Building a Cross-lingual Referential Knowledge Database using Dictionaries

Shigeko Nariyama*[•], Takaaki Tanaka*, Eric Nichols[υ],
Francis Bond*, Hiromi Nakaiwa*
Communication Science Lab, NTT Kyoto, Japan
*{shigekon, takaaki, bond, hiromi}@cslab.kecl.ntt.co.jp
[•] The University of Melbourne, Australia shigeko@unimelb.edu.au
[υ] Nara Institute of Science and Technology, Nara, Japan eric-n@is.naist.jp

Abstract

Referential knowledge is vital for resolving various problems in NLP, such as anaphora resolution. For example, we have the referential knowledge that *diagnose* is most likely a member of the referential relation '*doctor* diagnose *patient's illness'*. Nariyama et al. (2005) presented an inventory of such referents as *doctor,* collected from Japanese dictionary definition sentences. Such referential information is based on world knowledge and is applicable across languages. This paper describes our work using the inventory to build a cross-lingual referential database for multilingual applications.

1 Introduction

Natural language can be highly ambiguous. Utterances tend to avoid repeating information that is deducible from context or world knowledge. Furthermore, individual words include multiple senses (as opposed to being limited to one sense per word).

Various problems in NLP deriving from these ambiguities, such as anaphora resolution and word sense disambiguation, have been known to be prohibitively difficult to solve. The difficulties lie in the fact that the resolutions of these problems rely heavily on contextual information and world knowledge, for which even the state of the art in NLP cannot adequately account.

Nonetheless, words contain in their lexical semantics a large amount of inferences and entailments. When we hear words, we tend to make a strong association with certain referents. For example, the word *diagnose* prototypically appears in the referential relation, '*doctor* diagnose *patient's illness'*. With the word *diagnose* alone, we strongly associate two referents: one referent that is *doctor* as the subject of the sentence and another *patient's illness* as its object. Similarly, *arrest* prototypically has the referential relation as '*police* arrest *criminal*'. We refer to such relations of referents with a predicate as *referential knowledge.*

We contend that referential knowledge is a kind of contextual information or world knowledge, and it can be extracted from dictionary definition sentences. As such, referential knowledge captures referential relations of words based on heuristic and provides what we term 'representative arguments'.

Nariyama et al. (2005) presented an inventory of such representative arguments as *doctor* and *patient's illness* for *diagnose,* and *police* and *criminal* for *arrest,* collected from Japanese dictionary definition sentences. This referential knowledge makes a great contribution to resolving the aforementioned problems in NLP, such as zero pronoun resolution for languages, including Japanese and Chinese, that do not verbalise many referents (Isozaki and Hirano 2003, Nariyama 2003).

Since much of such referential information has its basis in world knowledge, it is quite possible to apply them across languages. In other words, if a language has an equivalent word of *arrest,* then the language is likely to use it with two referents: *police or person with a related authority* as the agent and *criminal or person with a suspect of crime* as the patient.

Arrest = arrestation (French), anhalten (German),
а р е с т о в а н и е (Russian),
검거 (Korean), 拘捕(Chinese), …

This paper describes our work using this inventory of representative arguments towards building a cross-lingual referential knowledge database that can be used in various languages. This database will save an enormous amount of work by eliminating the necessity to go through various steps for extracting representative arguments for each language. Moreover, the transfer of representative arguments to other languages automatically creates links among the languages, which is useful for multilingual applications.

Section 2 reviews the work on the inventory of the Japanese representative arguments (Nariyama et al. 2005). Section 3 examines the feasibility and the methods of building a cross-lingual referential knowledge database using the inventory. Section 4 describes related research, followed by Conclusions.

2 Inventory of representative arguments

Nariyama et al. (2005) presented an inventory of representative arguments collected from the Japanese semantic database, Lexeed (Bond et al. 2004). This is a hand-built self-contained lexicon, consisting of words and their definitions for the most familiar 28,000 words, as measured by native speakers, comprising a total of 46,347 different senses. This set is large enough to include most basic level words and covers 72.2% of the words in a typical Japanese newspaper.

Lexeed has been enhanced by manual word sense disambiguation of all the open class words. Furthermore, the senses are linked in an ontology (Nichols et al. 2005), which allows us to measure the semantic distance between words or senses using a variety of methods.

We see several advantages in using dictionary definition sentences for collecting referential knowledge. Dictionaries are created to provide information about words from cross-domain in lay terms with little contextual information to be comprehensible, while often providing world knowledge as well. For example, Lexeed provides the following definition about the word *taiho* 逮捕 'arrest', whereby we extract the referential information *police officer* and *criminal*. These extracted referents are 'representative arguments', prototypical examples of the real-world referents that are likely to fill the argument slots.

(1) *Taiho: Keisatsu ga hannin o toraeru koto.*
逮捕: 警察が、犯人 を 捕える こと 。
'Arrest: A police officer captures a criminal.'

It is a fact about the real-world that things like *police* are likely to be the subject of the verb *arrest* and things like *criminal* (or *someone with a suspect of crime*) are likely to be its object. These representative arguments can be used as the basis for selectional preferences, which allow room for any rhetorical and other deviated usages.

In general, we should prefer an interpretation where the referents of the arguments are semantically similar to the representative arguments. Because arguments only have to be similar, not subsumed by, it is possible for the representative arguments to be actual words, although word senses would be preferred.

In contrast, processing using selectional restrictions must use broader semantic classes, otherwise non-typical sentences would be rejected. For example, *Goi-Taikei*'s valency dictionary (Ikehara et al. 1997) has the semantic classes *agent* and *person* as selectional restrictions for *taiho* 'arrest'. These semantic classes are derived by most research (see Section 4). These subsume the words *police* and *criminal* but are much less informative.

2.1 Process of extracting arguments

We created an inventory of the representative arguments that are more specific than what is available in *Goi-Taikei (GT)*, currently the most informative resource available in Japanese. The process involved:

1) Automatic extraction of the representative arguments of definition words (i.e. words being defined) that are predicates (i.e. verbs, verbal nouns, and adjectives) from definition sentences in Lexeed, using both Shallow and Deep parsing;
2) Hand-selecting representative arguments from those extracted to make a reliable list;
3) Selecting only those that are more specific than what is provided by *GT*.

Deep parsing (DP) gives us the information we want immediately, but only for those sentences that can be parsed. Shallow parsing (SP), on the other hand, allows us to extract information from more data, but with less precision. For the optimal results, we combined DP and SP to extract arguments for greater quality and quantity. This technique of combining DP and SP has been proposed in the Deep Thought project and proved to be effective (Nichols et al. 2005, inter alia). For DP, we used a combination of the PET parsing system (Callmeier 2002) and the JaCY Japanese HPSG grammar (Siegel and Bender 2002).[1]

2.2 Results

The total number of extracted arguments was 10,076. Of these 6,550 (65.0%) were manually verified as representative arguments that are more specific than those in *Goi-Taikei* or new to *Goi-Taikei*. Table 1 gives the precision (the rate of representative arguments extracted over total extraction) per POS and parsing method. The results are promising.

[1] PET is an open source, highly efficient unification parser. JaCY is broad-coverage, freely available HPSG grammar that produces semantic analysis in Robust Minimal Recursion Semantics (RMRS, Frank 2004). See Nariyama et al. (2005) for detail.
PET can be downloaded at: http://wiki.delph-in.net/moin/PetTop

	Adjective	Verb	Verbal N	All
DP only	69.3%	76.6%	72.8%	74.1%
SP only	49.9%	63.7%	49.9%	55.3%
Extracted by Both	56.8%	72.4%	72.9%	70.0%
Total (number)	**57.8%** (841/1455)	**71.5%** (3041/4252)	**66.0%** (2883/4370)	**67.1%** (6765/10076)

Table 1: Precision per POS and parsing method

Filtering by *Goi-Taikei*

We compared the specificity of the extracted arguments with that of the corresponding words in *Goi-Taikei* (GT) with the following classification. The results are shown in Table 2.

- ▪ > GT: more specific than GT
- ▪ = GT: same as GT
- ▪ no entry of the definition word in GT
- ▪ no sense entry of the definition word in GT
- ▪ < GT: less specific than GT

	Adj.	Verb	VN	All
> GT ▪	**48.8%**	**57.4%**	**46.6%**	***51.7%***
= GT ▪	.8%	3.4%	3.1%	2.9%
no GT entry ▪	41.1%	22.7%	39.8%	32.3%
No sense GT ▪	9.3%	16.2%	10.2%	12.8%
< GT ▪	0%	.3%	.3%	.3%
∑	100% (841)	100% (3,041)	100% (2,883)	100% (6,765)
N(▪ + ▪ + ▪) / ∑ extracted	98.9%	98.2%	98.5%	*96.8% (6,550)*

Table 3: Comparing specificity of extracted arguments with that in Goi-Taikei (GT)

The results show that 51.7% of the arguments we selected provide more specific referential information than those in GT. If those arguments that are not listed in GT are to be included, i.e. ▪ + ▪ , it goes up to 96.8%. In other words, virtually every argument extracted from the proposed method provides new or more specific referential information than what exists in GT.

While there remain many areas of improvements that will increase the precision as discussed in Nariyama et al. (2005), we have extracted 6,550 referents. This result is a promising first step towards building an inventory of representative arguments.

Proportions of sense use

The definition words in Lexeed have one or more senses, with 53 senses being the highest. Lexeed has been enhanced through manual word sense disambiguation. This enables us to measure the proportion of usage of a particular sense of a definition word among the other senses. That tells us how often a definition word is likely to come with the representative arguments extracted. For example, *taiho* 'arrest' has a single sense, while *umu* 'give birth' has two senses, and the sense with the representative arguments *mother* and *child/egg* is used 87.0% of the time in our corpus.

Table 3 shows the average proportions of sense use per POS. 'Mono' refers to words with a single sense (i.e. unambiguous) and the rest having multiple senses. '1st' refers to the most frequently used sense, '2nd' the next, and so forth. We accounted for up to the '3rd' most common sense, where the proportions plateau after the 2nd, 3rd for Verbs.

Figure 1 reports the cumulative frequencies computed from Table 3 by using the ordering: mono>1st>2nd; namely, '+1st' means the total proportion of 'mono' and '1st', and those plus '2nd' is shown by '+2nd'. It shows that many of the representative arguments we extracted have a single sense. For those words with multiple senses, the great majority of the representative arguments appear for either the most frequently used sense or the second highest sense, and few appear with the senses less frequent than the 3rd sense.

\ senses	Mono	1st	2nd	3rd	∑
Adjective	19.4 %	45.2%	16.9%	6.9%	88.4%
Verb	8.1%	28.3%	19.7%	19.7%	75.8%
Verb noun	34.6%	50.8%	12.6%	1.7%	99.7%

***Table 3: The proportions of sense use per POS**: mono-sense, 1^{st} (most frequently used sense), 2^{nd}, 3^{rd}, and Total*

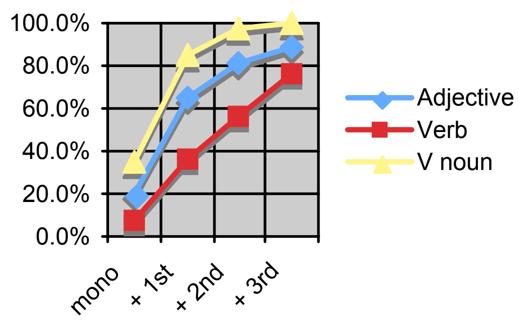

Figure 1: Representation of Table 3

3 Building a cross-lingual referential knowledge database

We aim to create a cross-lingual referential database, whereby the representative arguments for a word are shared across languages. There are two ways to approach this task. The first is to extract representative arguments from each language individually, analogous to the way we did for Japanese. Then we list those definition words that take the same representative arguments shared by other languages. The other approach is to select only those representative arguments from Japanese that have their basis in world knowledge, and to transfer the information across to other languages.

We show in Subsection 3.1 that the observation from English dictionaries and a (Mandarin) Chinese dictionary indicates that the first option is not viable, so the second option should be taken. Accordingly, Subsection 3.2 discusses the process of classifying the Japanese representative arguments into two classes: Language independent referents (i.e. representative arguments that are shared across languages) and Language specific referents. Subsection 3.3 gives verification of the classification in two stages: first through human judgement, and secondly by hand-checking the Language independent referents using the English dictionary. The results are presented in Subsection 3.4.

3.1 English dictionaries

To make a comparison with Lexeed, we examined definition sentences from three machine-readable English electric dictionaries: Oxford Advanced Learner's dictionary 2000, Webster's dictionary 1913 (GCIDE http://www.ibiblio.org/webster/ in the public domain; 130,633 definition words), and Collins Cobuild Advance learner's English dictionary (Fourth edition 2003). In addition, a Chinese dictionary (现代汉语词典 *Xiandai hanyu cidian* 2002 by *Shang-wu-yin-shu-guan*) was referred to in comparison.

As an example, Figures 2a and 2b list the definition sentences for two definition words: *diagnose* and *marry* respectively. It is clear from the definition sentences for English and Chinese that Lexeed provides referential information more concisely, and that automatic extraction of the representative arguments in other languages will be not only difficult but also not fruitful. Most explanations are not sentences, but phrases with infinitive forms. This means that the subjects of sentences, one of the most important sources of representative arguments, are not expressed.

Furthermore, the referents are often very general as 'somebody' and 'something'. Many of the explanations also provide encyclopaedic information, which further complicates the process for automatically extracting representative arguments.

Lexeed: 診断：医師が患者を調べその病状を判断すること．
Shindan: Ishi-ga kanja-o shirabe sono byoujou-o handansuru.
'A doctor examines a patient and gives an option about the patient's illness.'

Oxford: To say exactly what an illness or the cause of a problem is.

Webster: To ascertain by diagnosis; to diagnosticate.

Cobuild: (1) To diagnose an illness or a problem means to discover and identify exactly what is wrong. e.g. 'Doctor has diagnosed it as rheumatism.'

Chinese: 诊断 *zhen-duan* 'diagnose'
在 检查 病人 的 症状 之后 判定 病人 的 病症 及 其 发展 情况
zai- jian-cha bing-ren de zheng-chuang zhi-hou pan ding bing-ren de bing-zheng ji qi fa-zhan qing-kuang
'(lit.) at examine patient DE symptom after decide patient DE disease and its development status'

Figure 2a: Definition sentences for 'diagnose'

Lexeed: 結婚：男女が夫婦になること。
Kekkon: Danjo-ga fuufu-ni naru koto.
'A man and a woman become a married couple.'

Oxford: To become the husband or wife of somebody; to get married to somebody.

Webster: To unite in wedlock or matrimony; to perform the ceremony of joining, as a man and a woman, for life.

Cobuild: If you marry someone, or if you get married, you form a legal relationship with a person of the opposite sex in a ceremony during which you make particular promises to that person and become their husband or wife. EG. 'I want to marry him.'

Chinese: 结婚 *jiehun* 'marriage'
男子 和 女子 经过 合法 手续 结合 成为 夫妻
nan-zi he nv-zi jing-guo he-fa shou-xu jie-he cheng-wei fu-qi
'(lit.) man and women through legal procedure combine become husband and wife'

Figure 2b: Definition sentences for 'marry'

Thus, we opted for taking the second approach; that is, to take those Japanese representative arguments that are judged as language independent referents from the experiments using English dictionaries (see Subsections 3.2 and 3.3) as seeds, and to use the resources from Japanese to bootstrap coverage of other languages.

Although the work up to this stage involves a certain amount of manual work, benefits of this approach are substantial:

1) It enables collection of representative arguments otherwise not possible, because they are not specifically written in the dictionaries of the other languages, or for languages with no dictionaries.

2) It lessens the amount of work on other languages, eliminating the various stages of extractions and hand verifications.

3) It will become even more cost effective as the amount of resources available increases and the number of language transfer increases.

4) It can automatically create a cross-lingual link that is useful for multi-lingual applications.

3.2 Classifying representative arguments

We manually classified the Japanese representative arguments that we extracted into two classes:

(a) **Language independent referents**
(b) **Language dependent referents**

The criteria for the distinction used in this verification are as follows, although they should be more objective and clear, requiring improvements. (a) was classified as such if the representative arguments in Japanese are either [1] not (b), or [2] based on scientific facts (e.g. physics, biology of animals, physiology of humans) and common knowledge (believed to be commonly known or agreed by adults of the world). For example, a word *umu* 生む 'give birth' is likely to have *mother* as the subject of the verb and *child/egg* as its object, and this referential relation is cross-linguistically valid.

The referents under (a) includes 'referent incorporation'. For example, *kyuukon* 求婚 'propose (a marriage)' includes the object 'a marriage' in the definition word in itself, whereas the English equivalent does not.

(b) Language dependent referents mainly involve three types. The definition words express:

(b-1) *Language specific concepts and idioms*: e.g. *katazukeru* (娘を) かたずける 'to get rid of' implies 'to get rid of (one's daughter by marrying her off');

(b-2) *Honorifics and other social ranking*: e.g. *insotsu* 引率 'to take' is used as 'a (higher ranked) person takes a (lower ranked) person', instead of the neutral form *tsurete-iku* 連れて行く 'to take';

(b-3) *Specialised or domain specific terms* e.g. *aisatsu* あいさつ「俳優が観客に」 'to greet' is used as '(In performing arts and theatre plays), the actors/actresses greet the audience'.

3.3 Verification

Using the criteria described in Subsection 3.2, we conducted the following hand verification in two stages in order to ascertain how many of the Japanese representative arguments are identified as 'language independent' referents.

[1] Among the 4,099 representative arguments we extracted that are more specific than *Goi-Taikei*, we have identified only 421 are language specific, i.e. 3,678 arguments (89.7%) are judged as language independent.
[2] 10% of those arguments were randomly selected and hand-checked in Cobuild[2] CD-ROM and/or Google search in order to further verify.

The results are classified as follows:

Language independent

i. Lexeed's referent is found in the Cobuild definition sentence.
ii. When not i, Lexeed's referent is found in the Cobuild example sentence.
iii. When not i or ii, Cobuild lists a referent that is of the same semantic class as Lexeed's referent.
iv. Lexeed's referent is not found in Cobuild, but the collocation of the Lexeed's referent with its definition word exceeds 100,000[3] hits in Google search (which indicates that the referent is most likely to appear with its definition word).

Language specific

v. Lexeed's referent is not found in Cobuild, and the collocation of the referent with its definition word is less 100,000 hits in Google search.
vi. No match of English translation of the definition word, or referent.

We make an assumption that the referents under i, ii, iii, and iv are likely to be language independent for the definition words and that they are readily

[2] Cobuild was chosen for this verification, as it is somewhat different from other English dictionaries. It uses full sentences, not phrases, and focuses on providing frequently used examples taken from a corpus (a collection of British and American newspapers, books, TV programs, real-life conversations, etc). In other words, Cobuild entries explain their usage in discourse, unlike the traditional dictionaries that focuses on precise definitions of words. In addition, Cobuild has been used in various NLP work (e.g. Hoelter 1999).

[3] The figure of 100,000 was heuristically chosen as the cut-off point.

transferable to other languages. The referents under v are not verified in English as language independent referents, requiring another mode of verification. Many of the referents under vi should have been classified as language dependent in the first stage of verification (see the next subsection).

3.4 Results and discussion

The results were significant, as shown in Table 4. Types i - iv (i.e. ∑ i-iv for all POSs) amount to 90.7%, which are confirmed to be language independent in the second verification.

	i	ii	iii	iv	∑ i-iv	v	vi
Adj	44.4	7.1	12.2	24.4	87.8%	7.8	4.4
Verb	44.2	3.3	20.8	16.7	85.0%	5.8	9.2
VN	58.1	3.2	21.8	13.7	96.8%	2.4	.8
All POS	50.1	4.0	19.0	17.6	***90.7%***	4.2	5.1

Table 4: Types of Japanese representative arguments compared with English per POS

We find that when there is a good match of English translation that is expressed by one or two words of English and semantically maps well, Lexeed's referent is generally found in the Cobuild or confirmed in Google search.

This gives rise to two implications.

One is that the exact semantic transfer between languages can be difficult for some words, and those words that don't transfer cleanly in translation tell us much about the culture of the speakers and are, thus, language specific (Bond 2005).

On this note, iv includes the cases where there is a translation but it takes, say, more than three words to *explain* the Japanese definition word; e.g. 論外 *rongai* is translated as 'be out of question'. Practically speaking, one word has to be chosen to be able to look it up in the Cobuild dictionary, and even then the arguments listed there are often too general or unrelated. In other words, referents can be automatically determined as language independent referents discerning from specific, when the appropriate translation of words are found and they comprise no more than three.

The other implication is that finding the optimal match of translation requires high command of the two languages: English and Japanese in this case. The inadequacy of this increased the number for vi ' no match', which could have been eliminated on the first verification. Nonetheless, the results from the two methods of verification showed that approximately 80% of arguments can be expected to be language independent.[4]

We can, thus, conclude that we can expect the majority of the representative arguments extracted from Lexeed to be language independent. Although this experiment is based on only two languages, considering the fact that Japanese and English are linguistically and culturally quite distant, the results are intriguing and promising. If arguments are manually confirmed as being language independent by the two very distinct languages, it is likely that those referents are shared across languages, although further verifications using a third language will be more assuring.

Interesting to note that some referents showed a case of 'partial mismatch of referents'. For example. the word *iroppoi* refers to women in Japanese, while the English equivalent 'sexy' is used for both sexes. The reverse is also true: 'pretty' in English is generally referred to girls, while in Japanese both sexes.

4 Related work

The series of our work is summarised as involving:

- the extraction of referential knowledge in the form of <u>representative arguments</u>,
- from <u>Japanese</u> dictionary definition sentences,
- using <u>machine-readable dictionaries</u>,
- investigating the feasibility for extending its referential knowledge <u>across languages</u>.

Since the inception of electronic lexical databases, such as WordNet (for English), *Goi-Taikei* (for Japanese), and HowNet (for Chinese), the use of machine-readable dictionaries for acquiring ontology has been the method taken by many in various languages (e.g. Tsurumaru 1991, Wilks et al. 1996, Nichols et al. 2005, inter alia). The majority of work, however, has concentrated on extracting semantic relations of words, such as synonym, hypernym, and meronym (Wilks et al. 1996, Fellbaum 1998).

In terms of work that focuses on extracting referential information, many studies use newspaper corpora instead of dictionaries. The two notable works in Japanese are the new EDR *Verb valency dictionary* (Hagino et al. 2003, listing verbs only) and a case frame dictionary (Kawahara and Kurohashi 2004). Utsuro et al. (1992) use bilingual corpora to acquire lexical knowledge.

Similar work for English has also been reported (Resnik 1997, inter alia). Slightly different is the work by Agirre and Martinez (2002) that focuses on class-to-

[4] We manually examined 3,678 representative arguments, 421 of these are found to be language specific by the first verification and another 342 from the second verification, that amount to 20.7% total (421+342/3,678).

class (class of verbs – class of nominals) relations instead of usual word-to-class (verb – a nominal class) relations.

The notable works that aim to process knowledge are CYC (Lenat 1995), Harabagiu and Moldovan (1998), and MindNet (Richardson et al. 1998). Although all of them are designed for English, we can improve our work from their approaches, which is our future work. Different in its approach but closer to the interest of our work is the work by Elouazizi (2004). It tries to formalise a universal ontology of referring modes to capture an optimal referential relations from the perspective of cognitive semantics.

What is different about our research is that while others extract general semantic classes of referents (e.g. 'person'), we extract specific referents that are representative for the predicate (e.g. 'police').

Our approach, however, has one disadvantage in terms of coverage. It cannot hold for all referents, since not every definition word has 'representative' arguments or dictionaries ensure to list them. As no single method is perfect per se, it is deemed beneficial that we consider merging the positives from various methods to further improve.

5 Conclusions

The output of this paper is that following the work on extracting the referential knowledge in the form of representative arguments from a machine readable Japanese dictionary (Nariyama et al. 2005), we examined the feasibility for extending its application across languages. The initial results show substantial promise.

Accounting for contextual information and world knowledge seems a prohibitive task at present. This paper has made a first step forward towards dealing with these issues by proposing a method to create a cross-lingual referential knowledge database. This linkage is of significant importance for multi-lingual applications, such as machine translation systems.

As another future work, we plan to formulate additional inferences drawing from representative arguments. For example, 'Mary gave birth to a baby' entails that Mary is the *mother* of the *baby*, and this knowledge is cross-linguistically true. This knowledge is particularly of importance for Question and answering tasks. It enables to find the answer for questions, such as 'Who is the mother of the baby?'

References

E Agirre and D. Martinez. 2002. Integrating selectional preferences in WordNet. The first *International WordNet Conference*.

F. Bond. 2005. *Translating the untranslatable: A solution to the problem of generating English determiners.* CSLI Publications. California

F. Bond et al. 2004. The Hinoki Treebank: A Treebank for Text Understanding, In Proc. of the First *IJCNLP*, *Lecture Notes in Computer Science.* Springer Verlag

U. Callmeier. 2002. Preprocessing and encoding techniques in PET. In Oepen et al. (eds), *Collaborative Language Engineering*, 127–143. CSLI, Stanford

N. Elouazizi. 2004. Towards an ontology of referential relations in natural language and its implication for NLP. Workshop on the Potential of Cognitive semantics for ontology (FOIS 2004)

C. Fellbaum (ed.). 1998. *WordNet: An electronic lexical database*. The MIT Press, Cambridge

A. Frank. 2004. Constraint-based RMRS construction from shallow grammars. In Proc. of *COLING*. 1269–1272, Geneva

T. Hagino et al. 2003. *Japanese verb valency dictionary.* Sanseido Publishing

S. M. Harabagiu and D. Moldovan. 1998. Knowledge processing on an extended WordNet. In Fellbaum (ed). 379-405

M. Hoelter. 1999. *Lexical-semantic information in Head-Driven Phrase Structure grammar and natural language processing.* Lincom Theoretical Linguistics

S. Ikehara et al. (eds). 1997. *Japanese Lexicon.* Iwanami

H. Isozaki and T. Hirano. 2003. Japanese zero pronoun resolution based on ranking rules and machine learning. In Proc. of *EMNLP*.184-191

D. Kawahara and S. Kurohashi. 2004. Improving Japanese zero pronoun resolution by global word sense disambiguation. In Proc. of *COLING*. 343-349. Geneva

D. B. Lenat. 1995. CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11), 32-38

S. Nariyama. 2003. *Ellipsis and Reference-tracking in Japanese*. SLCS 66, Amsterdam: John Benjamins

S. Nariyama et al. 2005. Extracting Representative Arguments from Dictionaries for Resolving Zero Pronouns. MT Summit X. Phuket Thailand

E. Nichols, F. Bond, and D. Flickenger. 2005. Robust Ontology Acquisition from Machine-Readable Dictionaries. In Proc. of *the International Joint Conferences on Artificial Intelligence.* Edinburgh

P. Resnik. 1997. Selectional preference and sense disambiguation. In proceedings of the ANLP Workshop *Tagging text with lexical semantics: Why What and How?*

S. Richardson et al. 1998. MindNet: acquiring and structuring semantic information from text. In Proc. of *COLING*

M. Siegel and E M. Bender. 2002. Efficient deep processing of Japanese. In Proc. of the 3rd Workshop on Asian Language Resources and International Standardization at *COLING*, Taipei

H. Tsurumaru et al. 1991. An approach to thesaurus construction from Japanese language dictionary. In IPSJ SIG Notes Natural Language, vol.83-16. 121–128 (in Japanese)

T.Utsuro, Y.Matsumoto, M.Nagao. 1992. Lexical knowledge acquisition from bilingual corpora. COLING. 581-587

Y.A. Wilkes et al. 1996. *Electric Words*. MIT Press