Japanese Idiom Recognition:
Drawing a Line between Literal and Idiomatic Meanings

Chikara Hashimoto* **Satoshi Sato**† **Takehito Utsuro**‡

* Graduate School of Informatics
Kyoto University
Kyoto, 606-8501, Japan

† Graduate School of Engineering
Nagoya University
Nagoya, 464-8603, Japan

‡ Graduate School of Systems and Information Engineering
University of Tsukuba
Tsukuba, 305-8573, Japan

Abstract

Recognizing idioms in a sentence is important to sentence understanding. This paper discusses the lexical knowledge of idioms for idiom recognition. The challenges are that idioms can be ambiguous between literal and idiomatic meanings, and that they can be "transformed" when expressed in a sentence. However, there has been little research on Japanese idiom recognition with its ambiguity and transformations taken into account. We propose a set of lexical knowledge for idiom recognition. We evaluated the knowledge by measuring the performance of an idiom recognizer that exploits the knowledge. As a result, more than 90% of the idioms in a corpus are recognized with 90% accuracy.

1 Introduction

Recognizing idioms in a sentence is important to sentence understanding. Failure of recognizing idioms leads to, for example, mistranslation.

In the case of the translation service of Excite[1], it sometimes mistranslates sentences that contain idioms such as (1a), due to the recognition failure.

(1) a. Kare-wa mondai-no kaiketu-ni
he-TOP problem-GEN solving-DAT
hone-o *o*-tta.
bone-ACC *break*-PAST
"He *made an effort* to solve the problem."

b. "He broke his bone to the resolution of a question."

(1a) contains an idiom, *hone-o oru* (bone-ACC break) "make an effort." (1b) is the mistranslation of (1a), in which the idiom is interpreted literally.

In this paper, we discuss lexical knowledge for idiom recognition. The lexical knowledge is implemented in an idiom dictionary that is used by an idiom recognizer we implemented. Note that the idiom recognition we define includes distinguishing literal and idiomatic meanings.[2] Though there has been a growing interest in MWEs (Sag et al., 2002), few proposals on idiom recognition take into account ambiguity and transformations. Note also that we tentatively define an idiom as a phrase that is semantically non-compositional. A precise characterization of the notion "idiom" is beyond the scope of the paper.[3]

Section 2 defines what makes idiom recognition difficult. Section 3 discusses the classification of Japanese idioms, the requisite lexical knowledge, and implementation of an idiom recognizer. Section 4 evaluates the recognizer that exploits the knowledge. After the overview of related works in Section 5, we conclude the paper in Section 6.

2 Two Challenges of Idiom Recognition

Two factors make idiom recognition difficult: **ambiguity** between literal and idiomatic meanings and "**transformations**" that idioms could undergo.[4] In fact, the mistranslation in (1) is caused by the inability of disambiguation between the two meanings. "Transformation" also causes mistrans-

[1] http://www.excite.co.jp/world/

[2] Some idioms represent two or three idiomatic meanings. But those meanings in an idiom are not distinguished. We concerned only whether a phrase is used as an idiom or not.

[3] For a detailed discussion of what constitutes the notion of (Japanese) idiom, see Miyaji (1982), which details usages of commonly used Japanese idioms.

[4] The term "transformation" in the paper is not relevant to the Chomskyan term in Generative Grammar.

Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pages 353–360,
Sydney, July 2006. ©2006 Association for Computational Linguistics

lation. Sentences in (2) and (3a) contain an idiom, *yaku-ni tatu* (part-DAT stand) "serve the purpose."

(2) Kare-wa *yaku-ni tatu*.
he-TOP *part*-DAT *stand*
"He *serves the purpose*."

(3) a. Kare-wa *yaku-ni* sugoku *tatu*.
he-TOP *part*-DAT very *stand*
"He really *serves the purpose*."

b. "He stands enormously in part."

Google's translation system[5] mistranslates (3a) as in (3b), which does not make sense,[6] though it successfully translates (2). The only difference between (2) and (3a) is that bunsetu[7] constituents of the idiom are detached from each other.

3 Knowledge for Idiom Recognition

3.1 Classification of Japanese Idioms

Requisite lexical knowledge to recognize an idiom depends on how difficult it is to recognize it. Thus, we first classify idioms based on **recognition difficulty**. The recognition difficulty is determined by the two factors: ambiguity and transformability.

Consequently, we identify three classes (Figure 1).[8] **Class A** is not transformable nor ambiguous. **Class B** is transformable but not ambiguous.[9] **Class C** is transformable and ambiguous. Class A amounts to unambiguous single words, which are easy to recognize, while Class C is the most difficult to recognize. Only Class C needs further classifications, since only Class C needs disambiguation and lexical knowledge for disambiguation depends on its **part-of-speech** (POS) and **internal structure**. The POS of Class C is either verbal or adjectival, as in Figure 1. Internal structure represents constituent words' POS and a dependency between bunsetus. The internal structure of *hone-o oru* (bone-ACC bone), for instance, is "(Noun/Particle Verb)," abbreviated as "(N/P V)."

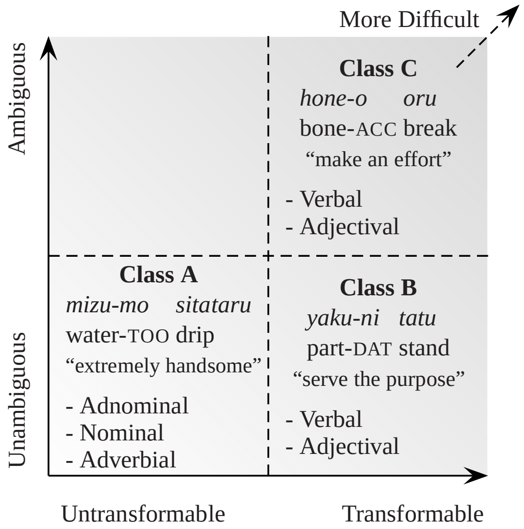

Figure 1: Idiom Classification based on the Recognition Difficulty

Then, let us give a full account of the further classification of Class C. We exploit grammatical differences between literal and idiomatic usages for disambiguation. We will call the knowledge of the differences the **disambiguation knowledge**. For instance, a phrase, *hone-o oru*, does not allow passivization when used as an idiom, though it does when used literally. Thus, (4), in which the phrase is passivized, cannot be an idiom.

(4) *hone*-ga *o*-rareru
bone-NOM *break*-PASS
"A bone is broken."

In this case, passivizability can be used as a disambiguation knowledge. Also, detachability of the two bunsetu constituents can serve for disambiguating the idiom; they cannot be separated. In general, usages applicable to idioms are also applicable to literal phrases, but the reverse is not always true (Figure 2). Then, finding the disambiguation knowledge amounts to finding usages applicable to only literal phrases.

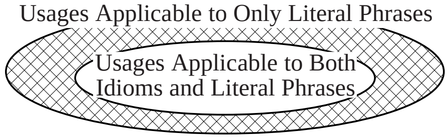
Usages Applicable to Only Literal Phrases
Usages Applicable to Both
Idioms and Literal Phrases

Figure 2: Difference of Applicable Usages

Naturally, the disambiguation knowledge for an idiom depends on its POS and internal structure.

[5] http://www.google.co.jp/language_tools

[6] In fact, the idiom has no literal interpretation.

[7] A bunsetu is a syntactic unit in Japanese, consisting of one independent word and more than zero ancillary words. The sentence in (3a) consists of four bunsetu constituents.

[8] The blank space at the upper left in the figure implies that there is no idiom that does not undergo any transformation and yet is ambiguous. Actually, we have not come up with such an example that should fill in the blank space.

[9] Anonymous reviewers pointed out that Class A and B could also be ambiguous. In fact, one can devise a context that makes the literal interpretation of those Classes possible. However, virtually no phrase of Class A or B is interpreted literally in real texts, and we think our generalization safely captures the reality of idioms.

As for **POS**, disambiguation of verbal idioms can be performed by the knowledge of passivizability, while that of adjectival idioms cannot. Regarding **internal structure**, detachability should be annotated on every boundary of bunsetus. Thus, the number of annotations of detachability depends on the number of bunsetus of an idiom.

There is no need for further classification of Class A and B, since lexical knowledge for them is invariable. The next section mentions their invariableness. After all, Japanese idioms are classified as in Figure 3. The whole picture of the subclasses of Class C remains to be seen.

3.2 Knowledge for Each Class

What lexical knowledge is needed for each class?

Class A needs only a string information; idioms of the class amount to unambiguous single words.

A string information is undoubtedly invariable across all kinds of POS and internal structure.

Class B requires not only a string but also knowledge that normalizes transformations idioms could undergo, such as passivization and detachment of bunsetus. We identify three types of transformations that are relevant to idioms: **1)** Detachment of Bunsetu Constituents, **2)** Predicate's Change, and **3)** Particle's Change. Predicate's change includes inflection, attachment of a negative morpheme, a passive morpheme or modal verbs, and so on. Particle's change represents attachment of topic or restrictive particles. (5b) is an example of predicate's change from (5a) by adding a negative morpheme to a verb. (5c) is an example of particle's change from (5a) by adding a topic particle to the preexsistent particle of an idiom.

(5) a. Kare-wa *yaku-ni* *tatu*.
he-TOP *part*-DAT *stand*
"He *serves the purpose*."

b. Kare-wa *yaku-ni* *tat*-**anai**.
he-TOP *part*-DAT *stand*-**NEG**
"He does not *serve the purpose*."

c. Kare-wa *yaku-ni*-**wa** *tatu*.
he-TOP *part*-DAT-**TOP** *stand*
"He *serves the purpose*."

To normalize the transformations, we utilize a dependency relation between constituent words, and we call it the **dependency knowledge**. This amounts to checking the presence of all the constituent words of an idiom. Note that we ignore, among constituent words, endings of a predicate and case particles, *ga* (NOM) and *o* (ACC), since they could change their forms or disappear.

The dependency knowledge is also invariable across all kinds of POS and internal structure.

Class C requires the disambiguation knowledge, as well as all the knowledge for Class B.

As a result, all the requisite knowledge for idiom recognition is summarized as in Table 1.

	String	Dependency	Disambiguation
Class A	✔		
Class B	✔	✔	
Class C	✔	✔	✔

Table 1: Requisite Knowledge for each Class

As discussed in §3.1, the disambiguation knowledge for an idiom depends on which subclass it belongs to. A comprehensive idiom recognizer calls for all the disambiguation knowledge for all the subclasses, but we have not figured out all of them. Then, we decided to blaze a trail to discover the disambiguation knowledge by investigating the most commonly used idioms.

3.3 Disambiguation Knowledge for the Verbal (N/P V) Idioms

What type of idiom is used most commonly? The answer is the **verbal (N/P V)** type like *hone-o oru* (bone-ACC break); it is the most abundant in terms of both type and token. Actually, 1,834 out of 4,581 idioms (≃40%) in Kindaichi and Ikeda (1989), which is a Japanese dictionary with more than 100,000 words, are this type.[10] Also, 167,268 out of 220,684 idiom tokens in Mainichi newspaper of 10 years ('91–'00) (≃76%) are this type.[11]

Then we discuss what can be used to disambiguate the verbal (N/P V) type. First, we examined literature of linguistics (Miyaji, 1982; Morita, 1985; Ishida, 2000) that observed characteristics of Japanese idioms. Then, among the characteristics, we picked those that could help with the disambiguation of the type. (6) summarizes them.

[10] Counting was performed automatically by means of the morphological analyzer ChaSen (Matsumoto et al., 2000) with no human intervention. Note that Kindaichi and Ikeda (1989) consists of 4,802 idioms, but 221 of them were ignored since they contained unknown words for ChaSen.

[11] We counted idiom tokens by string matching with inflection taken into account. And we referred to Kindaichi and Ikeda (1989) for a comprehensive idiom list. Note that counting was performed totally automatically.

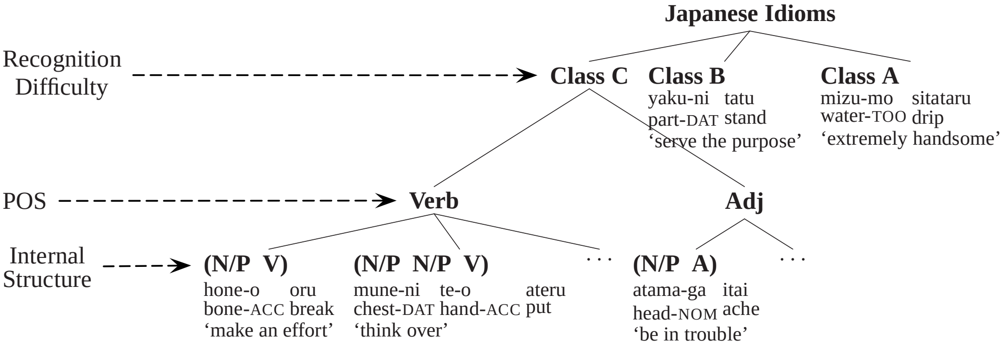

Figure 3: Classification of Japanese Idioms for the Recognition Task

(6) **Disambiguation Knowledge for the Verbal (N/P V) Idioms**

a. Adnominal Modification Constraints

I. Relative Clause Prohibition
II. Genitive Phrase Prohibition
III. Adnominal Word Prohibition

b. Topic/Restrictive Particle Constraints

c. Voice Constraints

I. Passivization Prohibition
II. Causativization Prohibition

d. Modality Constraints

I. Negation Prohibition
II. Volitional Modality Prohibition[12]

e. Detachment Constraint

f. Selectional Restriction

For example, the idiom, *hone-o oru*, does not allow adnominal modification by a genitive phrase. Thus, (7) can be interpreted only literally.

(7) **kare-no** *hone-o* *oru*
he-GEN *bone*-ACC *break*
"(Someone) breaks **his** bone."

That is, the Genitive Phrase Prohibition, (6aII), is in effect for the idiom. Likewise, the idiom does not allow its case particle *o* (ACC) to be substituted with restrictive particles such as *dake* (only). Thus, (8) represents only a literal meaning.

(8) *hone*-**dake** *oru*
bone-**ONLY** *break*
"(Someone) breaks **only** some bones."

This means the Restrictive Particle Constraint, (6b), is also in effect. Also, (4) shows that the Passivization Prohibition, (6cI), is in effect, too.

Note that the constraints in (6) are not always in effect for an idiom. For instance, the Causativization Prohibition, (6cII), is invalid for the idiom, *hone-o oru*. In fact, (9a) can be interpreted both literally and idiomatically.

(9) a. kare-ni *hone-o* *or*-**aseru**
he-DAT *bone*-ACC *break*-**CAUS**

b. "(Someone) makes him break a bone."

c. "(Someone) makes him *make an effort*."

3.4 Implementation

We implemented an idiom dictionary based on the outcome above and a recognizer that exploits the dictionary. This section illustrates how they work, and we focus on Class B and C hereafter.

The idiom recognizer looks up **dependency patterns** in the dictionary that match a part of the dependency structure of a sentence (Figure 4). A dependency pattern is equipped with all the requisite knowledge for idiom recognition. Rough sketch of the recognition algorithm is as follows:

1. Analyze the morphology and dependency structures of an input sentence.
2. Look up dependency patterns in the dictionary that match a part of the dependency structure of the input sentence.
3. Mark constituents of an idiom in the sentence if any.[13] Constituents that are marked are constituent words and bunsetu constituents that include one of those constituent words.

[12] "Volitional Modality" represents those verbal expressions of order, request, permission, prohibition, and volition.

[13] As a constituent marker, we use an ID that is assigned to each idiom in the dictionary.

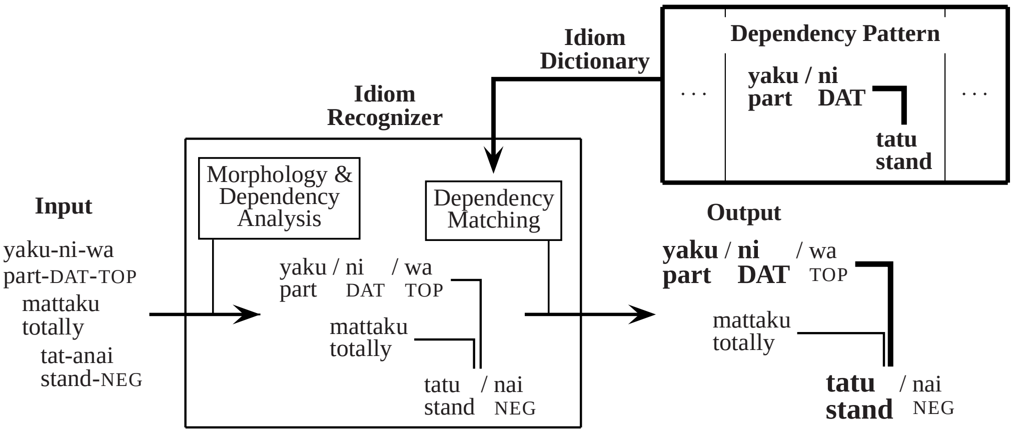

Figure 4: Internal Working of the Idiom Recognizer

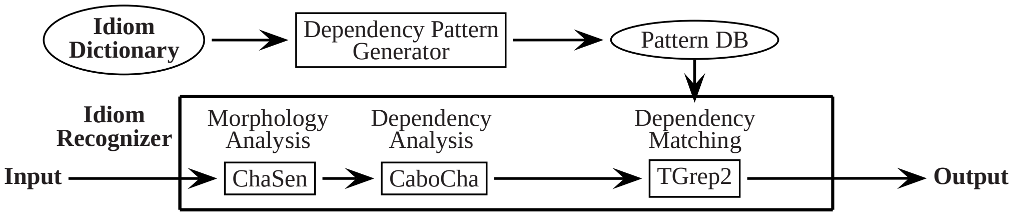

Figure 5: Organization of the System

As in Figure 5, we use ChaSen as a morphology analyzer and CaboCha (Kudo and Matsumoto, 2002) as a dependency analyzer. Dependency matching is performed by TGrep2 (Rohde, 2005), which finds syntactic patterns in a sentence or treebank. The dependency pattern is usually getting complicated since it is tailored to the specification of TGrep2. Thus, we developed the Dependency Pattern Generator that compiles the pattern database from a human-readable idiom dictionary.

Only the difference in treatments of Class B and C lies in their dependency patterns. The dependency pattern of Class B consists of only its dependency knowledge, while that of Class C consists of not only its dependency knowledge but also its disambiguation knowledge (Figure 6).

The idiom dictionary consists of 100 idioms, which are all verbal (N/P V) and belong to either Class B or C. Among the knowledge in (6), the Selectional Restriction has not been implemented yet. The 100 idioms are those that are used most frequently. To be precise, 50 idioms in Kindaichi and Ikeda (1989) and 50 in Miyaji (1982) were extracted by the following steps:[14]

1. From Miyaji (1982), 50 idioms that were used most frequently in Mainichi newspaper of 10 years ('91–'00) were extracted.
2. From Kindaichi and Ikeda (1989), 50 idioms that were used most frequently in the newspaper of 10 years but were not included in the 50 idioms from Miyaji (1982) were extracted.

As a result, 66 out of the 100 idioms were Class B, and the other 34 idioms were Class C.[15]

4 Evaluation

4.1 Experiment Condition

We conducted an experiment to see the effectiveness of the lexical knowledge we proposed.

As an **evaluation corpus**, we collected 300 example sentences of the 100 idioms from Mainichi newspaper of '95: three sentences for each idiom. Then we added another nine sentences for three idioms that are orthographic variants of one of the 100 idioms. Among the three idioms, one belonged to Class B and the other two belonged to Class C. Thus, 67 out of the 103 idioms were Class B and the other 36 were Class C. After all, 309

[14] We counted idiom tokens by string matching with inflection taken into account. Note that counting was performed automatically without human intervention.

[15] We found that the most frequently used 100 idioms in Kindaichi and Ikeda (1989) cover as many as 53.49% of all tokens in Mainichi newspaper of 10 years. This implies that our dictionary accounts for approximately half of all idiom tokens in a corpus.

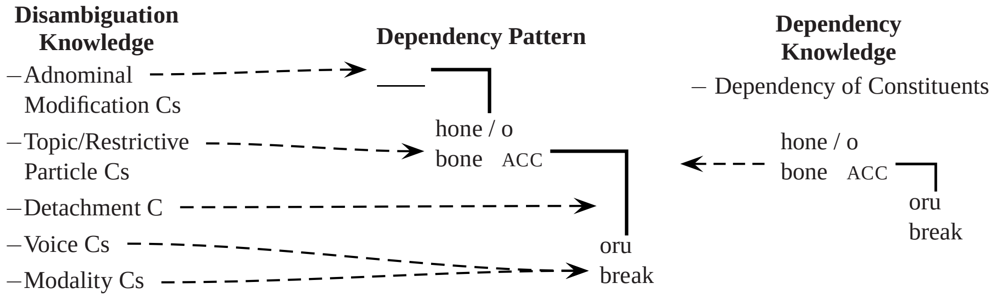

Figure 6: Dependency Pattern of Class C

sentences were prepared. Table 2 shows the breakdown of them. "Positive" indicates sentences including a true idiom, while "Negative" indicates those including a literal-usage "idiom."

	Class B	Class C	Total
Positive	200	66	266
Negative	1	42	43
Total	201	108	309

Table 2: Breakdown of the Evaluation Corpus

A **baseline system** was prepared to see the effect of the disambiguation knowledge. The baseline system was the same as the recognizer except that it exploited no disambiguation knowledge.

4.2 Result

The result is shown in Table 3. The left side shows the performances of the recognizer, while the right side shows that of the baseline. Differences of performances between the two systems are marked with **bold**. Recall, Precision, and F-Measure, are calculated using the following equations.

$$Recall = \frac{|Correct\ Outputs|}{|Positive|}$$

$$Precision = \frac{|Correct\ Outputs|}{|All\ Outputs|}$$

$$F\text{-}Measure = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

As a result, more than 90% of the idioms can be recognized with 90% accuracy. Note that the recognizer made fewer errors due to the employment of the disambiguation knowledge.

The result shows the high performances. However, there turns out to be a long way to go to solve the most difficult problem of idiom recognition: drawing a line between literal and idiomatic meanings. In fact, the precision of recognizing idioms of Class C remains less than 70% as in Table 3. Besides, the recognizer successfully rejected only 15 out of 42 negative sentences. That is, its success rate of rejecting negative ones is only 35.71%

4.3 Discussion of the Disambiguation Knowledge

First of all, positive sentences, i.e., sentences containing true idioms, are in the blank region of Figure 2, while negative ones, i.e., those containing literal phrases, are in both regions. Accordingly, the disambiguation amounts to **i)** rejecting negative ones in the shaded region, **ii)** rejecting negative ones in the blank region, or **iii)** accepting positive ones in the blank region. i) is relatively easy since there are visible evidences in a sentence that tell us that it is NOT an idiom. However, ii) and iii) are difficult due to the absence of visible evidences. Our method is intended to perform i), and thus has an obvious limitation.

Next, we look cloosely at cases of success or failure of rejecting negative sentences. There were 15 cases where rejection succeeded, which correspond to i). The disambiguation knowledge that contributed to rejection and the number of sentences it rejects are as follows.[16]

1. Genitive Phrase Prohibition (6aII) 6
2. Relative Clause Prohibition (6aI) 5
3. Detachment Constraint (6e) 2
4. Negation Prohibition (6dI) 1

This shows that the Adnominal Modification Constraints, 1. and 2. above, are the most effective.

There were 27 cases where rejection failed. These are classified into two types:

[16] There was one case where rejection succeeded due to the dependency analysis error.

	Class B	Class C	All	Class B	Class C	All
Recall	0.975 ($\frac{195}{200}$)	0.939 ($\frac{62}{66}$)	0.966 ($\frac{257}{266}$)	0.975 ($\frac{195}{200}$)	0.939 ($\frac{62}{66}$)	0.966 ($\frac{257}{266}$)
Precision	1.000 ($\frac{195}{195}$)	**0.697** ($\frac{62}{89}$)	**0.905** ($\frac{257}{284}$)	1.000 ($\frac{195}{195}$)	0.602 ($\frac{62}{103}$)	0.862 ($\frac{257}{298}$)
F-Measure	0.987	**0.800**	**0.935**	0.987	0.734	0.911

Table 3: Performances of the Recognizer (left side) and the Baseline System (right side)

1. Those that could have been rejected by the Selectional Restriction (6f) 5
2. Those that might be beyond the current technology 22

1. and 2. correspond to i) and ii), respectively. We see that the Selectional Restriction would have been as effective as the Adnominal Modification Constraints. A part of a sentence that the knowledge could have rejected is below.

(10) basu-ga *tyuu-ni* *ui*-ta
bus-NOM *midair*-DAT *float*-PAST
"The bus floated in midair."

An idiom, *tyuu-ni uku* (midair-DAT float) "remain to be decided," takes as its argument something that can be decided, i.e., ⟨`1000:abstract`⟩ rather than ⟨`2:concrete`⟩ in the sense of the *Goi-Taikei* ontology (Ikehara et al., 1997). Thus, (10) has no idiomatic sense.

A simplified example of 2. is illustrated in (11).

(11) ase-o nagasi-te huku-o
sweat-ACC shed-and clothes-ACC
kiru-yorimo, hadaka-ga gouriteki-da
wear-rather.than, nudity-NOM rational-DECL
"It makes more sense to be naked than wearing clothes in a sweat."

The phrase *ase-o nagasu* (sweat-ACC shed) could have been an idiom meaning "work hard." It is contextual knowledge that prevented it from being the idiom. Clearly, our technique is unable to handle such a case, which belongs to ii), since no visible evidence is available. Dealing with that might require some sort of machine learning technique that exploits contextual information. Exploring that possibility is one of our future works.

Finally, the 42 negative sentences consist of 15 sentences, which we could disambiguate, 5 sentences, which Selectional Restriction could have disambiguated, and 22, which belong to ii) and are beyond the current technique. Thus, the real challenge lies in 7% ($\frac{22}{309}$) of all idiom occurrences.

4.4 Discussion of the Dependency Knowledge

The dependency knowledge failed in only five cases. Three of them were due to the defect of dealing with case particles' change like omission. The other two cases were due to the noun constituent's incorporation into a compound noun. (12) is a part of such a case.

(12) kaihuku-*kidou-ni* *nori*-hajimeru
recovery-*orbit*-DAT *ride*-begin
"(Economics) *get* back *on* a recovery *track*."

The idiom, *kidou-ni noru* (orbit-DAT ride) "get on track," has a constituent, *kidou*, which is incorporated into a compound noun *kaihuku-kidou* "recovery track." This is unexpected and cannot be handled by the current machinery.

5 Related Work

There has been a growing awareness of Japanese MWE problems (Baldwin and Bond, 2002). However, few attempts have been made to recognize idioms in a sentence with their ambiguity and transformations taken into account. In fact, most of them only create catalogs of Japanese idiom: collecting idioms as many as possible and classifying them based on some general linguistic properties (Tanaka, 1997; Shudo et al., 2004).

A notable exception is Oku (1990); his idiom recognizer takes the ambiguity and transformations into account. However, he only uses the Genitive Phrase Prohibition, the Detachment Constraint, and the Selectional Restriction, which would be too few to disambiguate idioms.[17] As well, his classification does not take the recognition difficulty into account. This makes his idiom dictionary get bloated, since disambiguation knowledge is given to unambiguous idioms, too.

Uchiyama et al. (2005) deals with disambiguating some Japanese verbal compounds. Though verbal compounds are not counted as idioms, their study is in line with this study.

[17] We cannot compare his recognizer with ours numerically since no disambiguation success rate is presented in Oku (1990); only the overall performance is presented.

Our classification of idioms correlates loosely with that of MWEs by Sag et al. (2002). Japanese idioms that we define correspond to *lexicalized phrases*. Among lexicalized phrases, *fixed expressions* are equal to Class A. Class B and C roughly correspond to *semi-fixed* or *syntactically-flexible expressions*. Note that, though the three subtypes of lexicalized phrases are distinguished based on what we call **transformability**, no distinction is made based on the **ambiguity**.[18]

6 Conclusion

Aiming at Japanese idiom recognition with ambiguity and transformations taken into accout, we proposed a set of lexical knowledge for idioms and implemented a recognizer that exploits the knowledge. We maintain that requisite knowledge depends on its transformability and ambiguity; transformable idioms require the dependency knowledge, while ambiguous ones require the disambiguation knowledge as well as the dependency knowledge. As the disambiguation knowledge, we proposed a set of constraints applicable to a phrase when it is used as an idiom. The experiment showed that more than 90% idioms could be recognized with 90% accuracy but the success rate of rejecting negative sentences remained 35.71%. The experiment also revealed that, among the disambiguation knowledge, the Adnominal Modification Constraints and the Selectional Restriction are the most effective.

What remains to be done is two things; one is to reveal all the subclasses of Class C and all the disambiguation knowledge, and the other is to apply a machine learning technique to disambiguating those cases that the current technique is unable to handle, i.e., cases without visible evidence.

In conclusion, there is still a long way to go to draw a perfect line between literal and idiomatic meanings, but we believe we broke new ground in Japanese idiom recognition.

Acknowledgment A special thank goes to Gakushu Kenkyu-sha, who permitted us to use Gakken's Dictionary for our research.

[18] The notion of *decomposability* of Sag et al. (2002) and Nunberg et al. (1994) is independent of **ambiguity**. In fact, ambiguous idioms are either decomposable (*hara-ga kuroi* (belly-NOM black) "black-hearted") or non-decomposable (*hiza-o utu* (knee-ACC hit) "have a brainwave"). Also, unambiguous idioms are either decomposable (*hara-o yomu* (belly-ACC read) "fathom someone's thinking") or non-decomposable (*saba-o yomu* (chub.mackerel-ACC read) "cheat in counting").

References

Timothy Baldwin and Francis Bond. 2002. Multiword Expressions: Some Problems for Japanese NLP. In *Proceedings of the 8th Annual Meeting of the Association of Natural Language Processing, Japan*, pages 379–382, Keihanna, Japan.

Satoru Ikehara, Masahiro Miyazaki, Satoshi Shirai, Akio Yokoo, Hiromi Nakaiwa, Kentaro Ogura, Yoshifumi Ooyama, and Yoshihiko Hayashi. 1997. *Goi-Taikei — A Japanese Lexicon*. Iwanami Shoten.

Priscilla Ishida. 2000. Doushi Kanyouku-ni taisuru Tougoteki Sousa-no Kaisou Kankei (On the Hierarchy of Syntactic Operations Applicable to Verb Idioms). *Nihongo Kagaku (Japanese Linguistics)*, 7:24–43, April.

Haruhiko Kindaichi and Yasaburo Ikeda, editors. 1989. *Gakken Kokugo Daijiten (Gakken's Dictionary)*. Gakushu Kenkyu-sha.

Taku Kudo and Yuji Matsumoto. 2002. Japanese Dependency Analyisis using Cascaded Chunking. In *Proceedings of the 6th Conference on Natural Language Learning (CoNLL-2002)*, pages 63–69.

Yuji Matsumoto, Akira Kitauchi, Tatsuo Yamashita, Yoshitaka Hirano, Hiroshi Matsuda, Kazuma Takaoka, and Masayuki Asahara, 2000. *Morphological Analysis System ChaSen version 2.2.1 Manual*. Nara Institute of Science and Technology, Dec.

Yutaka Miyaji. 1982. *Kanyouku-no Imi-to Youhou (Usage and Semantics of Idioms)*. Meiji Shoin.

Yoshiyuki Morita. 1985. Doushikanyouku (Verb Idioms). *Nihongogaku (Japanese Linguistics)*, 4(1):37–44.

Geoffrey Nunberg, Ivan A. Sag, and Thomas Wasow. 1994. Idioms. *Language*, 70:491–538.

Masahiro Oku. 1990. Nihongo-bun Kaiseki-ni-okeru Jutsugo Soutou-no Kanyouteki Hyougen-no Atsukai (Treatments of Predicative Idiomatic Expressions in Parsing Japanese). *Journal of Information Processing Society of Japan*, 31(12):1727–1734.

Douglas L. T. Rohde, 2005. *TGrep2 User Manual version 1.15*. Massachusetts Institute of Technology. http://tedlab.mit.edu/˜dr/Tgrep2/.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Computational Linguistics and Intelligent Text Processing: Third International Conference*, pages 1–15.

Kosho Shudo, Toshifumi Tanabe, Masahito Takahashi, and Kenji Yoshimura. 2004. MWEs as Non-propositional Content Indicators. In *the 2nd ACL Workshop on Multiword Expressions: Integrating Processing*, pages 32–39.

Yasuhito Tanaka. 1997. Collecting idioms and their equivalents. In *IPSJ SIGNL 1997-NL-121*.

Kiyoko Uchiyama, Timothy Baldwin, and Shun Ishizaki. 2005. Disambiguating Japanese Compound Verbs. *Computer Speech and Language, Special Issue on Multiword Expressions*, 19, Issue 4:497–512.