinese.

Language	Number of ILLs	%
English (en)	107,358	7.0
Japanese (ja)	70,274	4.6
Germany (de)	66,524	4.3
French (fr)	66,378	4.3
Esperanto (es)	48,955	3.2
Polish (pl)	48,912	3.2
Dutch (nl)	46,903	3.1
Portuguese (pt)	44,499	2.9
Russian (ru)	43,282	2.8
Swedish (sv)	42,040	2.8

Details of ILLs in Japanese.

Language	Number of ILLs	%
English (en)	195,154	9.5
French (fr)	113,410	5.5
Germany(de)	113,295	5.5
Polish (pl)	78,697	3.8
Dutch (nl)	78,209	3.8
Italian (it)	76,653	3.7
Esperanto (es)	74,809	3.6
Portuguese (pt)	69,859	3.4
Chinese (zh)	68,497	3.3
Swedish (sv)	63,270	3.1

Details of ILLs in Korean.

Language	Number of ILLs	%
English (en)	51,171	4.8
Japanese (ja)	29,827	3.8
French (fr)	38,115	3.6
Germany (de)	36,390	3.4
Chinese (zh)	31,391	3.0
Russian (ru)	28,839	2.7
Esperanto (es)	27,749	2.6
Polish (pl)	26,591	2.5
Dutch (nl)	24,841	2.3
Swedish (sv)	24,552	2.3

2.Number of ILLs

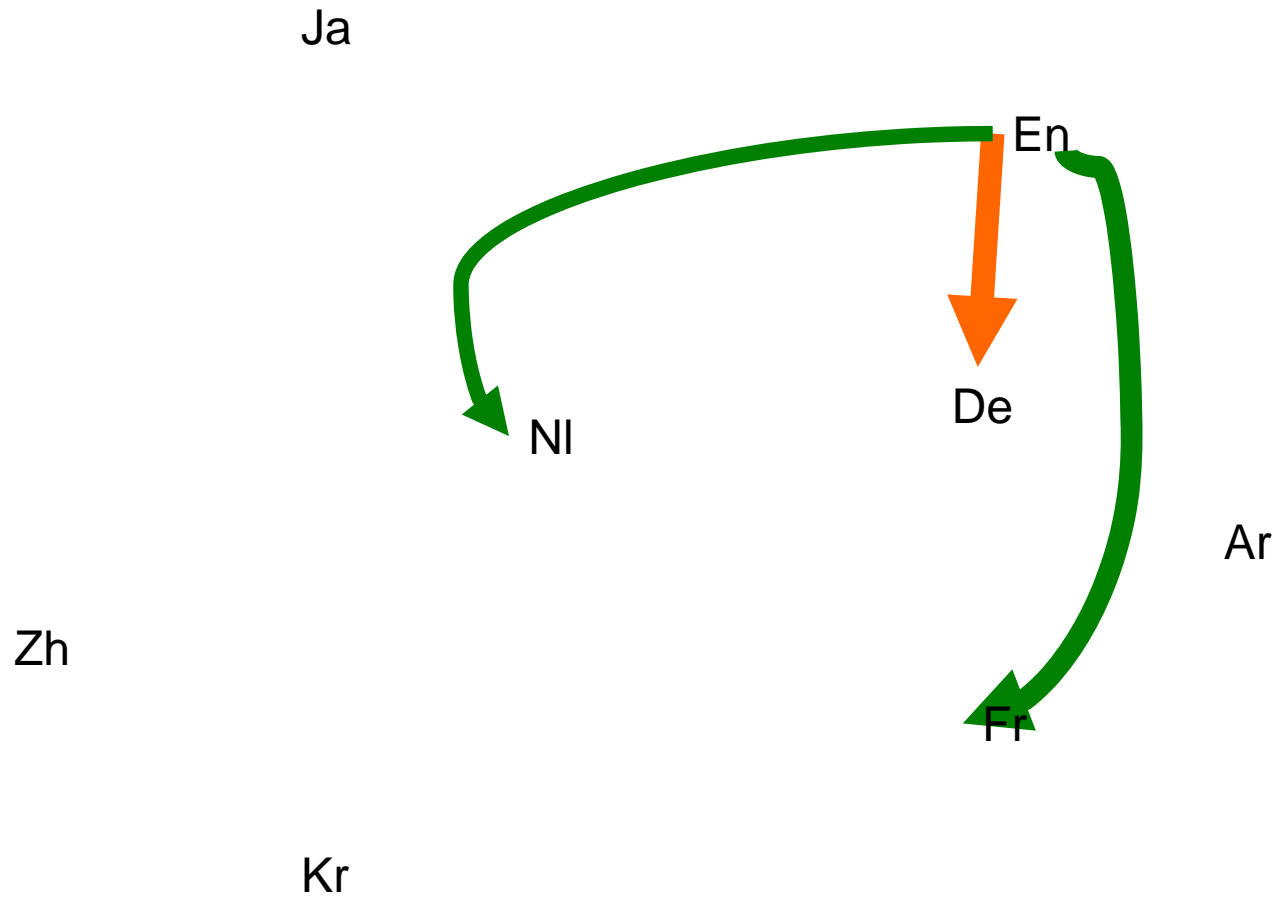
Details analysis of Interlanguage-links by languages (Arabic)

Details of ILLs in Arabic.

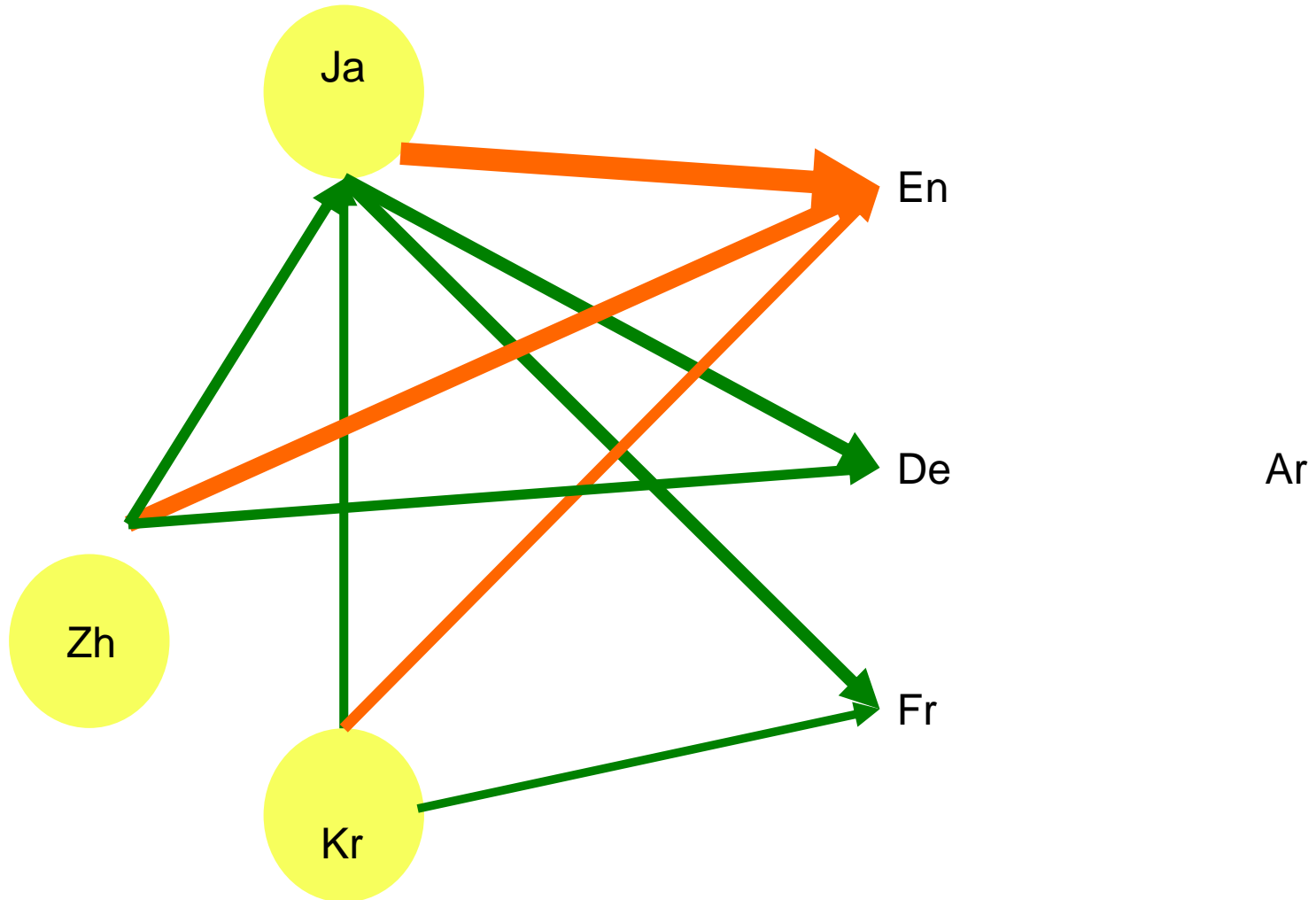
Language	Number of ILLs	%
English (en)	47,843	4.2
French (fr)	36,574	3.2
Germany(de)	28,805	3.1
Esperanto (es)	28,243	2.5
Dutch (nl)	28,120	2.5
Japanese (ja)	28,120	2.5
Italian (it)	27,505	2.4
Polish (pl)	27,241	2.4
Swedish (sv)	26,661	2.3
Chinese (zh)	26,387	2.3

2. Number of ILLs

Who follows whom? (English)

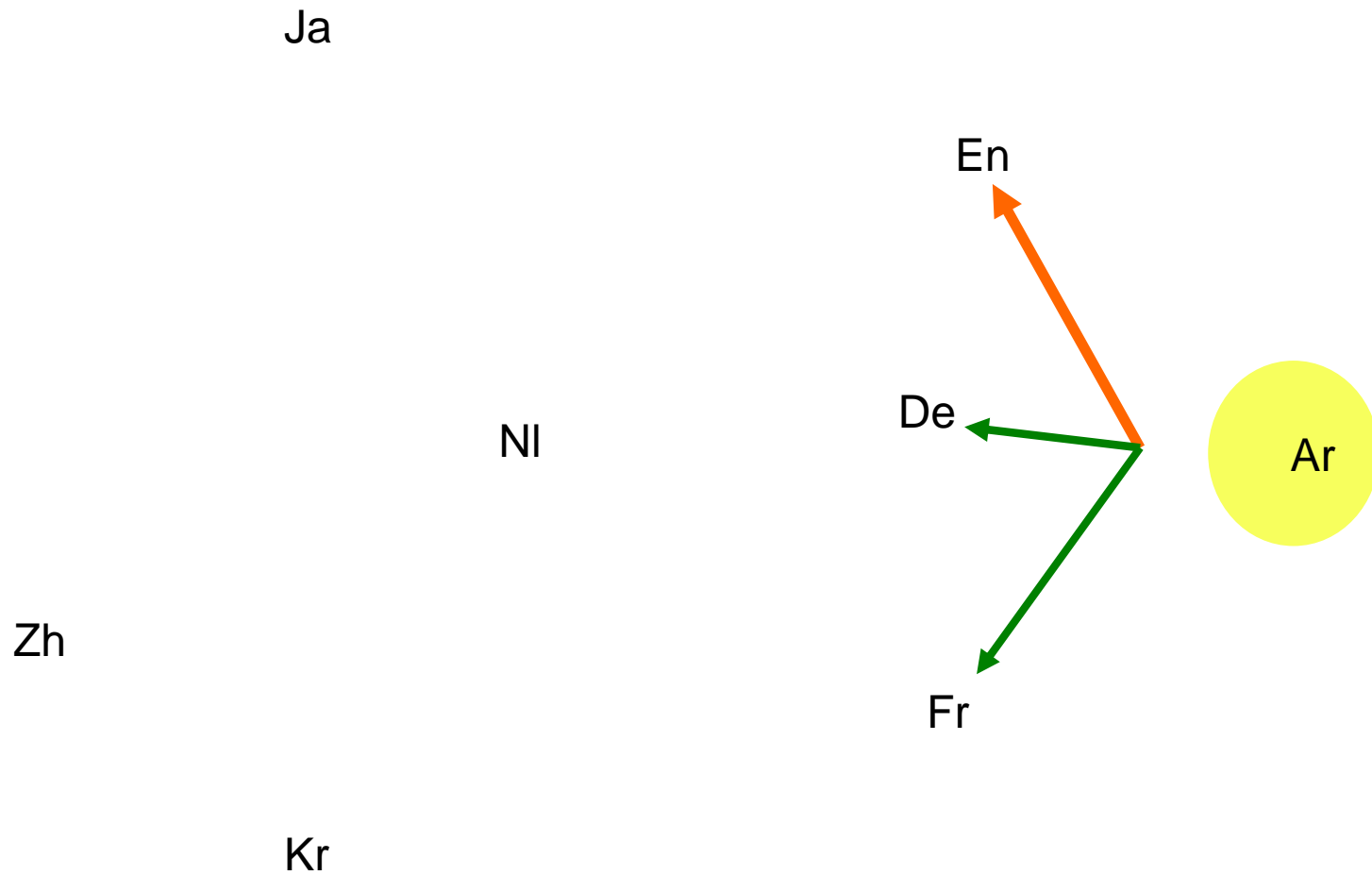


2. Number of ILLs Who follows whom?(CJK)



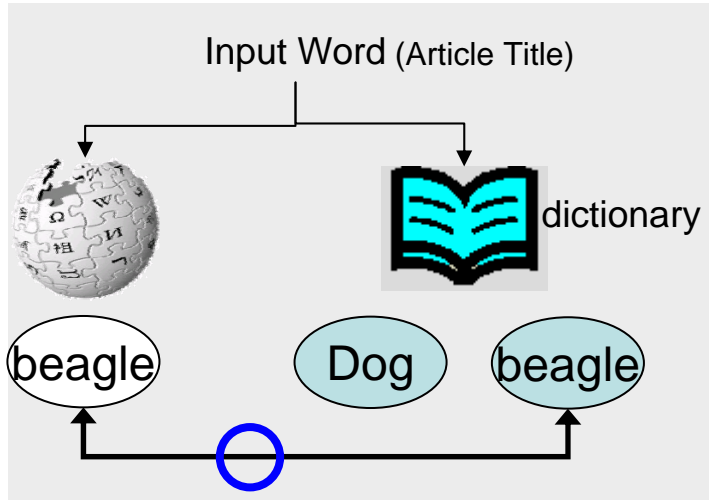
2. Number of ILLs

Who follows whom?(Arabic)

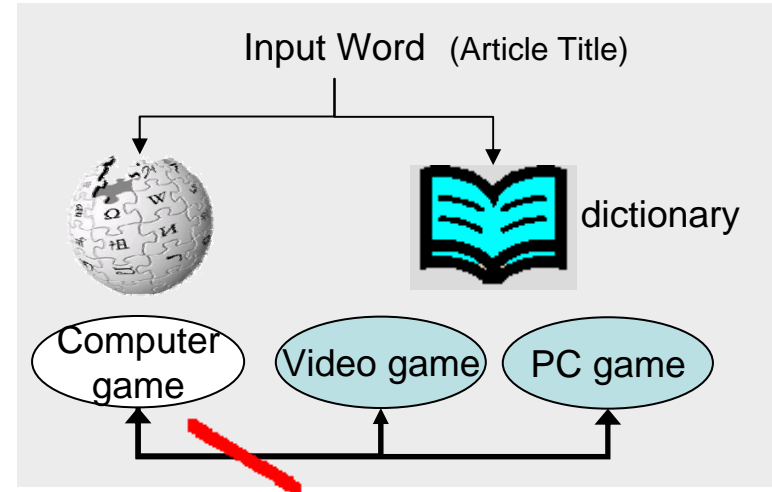


3. Comparison with dictionaries.

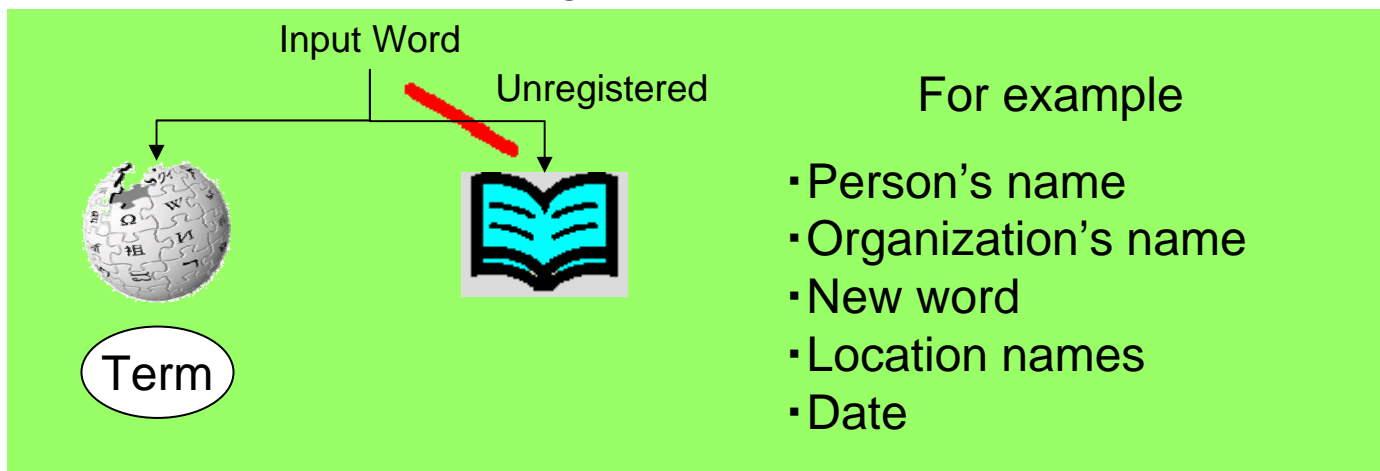
① Exact Match



② Non-Match



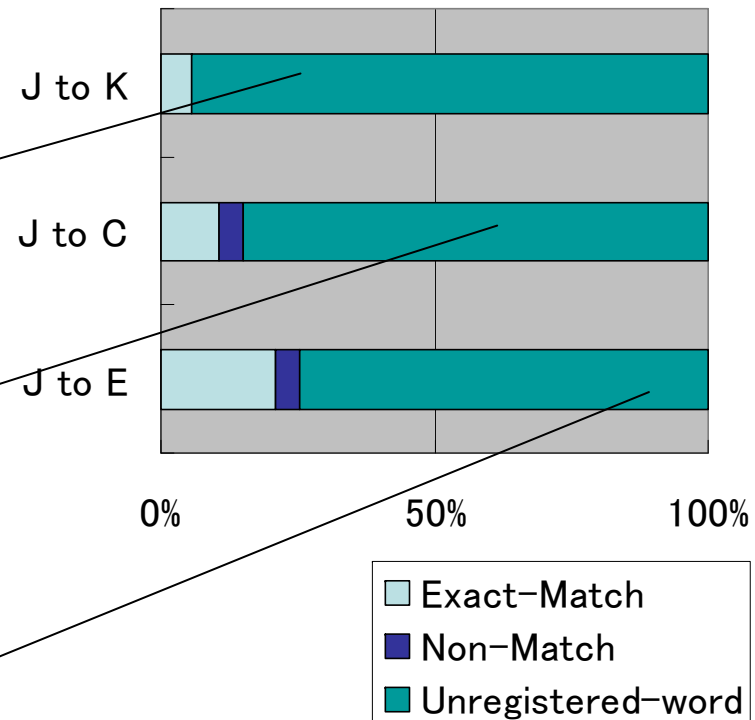
③ Unregistered-word



3. Comparison with dictionaries

Results of evaluation

	① Exact-Match	② Non-Match	③ Unregistered -word
J to E	42 (21%)	9 (4.5%)	149 (75%)
J to C	21 (11%)	9 (4.5%)	170 (85%)
J to K	11 (5.5%)	0 (0%)	189 (95%)



Cross-lingual keyword navigation system using ILLs

System interface

Dog

Enter a word to start your search.

Dog(General items)

•Related items

Canis familiaris.Dogs.Canis lupus familiaris.Canis Canis.Domestic Dog.

Man's_best_friend,A_man's_best_friend,Duppie,Doggy,Dog_(Domestic),Dog_groups,Dogs_as_our_pets,Domestic_dog,

•Affiliation category

[Dogs](#).[Animals kept as pets](#).[Cosmopolitan species](#)

•Link between languages (side-by-side information)

Korean 개

Japanese イヌ

Chinese 犬

•Link map between languages

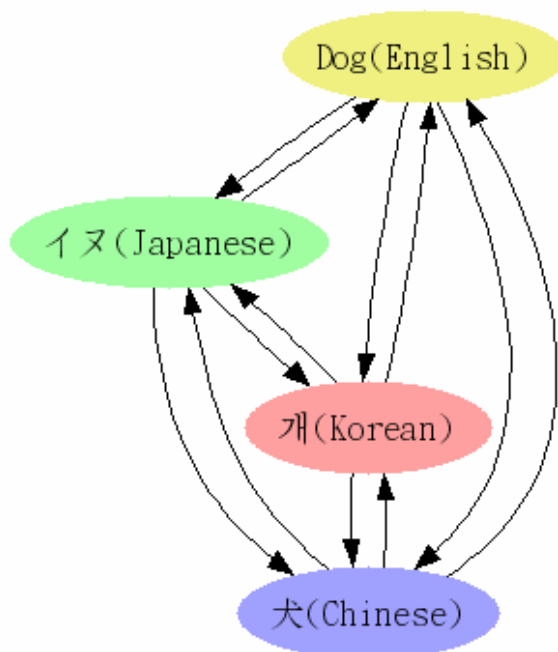
Displays the category.

Related-terms

Categories

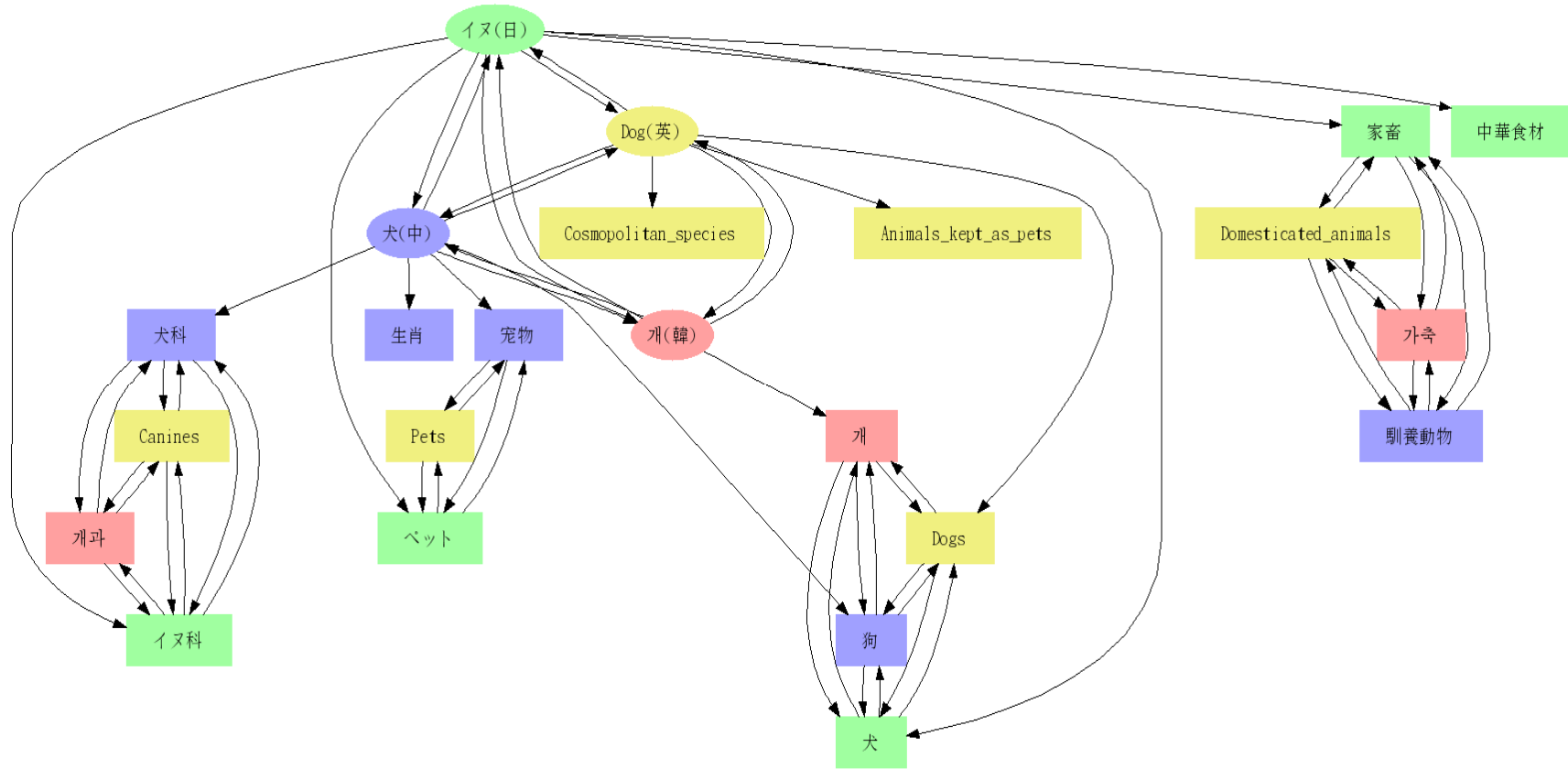
Translation-word
(JCKE)

ILL link state
(JCKE)



<http://arai.cdl.im.dendai.ac.jp/>

Visualization of Category with ILL



Conclusion

- We analyzed Interlanguage-Links of Wikipedias.
- We found:
 - five patterns of ILLs.
 - 92% of ILLs are mutual link.
 - 75 to 95% of ILLs are Unregistered in dictionaries.
- Future work
 - To Import and analyze other language data.
 - To Create multilingual Web search tool by using ILLs.

Thank you very much for your attention.

清听谢谢!

청취 감사합니다

ご清聴ありがとうございました

Grazie per la Sua attenzione!

Danke für Ihre Aufmerksamkeit!

Merci pour votre attention!

Gracias por su atención!

Obrigado por sua atenção.

أشْكُرُكَ شُكْرًا

С п а с и б о з а в н и м а н и е .