Utilizing Semantic Equivalence Classes of Japanese Functional Expressions in Translation Rule Acquisition from Parallel Patent Sentences

Taiji Nagasaka† **Ran Shimanouchi**† **Akiko Sakamoto**† **Takafumi Suzuki**†
Yohei Morishita† **Takehito Utsuro**† **Suguru Matsuyoshi**‡

†Graduate School of Systems and Information Engineering, University of Tsukuba, Tsukuba, 305-8573, JAPAN
‡Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma, Nara, 630-0192, JAPAN

Abstract

In the "Sandglass" MT architecture, we identify the class of monosemous Japanese functional expressions and utilize it in the task of translating Japanese functional expressions into English. We employ the semantic equivalence classes of a recently compiled large scale hierarchical lexicon of Japanese functional expressions. We then study whether functional expressions within a class can be translated into a single canonical English expression. Based on the results of identifying monosemous semantic equivalence classes, this paper studies how to extract rules for translating functional expressions in Japanese patent documents into English. In this study, we use about 1.8M Japanese-English parallel sentences automatically extracted from Japanese-English patent families, which are distributed through the Patent Translation Task at the NTCIR-7 Workshop. Then, as a toolkit of a phrase-based SMT (Statistical Machine Translation) model, Moses is applied and Japanese-English translation pairs are obtained in the form of a phrase translation table. Finally, we extract translation pairs of Japanese functional expressions from the phrase translation table. Through this study, we found that most of the semantic equivalence classes judged as monosemous based on manual translation into English have only one translation rules even in the patent domain.

1. Introduction

The Japanese language has various types of functional expressions, which are very important for understanding their semantic contents. Those functional expressions are also problematic in further applications such as MT of Japanese sentences into English. This problem can be partially recognized by the fact that the Japanese language has a large number of variants of functional expressions, where their total number is recently counted as over 10,000 in Matsuyoshi et al. (2006). Based on those recent development in studies on lexicon for processing Japanese functional expressions (Matsuyoshi et al., 2006), this paper studies issues on MT of Japanese functional expressions into English.

More specifically, in order to solve the problem of a large number of variants of Japanese functional expressions, in this paper, we employ the "Sandglass" MT architecture (Yamamoto, 2002). In the "Sandglass" MT architecture, variant expressions in the source language are first paraphrased into representative expressions, and then, a small number of translation rules are applied to the representative expressions. In this paper, we apply this architecture to the task of translating Japanese functional expressions into English, where we introduce a recently compiled large scale hierarchical lexicon of Japanese functional expressions (Matsuyoshi et al., 2006). We employ the semantic equivalence classes of the lexicon and examine each class whether it is monosemous or not. We realize this procedure by manually examining whether functional expressions within a class can be translated into a single canonical English expression. In this step, we refer to Japanese sentences randomly selected from a Japanese corpus of about 8,000 sentences for Japanese language grammar learners (Group Jamashii, 1998). English translation of those randomly selected Japanese sentences are manually annotated.

Based on the results of identifying monosemous semantic equivalence classes, this paper proposes how to extract rules for translating functional expressions in Japanese patent documents into English. In this study, we use about 1.8M Japanese-English parallel sentences automatically extracted from Japanese-English patent families, which are distributed through the Patent Translation Task at the NTCIR-7 Workshop (Fujii et al., 2008). Then, as a toolkit of a phrase-based SMT (Statistical Machine Translation) model, Moses (Koehn et al., 2007) is applied and Japanese-English translation pairs are obtained in the form of a phrase translation table. Finally, we extract translation pairs of Japanese functional expressions from the phrase translation table. Through this study, we found that most of the semantic equivalence classes judged as monosemous based on manual translation into English have only one translation rules even in the patent domain.

2. Japanese Functional Expressions

Even before Matsuyoshi et al. (2006) recently compiled the almost complete list of Japanese functional expressions, there had existed several collections which list Japanese functional expressions and examine their usages. For example, Morita and Matsuki (1989) examined 450 functional expressions and Group Jamashii (1998) also listed 965 expressions and their example sentences. Compared with those two collections, *Gendaigo Hukugouji Youreishu* (National Language Research Institute, 2001) concentrated on 125 major functional expressions which have non-compositional usages, as well as their variants (337 expressions in total)[1], and collected example sentences of those

[1]For each of those 125 major expressions, the differences between it and its variants are summarized as below: i) insertion/deletion/alternation of certain particles, ii) alternation of synonymous words, iii) normal/honorific/conversational forms, iv) base/adnominal/negative forms.

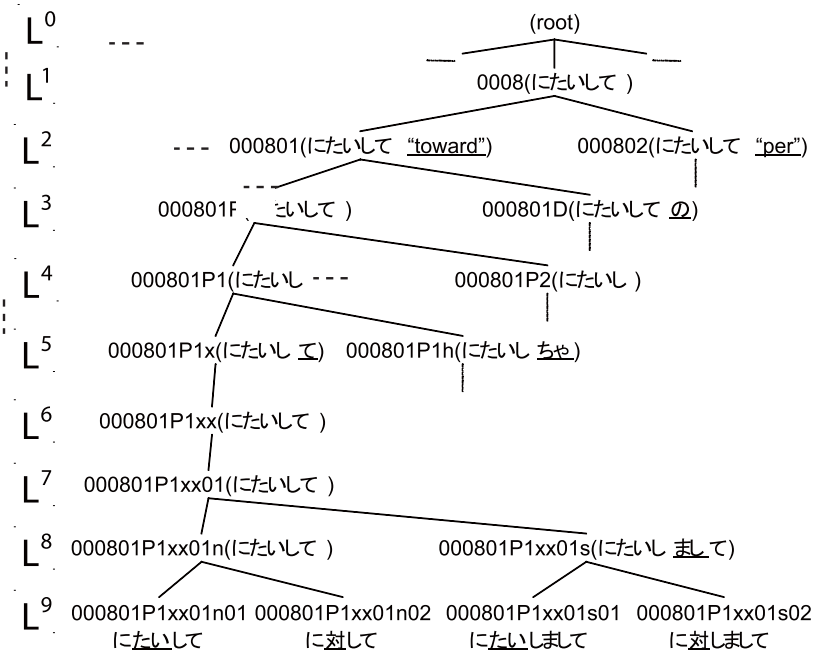

Figure 1: A Part of the Hierarchical Lexicon of Japanese Functional Expressions

expressions. For each of the 337 expressions, Tsuchiya et al. (2005) developed an example database which is used for training/testing a chunker of Japanese (compound) functional expressions. The corpus from which they collected example sentences is 1995 Mainichi newspaper text corpus. For each of the 337 expressions, 50 sentences were collected and labels for chunking were annotated.

3. Hierarchical Lexicon of Japanese Functional Expressions

3.1. Morphological Hierarchy

In order to organize Japanese functional expressions with various surface forms, Matsuyoshi et al. (2006) proposed a methodology for compiling a lexicon of Japanese functional expressions with hierarchical organization[2]. Matsuyoshi et al. (2006) compiled the lexicon with 341 headwords and 16,801 surface forms. The hierarchy of the lexicon has nine abstraction levels and Figure 1 shows a part of the hierarchy[3]. In this hierarchy, the root node (in L^0) is a dummy node that governs all the entries in the lexicon. A node in L^1 is an entry (headword) in the lexicon; the most generalized form of a functional expression. A leaf node (in L^9) corresponds to a surface form (completely-instantiated form) of a functional expression. An intermediate node corresponds to a partially-abstracted (partially-instantiated) form of a functional expression. The second level L^2 distinguishes senses of Japanese functional expressions. This level enables distinction of more than one senses of one functional expression. For example, “にたいして” (ni-taishi-te) has two different senses. The first sense is “to”; e.g., “彼は私 にたいして 親切だ。” (He is kind to me). The second sense is “per”; e.g., “一人 にたいして5つ。” (five per one person). This level is introduced to distinguish such ambiguities. On the other hand, L^3 distinguishes grammatical functions, L^4 distinguishes alternations of function words, L^5 distinguishes phonetic variations, L^6 distinguishes optional focus particles, L^7 distinguishes conjugation forms, L^8 distinguishes normal/polite forms, and L^9 distinguishes spelling variations.

3.2. Semantic Hierarchy

Along with the hierarchy of surface forms of functional expressions with nine abstraction levels, the lexicon compiled by Matsuyoshi et al. (2006) also has a hierarchy of semantic equivalence classes introduced from the viewpoint of paraphrasability. This semantic hierarchy has three abstraction levels, where 435 entries in L^2 (headwords with a unique sense) of the hierarchy of surface forms are organized into the top 45 semantic equivalence classes, the middle 128 classes, and the 199 bottom classes.

Figure 2 shows examples of the bottom 199 classes, where each of the leaf labels “B13”, “B31”, “B32”, “C11”, . . ., “d11”, “d12”, “d13”, . . . represents a label of the bottom 199 classes. In Matsuyoshi and Sato (2008), the bottom 199 semantic equivalence classes of Japanese functional expressions are designed so that functional expressions within a class are paraphrasable in most contexts of Japanese texts.

4. Two Types of Ambiguities of A Compound Expression

One of the most important issues in the processes of acquiring translation rules from parallel patent sentences and applying them is whether each compound expression to which those translation rules are applied is monosemous or not. Unless each compound expression is monosemous, it is necessary to apply certain disambiguation techniques

[2] http://kotoba.nuee.nagoya-u.ac.jp/tsutsuji/

[3] In this lexicon, following Sag et al. (2002), each functional expression is regarded as a fixed expression, rather than a semi-fixed expression or a syntactically-flexible expression.

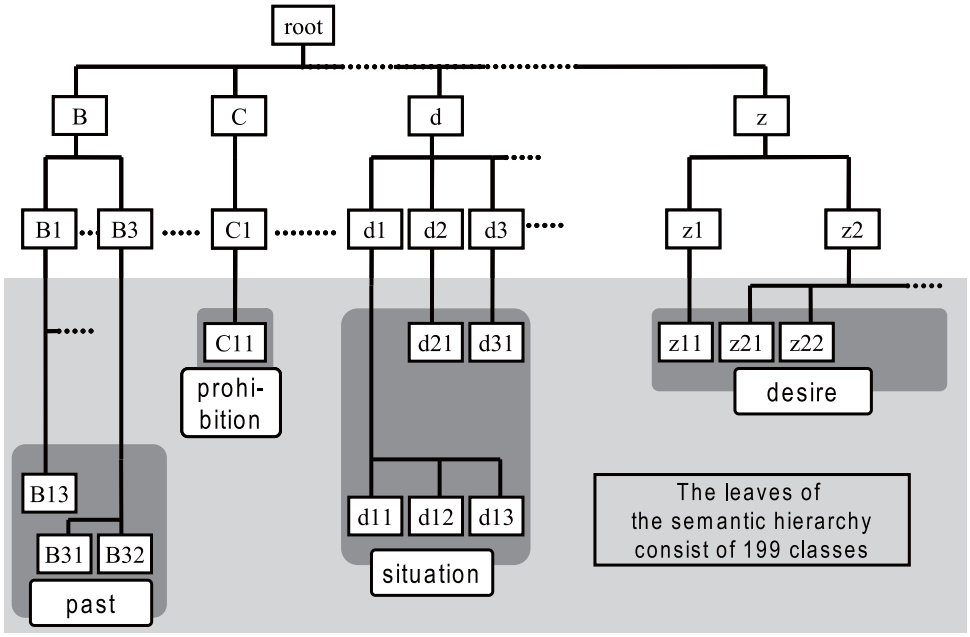

Figure 2: A Part of the Hierarchy of Semantic Equivalence Classes

and then apply translation rules that are appropriate for the actual usage of the target compound expression. Before we discuss how to consider ambiguities of compound expressions in the process of translation rule acquisition, this section first overviews two types of ambiguities of compound expressions.

4.1. Ambiguities of Functional/Content Usages

The first type of ambiguity is for the case that one compound expression may have both a literal (i.e. compositional) *content word* usage and a non-literal (i.e. non-compositional) *functional* usage. This type of ambiguity often happens when the surface form of a functional expression can be decomposed into a sequence of at least one content word and one or more function words. In such a case, the surface form of the compound expression may have both a literal (i.e. compositional) *content word* usage where each of its constituents has its own literal usage, and a non-literal (i.e. non-compositional) *functional* usage where its constituents have no longer their literal usages.

For example, Table 1 (b) shows two example sentences of a compound expression "と (to) は (ha) いえ (ie)", which consists of a post-positional particle "と (to)", a topic-marking particle "は (ha)", and a conjugated form "いえ (ie)" of a verb "いう (iu)". In the sentence (2), the compound expression functions as an adversative conjunctive particle and has a non-compositional functional meaning "*although*". On the other hand, in the sentence (3), the expression simply corresponds to a literal concatenation of the usages of the constituents: the post-positional particle "と (to)", the topic-marking particle "は (ha)", and the verb "いえ (ie)", and has a content word meaning "*can not say*". Compared to Table 1 (b), Table 1 (a) shows an example of a functional expression without ambiguity of functional/content usages. In this case, the compound expression "こと (koto) が (ga) できる (dekiru)" consists of a formal noun "こと (koto)", a post-positional particle "が (ga)", and an auxiliary verb "できる (dekiru)". In almost all the occurrences in a newspaper corpus, the surface form of this compound expression functions as an auxiliary verb and has a non-compositional functional meaning "*can*".

4.2. Ambiguities of Functional Usages

The second type of ambiguity is for the case that the surface form of a functional expression has more than one *functional* usages. For example, Table 1 (c) shows two example sentences of a compound expression "ため (tame) に (ni)", which consists of a noun "ため (tame)" and a post-positional particle "に (ni)". In the sentence (4), the compound expression functions as a case-marking particle and has a non-compositional functional meaning "*for the purpose of*". Also in the sentence (5), the compound expression functions as a case-marking particle, but in this case, has another non-compositional functional meaning "*because of*". Compared to Table 1 (c), Table 1 (a) shows an example of a functional expression without ambiguity of functional usages. In this case, the functional expression "こと (koto) が (ga) できる (dekiru)" has only one non-compositional functional meaning "*can*".

This type of ambiguity includes issues on typical polysemies and homographs, where the issues on sense disambiguation of content words have been well studied in NLP community (e.g, in SENSEVAL tasks (Kilgarriff and Palmer, 2000; Kurohashi and Uchimoto, 2003)). However, in the areas of semantic analysis of Japanese sentences as well as machine translation of Japanese sentences, the issue of sense disambiguation of functional expressions has not been paid much attention so far, and any standard tool for sense disambiguation of Japanese functional expressions have not been publicly available.

Table 1: Two Types of Ambiguities of A Compound Expression

(a) *w/o* ambiguity of functional usages NOR *w/o* ambiguity of functional/content usages

	Expression	Example sentence (English translation)	Usage
(1)	ことができる (koto-ga-dekiru)	彼は英語を話す[ことができる]。 (He *can* speak English.)	functional, semantic class = *possible* (ことができる (koto-ga-dekiru) = *can*)

(b) *w/o* ambiguity of functional usages AND *with* ambiguity of functional/content usages

	Expression	Example sentence (English translation)	Usage
(2)	とはいえ (to-ha-ie)	状況は改善している[とはいえ]、まだ安心できない。 (*Although* it has become better, we can not feel easy.)	functional, semantic class = *adversative* (～ とはいえ (to-ha-ie) = *although* ～
(3)	とはいえ (to-ha-ie)	状況が改善した[とはいえ]ない。 (We *can not say* that it has become better.)	content (～ とはいえ (ない) (to-ha-ie(-nai))) = *can not say* ～)

(c) *with* ambiguity of functional usages

	Expression	Example sentence (English translation)	Usage
(4)	ために (tame-ni)	世界平和の[ために]国際会議が開かれる。 (An international conference is held *for the purpose of* world peace.)	functional, semantic class = *purpose* (ために (tame-ni) = *for the purpose of*
(5)	ために (tame-ni)	雨の[ために]彼の到着が遅れた。 (He arrived late *because of* rain.)	functional, semantic class = *reason* (ために (tame-ni) = *because of*

5. Manually Identifying Monosemous Semantic Equivalence Classes of Functional Expressions in Translation

In terms of translation in English, we manually identify monosemous semantic equivalence classes of Japanese functional expressions (Sakamoto et al., 2009). First, we use the Japanese corpus of about 8,000 sentences for Japanese language grammar learners (Group Jamashii, 1998) as a repository for collecting example sentences of Japanese functional expressions. For each of the 199 semantic equivalence classes, we collect example sentences from this corpus. Here, for each of the 199 classes, we manually judge whether the sense of the functional expression in each sentence corresponds to that of the target class. Then, we keep 91 classes that are with at least five example sentences and we use the total 455 (5 sentences for each of the 91 classes) collected example sentences in further examination for translation into English.

The 455 example sentences are next manually translated into English. Then, for each of the 91 classes, English translation of the Japanese functional expressions in the collected five sentences are compared. Here, if all of the five Japanese functional expressions can be translated into a single canonical English expression, we classify the class as "single translation", and otherwise, as "multiple translations". The "single translation" semantic equivalence classes are considered as monosemous. 49 out of the 91 classes are classified as "single translation", and the remaining 42 as "multiple translations". The 49 "single translation" classes cover more than 6,000 functional expressions[4].

6. Acquiring Translation Rules from Parallel Patent Sentences

6.1. Japanese-English Parallel Patent Documents

In the Japanese-English patent translation task of the NTCIR-7 workshop (Fujii et al., 2008), parallel patent documents and sentences were provided by the organizer. Those parallel patent documents are collected from the 10 years of unexamined Japanese patent applications published by the Japanese Patent Office (JPO) and the 10 years

[4]It is important to note here that, out of those more than 6,000 functional expressions, many have ambiguities of functional/content usages (introduced in section 4.1.) and / or those of functional usages (introduced in section 4.2.). In the next section, we take care of those ambiguities in the process of acquiring translation rules from parallel patent sentences.

Table 2: # of Translation Rules and Functional Expressions for the 17 Semantic Equivalence Classes

# of translation rules per semantic equivalence class	1	2	Total
# of semantic equivalence classes	15	2	17
# of translation rules	15	4	19
# of Japanese functional expressions	33	4	37
# of pairs of Japanese functional expressions and English translation	62	6	68

patent grant data published by the U.S. Patent & Trademark Office (USPTO) in 1993-2000. The numbers of documents are approximately 3,500,000 for Japanese and 1,300,000 for English. Because the USPTO documents consist of only patent that have been granted, the number of these documents is smaller than that of the JPO documents.

From these document sets, patent families are automatically extracted and the fields of "Background of the Invention" and "Detailed Description of the Preferred Embodiments" are selected. This is because the text of those fields is usually translated on a sentence-by-sentence basis. Then, the method of Utiyama and Isahara (2007) is applied to the text of those fields, and Japanese and English sentences are aligned.

6.2. Phrase Translation Table of an SMT Model

As a toolkit of a phrase-based statistical machine translation model, we use Moeses (Koehn et al., 2007) and apply it to the whole 1.8M parallel patent sentences. In Moses, first, word alignment of parallel sentences are obtained by GIZA++ (Och and Ney, 2003) in both translation directions and then the two alignments are symmetrised. Next, any phrase pair that is consistent with word alignment is collected into the phrase translation table and a phrase translation probability is assigned to each pair (Koehn et al., 2003). We finally obtain 76M translation pairs with 33M unique Japanese phrases, i.e., 2.29 English translations per Japanese phrase on average, with Japanese to English phrase translation probabilities $P(p_E \mid p_J)$ of translating a Japanese phrase p_J into an English phrase p_E. For each Japanese phrase, those multiple translation candidates in the phrase translation table are ranked in descending order of Japanese to English phrase translation probabilities.

6.3. The Procedure and the Results of Translation Rule Acquisition

Out of the 49 "single translation" classes, with the lower bound of the phrase translation probability as 0.05 and that of the phrase translation frequency as 10, we extract translation rules for 27 semantic equivalence classes, where, for each of the 27 classes, extracted translation rules include at least one rule which corresponds to the sense of the target class[5]. The 27 classes are divided into two groups; 17 classes and the remaining 10 classes. For each of the 17 classes, all of the extracted translation rules correspond to the sense of the target class. For each of the remaining 10 classes, on the other hand, the extracted translation rules correspond to a mixture of the sense of the target class and other senses/usages. For them, some of the extracted translation rules correspond to content word usages of compound expressions included in one class. Or, functional expressions included in each class have ambiguities of functional usages, and some of the extracted translation rules correspond to a functional usage other than that of the target class.

In the following, we concentrate on the 17 semantic equivalence classes, for which all of the extracted translation rules correspond to the sense of the target class. Within the 17 classes, as shown in Table 2, we actually extract translation pairs for 37 Japanese functional expressions, where the number of extracted translation pairs for those 37 expressions is 68. Here, it is quite important to note that, in the parallel patent sentences, two semantic equivalence classes out of the 17 are not actually "single translation" classes. To put it another way, 15 classes (listed in Table 3) out of the 17 are actually "single translation" classes in the parallel patent sentences. This means that the result of the procedure in section 5. based on the corpus for Japanese language grammar learners (Group Jamashii, 1998) is reliable also in the patent domain to the extent that 15 out of the 17 "single translation" classes are actually with *single translation into English*.

For each of the two "multiple translations" classes, the following lists its sense description as well as multiple translations into English. As shown in Table 4, in the class with a label "m21" with the sense of "restriction", a Japanese functional expression "ほか (hoka)" is translated into an English prepositional phrase "in addition to", while another Japanese functional expression "以外 (igai)" is translated into an English preposition "except" and so on. In the class with a label "P21" with the sense of "exemplification - extreme case", a Japanese functional expression "さえ (sae)" is translated into an English adverb "only", and another Japanese functional expression "でも (demo)" is translated into an English adverb "even". In the hierarchical lexicon of Matsuyoshi et al. (2006), the 17 semantic equivalence classes cover about 541 functional expressions, which can be translated into English by 15+2+2=19 translation rules in the patent domain.

7. Related Works

(Yamamoto, 2002) proposed the "Sandglass" machine translation architecture in which variant expressions in the source language are first paraphrased into representative expressions, and then, a small number of translation rules are

[5]Other than the 27 classes, we extracted translation rules for two more classes, although all of those extracted rules correspond to content word usages of compound expressions included in the two classes. We also extracted translation rules for two more classes, where all of the extracted rules correspond to functional usages other than those of the target classes.

Table 3: 15 “Single Translation” Semantic Equivalence Classes in the Patent Demain

Semantic Equivalence Class	Japanese functional expressions	English Translation
D11 (judgement — necessary)	なければならず, なければならない, べき,	must, must be, to be
G11 (will)	うとし, うとする	to, to be
I12 (estimation — uncertain)	よう	as, so as
R11 (analogy)	のごとく	as, as shown
b11 (object — about)	について, についての, につき, に関し, に関して, に関する	for, in, of, with reference to, with respect to, concerning, related to, relates to, relating to
f12 (range)	にわたって, にわたり, にわたる, に亘って	over
h11 (rate)	に対し, に対しての, に対する	to, for, relative to, with respect to
o11 (simultaneous)	と同時に	at the same time, simultaneously with, together with
p12 (after)	た後で, た後に, てから, てからの	after, after the, from
r12 (conjunction — limitation)	ないと, なければ	unless
r21 (conjunction — condition)	とすれば	if, when
r22 (conjunction — condition)	なら	case, if
r31 (conjunction)	たら	after, if, when
t24 (adversative)	ものの	although, but, though
v21 (parallel)	ながら	while, with

applied to the representative expressions. In this paper, we apply the “Sandglass” architecture to the task of translating Japanese functional expressions into English, where we introduce a recently compiled large scale hierarchical lexicon of Japanese functional expressions (Matsuyoshi et al., 2006; Matsuyoshi and Sato, 2008).

Ambiguities of functional/content usages has been well studied in (Tsuchiya et al., 2005), (Tsuchiya et al., 2006), and (Shudo et al., 2004). (Tsuchiya et al., 2005) reported that, out of about 180 compound expressions which are frequently observed in the newspaper text, one third (about 60 expressions) have this type of ambiguity. Next, (Tsuchiya et al., 2006) formalized the task of identifying Japanese compound functional expressions in a text as a machine learning based chunking problem. The proposed technique performed reasonably well, while its major drawback is in its scale. So far, the proposed technique has not yet been applied to the whole list of over 10,000 Japanese functional expressions. (Shudo et al., 2004) also studied applying manually created rules to the task of resolving functional/content ambiguities, where their approach has limitation in that it requires human cost to create manually and to maintain those rules.

(Utsuro et al., 2007) and (Nivre and Nilsson, 2004) studied syntactic analysis of functional expressions in sentences. (Utsuro et al., 2007) studied how to incorporate the process of analyzing compound non-compositional functional expressions into the framework of Japanese statistical dependency parsing. (Nivre and Nilsson, 2004) also reported improvement of Swedish parsing when multi word units are manually annotated.

8. Concluding Remarks

In the “Sandglass” MT architecture (Yamamoto, 2002), we identified the class of monosemous Japanese functional expressions and utilized it in the task of translating Japanese functional expressions into English. We employed the semantic equivalence classes of a recently compiled large scale hierarchical lexicon of Japanese functional expressions. We then studied whether functional expressions within a class can be translated into a single canonical English expression. Based on the results of identifying