Relative Clause Coordination and Subordination in Japanese

Timothy Baldwin
Tokyo Institute of Technology
tim@cs.titech.ac.jp

Abstract

The research described in this paper is a direct extension of Baldwin et al. (1997), which proposed a declarative rule-based system to analyse gapping in simple Japanese relative clauses. Two shortcomings of this original framework are its inability to handle complex relative clauses, and its non-deterministic nature in choosing between multiple parses for the main verb of the relative clause. Here, we first propose two methods of scoring verb parses to make the system deterministic, and then apply the most successful verb scoring method in the analysis of relative clause subordination and coordination.

1 Introduction

A feature of Japanese relative clauses is their remarkable consistency of surface structure in realising an astounding variety of semantic types. This has made them the target of considerable descriptive literature (Sato, 1989; Kanzaki, 1997; Matsumoto, 1997), but little literature exists on analytical methods to differentiate their full spectrum of use.

In this paper, we first outline the intricacies of Japanese relative clauses, and the basic gapping dichotomy. Next, we introduce the existing system as detailed in Baldwin et al. (1997), and provide explanation of its primary shortcomings. Section 4 discusses types of verb-based lexical ambiguity, and proposes two methods of verb scoring to rank multiple parses. In section 6, we describe subordinated gapping and propose a basic add-on algorithm to identify the clause level at which to analyse the overall relative clause. Finally, we describe factors related to the analysis of coordinated relative clauses, including discussion of future extensions to the basic inter-clausal methodology introduced here.

2 Definitions

2.1 A basic model of Japanese syntax

Case and Valency provide valuable tools in describing Japanese syntax. The predicate is taken to be the nucleus of the clause, and relies on Valency to define case slot compatibility according to the predicate sense and modality. This provides a powerful mechanism to handle both the high levels of zero arguments in Japanese and the relative freedom of word order.

Case slots are made up of a filler ("case filler") and its adposition case marker, with local case slot ordering and the unmarked surface content of the case marker indicating the Case of that slot (see figure 1).

2.2 Relative clauses in Japanese

Japanese relative clauses immediately precede the modified noun head, are generally not inflectionally marked[1], and do not involve relative pronouns.[2]

(1)	*manzoku-sita*	*yūza*
	to be satisfied-PAST	user
	'a satisfied user' / 'a user who is satisfied'	

Semantically speaking, relative clauses can be classified as being either gapping or non-gapping.

Gapping relative clauses

Gapping relative clauses (such as (1)) contain a unique gap for the modified head, the associated case

[1]Inflectional marking does occur with verbal noun-type main verbs, but our research is currently restricted to the consideration of canonical verb-based relative clauses.

[2]The following case marker nomenclature is used in glosses: NOM = nominative, ACC = accusative, COM = comitative, DAT = dative, and QUOT = quotative. Deep case markers are indicated by: SBJ = subject, DO = direct object, and IO = indirect object. "ϕ" is used to indicate zero anaphoric verb complements, and "t_x" to indicate the trace for the head.

Timothy Baldwin (1998) **Relative Clause Coordination and Subordination in Japanese.** In D. Estival (ed.) *Abstracts for the ANLP Post Graduate Workshop,* pp 1-10.

$\text{Filler}_1\text{-marker}_1$	$\text{Filler}_2\text{-marker}_2$	$\cdots$	Predicate
case slot$_1$	*case slot*$_2$		

Figure 1: A case frame model of Japanese syntax

slot of which can be a complement or adjunct. The uniqueness of the case slot gap for the head can be illustrated with (2).

(2)	ϕ/t_i	ϕ/t_j	*au*	*hito*$_{i/j}$
	SBJ	IO	to meet-PRES	person

a. 'people who meet (her)'
OR
b. 'people (she) meets'

Here, the identity of the gap is ambiguous between the subject and indirect object case slots, but simultaneous gapping from these two case slots cannot occur, producing mutual exclusivity between interpretations (2a) and (2b). Even in the case of a reflexive pronoun occupying either case slot within the relative clause, the gap can be seen to be uniquely associated with the uninstantiated case slot, and coreference as being produced indirectly through the gap rather than direct binding by the head.

One feature of gapping relative clauses is that whereas the gap case slot is defined uniquely for a given interpretation, case marking information is not marked either as a gap trace in the relative clause, or as an adposition on the head. This results in the ambiguity seen for (2) above, which does not exist in the deep structure matrix clause counterparts for the respective interpretations:

(3)	*(sono-)hito-ga*	ϕ	*au*
	(that) person-NOM	DO	to meet-PRES

'(that) person meets (her)'

(4)	*(sono-)hito-to*	ϕ	*au*
	(that) person-COM	SBJ	to meet-PRES

'(she) meets (that) person'

Additionally, no distinction is made between topic-type instances of gapping and standard case frame-triggered adjuncts/complements.[3]

[3] In analysing topic-gapping, we classify the topic type as being either the *major subject* or *pure topic*, after Tateishi (1994). Refer to (Baldwin et al., 1997) for a discussion of major subject usages in 'indirect restrictive' clauses.

Non-gapping relative clauses

Non-gapping relative clauses display identical surface syntactic structure to gapping clauses. In this case, however, the head is not gapped from within the relative clause, but rather is a consequence, condition, requisite, simultaneous event, etc. of the modifying clause (Matsumoto, 1997, pp 103-130), or simply restricted by the semantic content of the relative clause. Examples of non-gapping relative clauses are:

(5)	ϕ	ϕ	*au*	*kikkake*
	SBJ	IO	meet-PRES	chance

'an excuse to meet (her)'

(6)	ϕ	*sakana-wo*	*yaku*	*kemuri*
	SBJ	fish-ACC	grill-PRES	smoke

'smoke from grilling fish'

In (5), *kikkake* is simply restricted by its modifying relative clause, whereas *kemuri* in (6) is an inferrable consequence of the event described by the associated relative clause. As is indicated by the zero pronominal complement case slots in both (5) and (6), there is syntactic ambiguity between the clauses being gapping and non-gapping, as a consequence of the scope to map the respective heads onto an uninstantiated case slot in the relative clause. In actual fact, the ellipted case slots in both sentences represent references to context/deixis-evoked entities, a fact which is recoverable only from general discourse processing and sortal restrictions on the various case slots.

3 Relative clause analysis

In our analysis of relative clauses, we apply the algorithm described in (Baldwin et al., 1997), which not only outputs a description of the gapping type (*gapping* or *non-gapping*), but also of the clause subtype within that main type. In the case of gapping clauses, this equates to identifying the case slot from which gapping occurred, whereas for non-gapping clauses, the output corresponds to the basic semantic type of that clause.

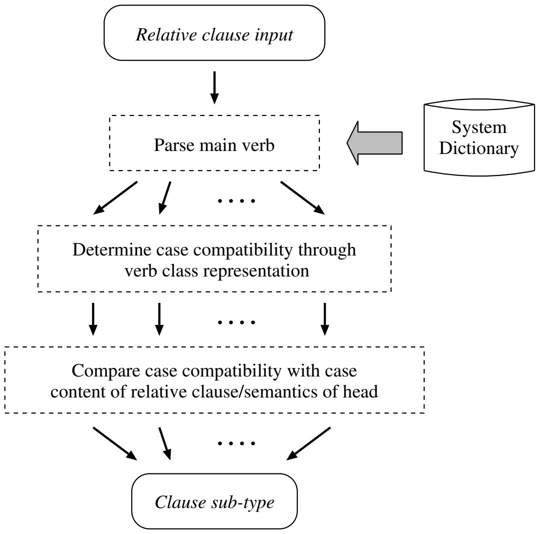

Figure 2: The gapping resolution algorithm

3.1 The algorithm

The algorithm is designed as a declarative rule set, guaranteed to produce a unique analysis for an arbitrary relative clause construction input and corresponding verb entry. The underlying mechanism employed in the algorithm is to apply local heuristics to a complement-based case frame representation of the verb, combined with a description of adjunct potentiality described through verb classes. This combined valency description of the verb interfaces with the inflectional content of the verb, and is combined with a low-level semantic analysis of the head to determine adjunct compatibility and local preferences within the complement component of the case frame (see Figure 2). Essentially, type preferences are determined by way of adjunct prototypicality, and through a balance of the 'most recent filler' strategy (Sakamoto, 1995) and the 'topic worthiness hierarchy' (Hasegawa, 1996; Kuno and Kaburaki, 1977; Kuno, 1978; Silverstein, 1976).

Despite the guarantee of a unique output being returned for a given input and verb entry, the rule set is applied to the full set of successfully parsed verb entries from the system dictionary, which results in potential ambiguity between analyses for distinct entries. Static entry type preferences help to diminish this ambiguity, but beyond this the system has no means of fully alleviating any remaining multiplicity of analysis, and final evaluation of the original system in (Baldwin et al., 1997) is based on the existence of the correct analysis within the candidate outputs produced by the system.

In addition, the system is based on the consideration of a simple matrix relative clause, and as such it is unable to analyse (a) coordinated relative clauses, and (b) cases of gapping from within a subordinate clause.

4 Verb lexical ambiguity

Plurality of successfully parsed entries results from a combination of full inflectional heteronomy/homophony and partial inflectional heteronomy/homophony of verbs.

Full inflectional heteronomy is a direct result of the existence of multiple inter-replaceable writing systems within Japanese (Shibatani, 1990, pp 125-130), and occurs when two distinct verb entries coincide in both conjugational type and phonetic content of the verb stem/auxiliary verb complex. It is distinguishable from polysemy by virtue of the fact that disambiguation is achievable through use of the *kanji* form of the verb stem. An example of verb stem-level full inflectional heteronomy is "*au*", for which three heterogeneous kanji forms produce the distinct entries corresponding to the generic glosses of 'to meet', 'to coincide' and 'to encounter'. Full inflectional heteronomy can alternatively be produced through combinations of auxiliary verb morphemes, such that "*miau*" is ambiguous between *mi-au* 'to see-MUTUAL' and *miau* 'to correspond'.

Full inflectional homophony is analogous to full inflectional heteronomy, except that the ambiguity exists in the kanji-based representation for coinciding conjugational types. In this case, disambiguation is possible through the kana phonetic version of the verb in question. An example of a full inflectional homograph occurs for the verbs *tomeru* 'to stop$_{TRANS}$' and *yameru* 'to quit/put an end to', for which a common kanji corresponds to the "*to*" and "*ya*" prefixes, respectively.

Partial inflectional heteronomy, meanwhile, occurs for verbs which differ in conjugational type, but agree in phonetic content of the verb stem. In this case, heteronomy of kana representation is produced for only certain inflectional types. In the case of our example of *au*, *aru* 'to have' shares the verb stem of *a*, and a heteronym is produced in the simple past tense, in the form of *atta*. Here again, however, kanji representation allows us to resolve the lexical ambiguity. Partial inflectional homonymy closely resembles partial inflectional heteronomy, except that the lexical ambiguity is produced in the kanji form, and resolvable through the use of kana. One example of partial inflectional heteronomy is produced for the simple past tense verbs *i-tta* 'to go-PAST' and *okonatta* 'to carry out/hold-PAST', in that a single kanji is used to represent both "*i-*" and "*okona-*", respectively.

Note that in both of the classifications of partial inflectional correspondence, the degree of coincidence is usually highly restricted, unlike full inflectional heteronomy. For the *i-tta*/*okona-tta* ambiguity, for example, partial inflectional homonymy occurs only in the simple past tense or for progressive/perfective aspect.

5 Resolving verb lexical ambiguity

The most immediate method of resolving representational ambiguity is through statistical means. In this, we tested two methods of statistical weighting, the first based on naive probability, and the second on representational preferences. For both methods, statistical scores were computed only in cases where multiple non-idiomatic entries[4] existed for a common verb stem. Idiomatic entries were automatically allocated a score of one, on the assumption that their fixed case element content is mutually exclusive, and that the system should prefer idiomatic entries over generalised entries.

Due to the difficulty in predicting partial inflectional heteronomy/homophony of verbs, all verbs sharing a common stem are treated as being fully inflectional heteronomic. Note however that for most inflectional types, coincidence of inflectional form does not result. In this case, the preferred entry is the one which has the highest relative score, ignoring the fact that the various scores in question may not total to one.

In terms of the interface between statistical weighting and the rule set, the rule set is applied as is for each parseable verb entry, and weights are summed for each resultant output. The unique system output is determined simply by calculating the highest summed weight, and randomly selecting between multiple analysis types of highest score.

5.1 Calculation of verb scores

The collation of frequencies is based on the EDR corpus (EDR, 1995), and the verb sense annotations given for each verb occurrence. This is the same corpus as was used to extract all relative clause test sets described in this paper, and hence forms a closed test set. Whereas no direct reliance is made on verb sense by our system, the EDR corpus provides a means of determining lexical correspondences between different verb forms. To return to our example of *atta* above, all occurrences of *atta* are attributed a verb sense index, which correspond to different verb 'sense sets'. Contained in these sense sets are one or more representational alternatives of the verb stem *a-*, detectable through the original system dictionary entry. While there is not guarantee of disjunction between the alternative forms of *atta* and their respective sense sets, in almost all cases seen in the EDR corpus, full disambiguation was possible through the granularity of the verb sense index. In cases where sense ambiguity remained, the frequency of the original verb index was equally distributed between polysemous candidates.

One unfortunate characteristic of the EDR corpus is the uncommonly high numbers of index mismatches and 'nil' verb senses (unanalysed/unanalyseable verb senses). In the calculation of verb scores, index mismatches were simply disregarded from the data, while 'nil' indices were treated as described below for the separate scoring methods.

Frequencies are calculated *a priori* and normalised (significant to three figures) to produce the probability of occurrence of that form of the given stem verb.

Naive probability of occurrence

The naive probability of occurrence (NPO) of lexical form a of verb entry f (represented as a_f) is computed simply by totalling the number of usages of verb senses corresponding to a_f, and normalising over the total occurrences of a. Smoothing is achieved by evenly distributing 'nil' occurrences for a between entries a_i, where the total number of distinct entries a_i is represented as $|a|$ in equation (1). Thus, high levels of 'nil' occurrences will produce roughly standardised probabilities for all entries a_i, whereas lower levels of 'nil' occurrences will lead to nearer correspondence between relative frequency and normalised probability. This is intended to reflect the assumption that 'nil' senses suggest inherent ambiguity, and that higher levels of 'nil' values indicate lower confidence on the part of the EDR developers in annotating usages of a.

$$NPO(a_f) \quad = \frac{\frac{freq(a_{nil})}{|a|} + freq(a_f)}{\sum_i freq(a_i)} \tag{1}$$

Normalised representational preference

The representational preference (RP) of lexical form a of verb entry f (i.e. a_f) is defined as the confidence with which we can predict that a will be used to represent f, with the mean confidence predicted as 1. Smoothing is carried out through a double application of Jeffrey's estimate (Good, 1965), that is by adding one to both the numerator and denominator. In this way, low-frequency verb entries and lexical

[4]Non-idiomatic entries refers to 'generalised' entries in (Baldwin et al., 1997).

forms can be smoothed to a value near the mean confidence of one (or to exactly one for zero-frequency entries), but at the same time high-frequency items are relatively unaffected. Additionally, instances of zero denominators are avoided, and the confidence is guaranteed to be strictly greater than zero.

Occurrences of the 'nil' index are not included in the *RP* calculation, such that entries found only with the 'nil' index return a representational preference of one.

$$RP(a_f) \quad = \quad \frac{1+freq(a_f)}{1+\sum_{i\neq a} freq(i_f)} \tag{2}$$

This is normalised over the representational preference for all source entries a_i, to produce the normalised representational preference $NRP(a_f)$.

$$NRP(a_f) \quad = \quad \frac{RP(a_f)}{\sum_i RP(a_i)} \tag{3}$$

5.2 Complexity of inflectional content

The only representational ambiguity not covered by these two scoring systems is instances where inflectional morphemes have produced an ambiguity which was not predictable from the stem verb (see the example of *miau* in section 4). This shortcoming is resolved by introducing the concept of 'complexity of inflectional content' (*CIC*), in which we penalise higher numbers of component inflectional morphemes. The penalty is computed *in situ* based on the number of inflectional morphemes contained in the verb, relative to the parse of simplest inflectional content (*min_infl*); the simplest parse receives a complexity of one. Thus, in the case of "*miau*", *mia-u* 'to correspond-PRES' has a complexity of one, and *mi-a-u* 'to see-MUTUAL-PRES' has a complexity of two. Weighting is achieved through the use of the constant parameter α. That is, the relative contribution of CIC can be enhanced by increasing α, hence exponentially increasing the value of the denominator and reducing the overall verb score (*VS*). At the same time, the parse of simplest inflectional content receives a complexity of one, and its *VS* is hence unaffected by variation in the value of α.

Complexity of inflectional content is compatible with both methods of statistical weighting given above, such that the *VS* for lexical form a of entry f (i.e. a_f) using statistical weighting measure *SW* is computed by:

$$VS(a_f) \quad = \frac{SW(a_f)}{(CIC(a_f)-min\_infl+1)^{\alpha}} \tag{4}$$

5.3 Evaluation of verb scoring

Preliminary evaluation was carried out to determine the relative effectiveness of the naive probability of occurrence (NPO) and normalised representational preference (NRP) methods, and contribution of CIC. The test sets used for this purpose were the full set of annotated relative clauses used in developing the system, and the subset of gapping relative clauses. The sizes of the two test sets are indicated in brackets below each heading.

	Overall (4411)	**Gapping** (3650)
Baseline	84.6%	90.8%
NPO ($\alpha = 1$)	86.0%	92.3%
NPO ($\alpha = 10$)	86.0%	92.3%
NRP ($\alpha = 0$)	85.9%	92.1%
NRP ($\alpha = 1$)	85.8%	92.2%
NRP ($\alpha = 10$)	85.9%	92.2%
Optimal	88.4%	94.5%

Table 1: Results for the verb scoring methods

The baseline method for evaluation purposes simply selects the entry of highest probability when multiple parses are produced, which equates to utilising the naive probability method in computing the verb score, with α set to zero. The optimal achievable result for the system is determined by testing for membership of the correct analysis in the full set of analysis types produced for all successful parses. Given that verb scores simply rank these candidates, it is impossible for the other methods to better this non-deterministic method.

Table 1 lists the comparative results for the various methods, including evaluation of varying values of α for both the NPO and NRP methods. The 1.4% point difference between the overall accuracy for the baseline method and that for the NPO method with various values of α is a direct indication of the effects of weighting according to inflectional complexity, although the ineffectiveness of an increased α value is unexpected.

Likewise for the NRP method, whereas results are significantly higher than those for the baseline method, altering α produced only minor improvement. Indeed, performance with α set to zero (i.e. without consideration of CIC) marginally outperformed NRP with α set to one, although the statistical significance of this difference is questionable. This would tend to suggest that there is some interference in the choice of representational form of the verb stem given complex inflection, a fact which was borne out on summary inspection of the data. That is, the kanji form of the verb stem is generally utilised if auxiliary verbs are also given in a kanji representation, and full hiragana representation is gen-

(7)	(	t_i	*100-ton-izyō*	*aru*)	*to*	*mi-rare-ru*	*zaiko$_i$*
		SBJ	over 100 tonnes	to be-PRES	QUOT	consider-PASS-PRES	stock

'stock considered to be over 100 tonnes (in quantity)'

(8)	(	t_i	*ziken-ni*	*kanyosi-ta*)	*to*	*nihon-ga*	*mite-i-ru*	*kuni$_i$*
		SBJ	incident-DAT	contribute-PAST	QUOT	Japan-NOM	consider-PROG-PRES	country

'countries which Japan considers to have contributed to the incident'

Figure 3: Subordinate gapping clause examples

```
IF (indirect quotational main verb)
    IF (passive or potential main verb OR superordinate subject position instantiated)
       Mark any subordinate gap incompatibilities based on superordinate case content
       IF (gapping resolution of the subordinate clause identifies a gap α) RETURN SUB-α
       ELSE RETURN NON_GAPPING
    ELSE mark any superordinate gap incompatibilities based on subordinate case content
```

Figure 4: The subordinate gapping resolution sub-algorithm

erally reserved for simple inflection uses, such that a hiragana occurrence of "miau" would tend to point to the simple inflectional 'mia-u' stem (see section 4).

Perhaps more noticeable, however, is that the NPO method slightly outperforms NRP, which would tend to suggest that representational preference in isolation is outweighed by the brute force of likelihood of sense.

Based on these results, we adopt the NPO method for the remainder of this paper, with α set to one.

6 Subordinate clause gapping

One important qualification which must be made to our definition of 'gapping' in the context of Japanese relative clauses is that the gapping can occur across a 'bridging' clause. Bridging clauses are defined as containing a suitably marked subordinate clause from which gapping has occurred, and being headed by a main verb which supports the gapping process. Members of this well-defined class of bridging verbs are termed 'indirect quotational', and rely on the subordinate clause being marked with the 'quotative' case marker (*to* - see sentences (7) and (8) in Figure 3). Examples of indirect quotational verbs are *iu* 'to say', *omou* 'to think' and *tutaeru* 'to report'.

In order for subordinate gapping to occur, the main verb in the superordinate relative clause must be potential or passive, or alternatively the superordinate relative clause must contain a surface representation of the clause subject. If these inflectional/syntactic requirements are met, gapping resolution takes place at the subordinate clause level, based on the case frame and inflectional content of its main verb. Interestingly, the same scope of gap types exists at the subordinate level as at the matrix relative clause level. We can thus reuse our original resolution algorithm, excepting that subordinate gapping can only occur across a single 'bridging clause' and hence recursion must be limited to a depth of one.

If a gap is detected within the subordinate clause, the system returns not only the deep case identity of the case slot, but the fact that the gap is subordinate rather than superordinate. In the instance that the subordinate clause is analysed as being non-gapping, the system treats the full relative clause as being non-gapping. The justification behind this analysis is that inflectional constraints simply stipulate the potential source of the gap (subordinate and superordinate) for indirect quotational verbs, without any guarantee of the relative clause being gapping. Additionally, in the case of non-gapping

(9)	(	*pasukaru-ga*	t_i	*kōan-si,*)	ϕ	t_i	*seisaku-si-ta*	*keisan-kikai*$_i$
		Pascal-NOM	DO	design-REN	SBJ	DO	make-PAST	computing device

‘a computing device designed and produced by Pascal’

(10)	(	t_i	*arubaitō-wo*	*si-nagara*)	t_i	*gakkō-ni*	*kayo-u*	*gakusei*$_i$
		SBJ	part-time work-ACC	to do-WHILE	SBJ	school-DAT	attend-PRES	student

‘students who work part-time while at school’

(11)	((	t_i	*i-na-i*)	*to*	*mi-rare*)	t_i	*renraku-sare-na-katta*	*hito*$_i$
		SBJ	to be-NEG-PRES	QUOT	consider-PASS-REN	SBJ	to contact-PASS-NEG-PAST	person

‘a person who was assumed not to be in and (hence) not contacted’

(12)	(	*kankeisya-wo*	*nozo-ki*)	t_i	*pāti-ni*	*syussekisi-ta*	*ninsū*$_i$
		organiser-ACC	exclude-REN	SBJ	party-DAT	attend-PAST	number of people

‘the number of people who attended the party, excluding organisers’

(13)	*yoru*	(	ϕ	*tōkyōwan-wo*	*watari-nagara*)	ϕ	t_i	mi-ru	*reinbōburizzi*$_i$
	night		SBJ	Tokyo Bay-ACROSS	to cross-WHILE	SBJ	DO	to see-PRES	Rainbow Bridge

‘Rainbow Bridge as seen at night while crossing Tokyo Bay’

Figure 5: Coordinated relative clause examples

relative clauses, the head is ‘indirectly associated with the *total* event described by the relative clause’ (Kameyama, 1995, p 168 – my emphasis), making the subordinate/superordinate distinction irrelevant.

In the case of a passive main verb, the superordinate and subordinate clause subject positions become coindexed (see (7)), whereas for other instances of subordinate gapping, the subordinate subject becomes coreferent in content with the superordinate direct object. While recognising that this superficially contradicts our stipulation that gapping occurs from a unique case slot in a given interpretation, we consider the co-indexed case slots to have been merged into one, and analyse the gap as existing in the subordinate clause. Indeed, the only consideration of the corresponding superordinate case slots comes in checking for zero content during gapping resolution, and conversely, for stipulating local gapping incompatibility in the superordinate clause when instantiation of the subordinate-level subject is detected.

The above resolution process for indirect quotational verbs can be summarised by the algorithm given in Figure 4.

	Overall (51)	**Gapping** (45)
Original	49.0%	54.3%
Revised	90.2%	97.9%

Table 2: Results of subordinate gapping analysis

6.1 Evaluation of subordinate gapping

Basic evaluation of the above method was carried out on a set of 51 relative clauses containing an indirect quotational main verb. As with evaluation of verb scoring, the algorithm was further tested on the component subset of gapping relative clauses, with successful gap detection requiring correct identification of the level of embedding of the gap. This derivative test set of 45 gapping relative clauses included 16 *subordinate* gapping clauses. The original algorithm was evaluated on the same data sets to allow for direct comparison of the methods. Results are given in Table 2.

It is perhaps unrealistic to directly compare the results of the two algorithms for gapping clauses, in that the original algorithm is incapable of correctly analysing the 16 gapping subordinate-type clauses.

Having said this, the degree to which the subordinate gapping sub-algorithm outperformed the original algorithm goes beyond the scope of these 16 examples, most importantly as a result of gap incompatibility judgements realised through the revised sub-algorithm. Perhaps more important, however, is that the subordinate gapping sub-algorithm returned higher figures than the overall averages calculated during evaluation of verb scoring (see Table 1).

7 Relative clause coordination

As was seen in the discussion of subordinate gapping, one drawback of the original algorithm is its inability to handle the clause-level structure of relative clauses, a fact which leads to the loss of valuable syntactic restrictional information as to the clause type. This section is devoted to consideration of the further expansion of inter-clausal processing, and its expected benefits.

7.1 Clause coordination

Clause coordination in Japanese is indicated by the use of a coordinating conjunction of the type *nagara*, *te*, *tutu* and *si*, or through *ren'yō* type inflection (aka. *continuative* (Kuno, 1973)). Of these, (Kuno, 1973) observes that *si* and *ren'yō* must be subject coreferential, and (Yoshimoto, 1986) and (Minami, 1974) note that all coordinating connectives tend to coincide in subject or object content.

In terms of relative clause analysis, we wish to suggest (14) as a corollary of the mutual exclusivity of the gapping paradigm:

(14) All coordinated and subordinated clauses in a single relative clause must agree in gapping type.

That is, it is not possible to have a relative clause comprised of both gapping and non-gapping clause components. Additionally, we extend the above observations to hypothesise that:

(15) For semantically coordinated gapping relative clauses, the component clauses must agree in clause sub-type (i.e. gap identity).[5]

By semantic coordination, we wish to distance ourselves from peripheral subordinating usages of *nagara* and *tutu* (in which the *nagara*/*tutu* suffix is interchangeable with *nagaramo* in the contrastive sense and *toki* in the manner sense) and non-additive usages of *te* (Hasegawa, 1996, p 6). Note that as was the case for subordinate gapping, the scope of gapping is unrestricted between complement and adjunct case slots, and includes, in this case, subordinate gaps.

7.2 Processing of clause coordination

By way of accepting hypothesis (15) on gap type correspondence, we are able to extend our algorithm to consider case slot *incompatibilities*, in addition to the existing framework of case slot *compatibility* determination. Case slot incompatibilities stem from two sources: (i) directly from the content of the complement case frame, and (ii) from case slot instantiation. Given a tool set of complement case types, it is possible to determine inherent case incompatibilities directly from the case frame of the verb in question through a simple matching mechanism. This is combined with an analysis of those case slots instantiated in the input, and hence incompatible with that gap through the 'one-case-per-clause' constraint (Fillmore, 1968, p 22).[6] Given that we can expect multiplicity of analysis type due to multiple parses, we take the intersection of gap incompatibilities for each analysis type, and return the resultant set of incompatibilities for the highest scoring analysis type. On the inter-clausal level, the union is taken of the individual incompatibility set for each component clause, in determining the overall incompatibility set.

Determination of the **unique** overall analysis for the relative clause is facilitated through the same process as at the single clause level, in that the weighted outputs for each member clause are summed, and a final sorted list of analysis types determined. However, this is now combined with the incompatibility set to weed out incompatible case types, and the highest scoring compatible clause analysis is outputted. In the case that all analysis types are judged to be case incompatible, the overall clause is assumed to be non-gapping.

7.3 Gap correspondences

Coordination of canonical gapping and subordinated gapping clauses leads to an interesting effect, in that inter-clausal agreement occurs in terms of the gap type, but not as to the clause level from which gapping has occurred (see (11)). It is for this reason that our hypothesis stipulates agreement in clause **sub-type**, but makes no mention of clause **level**. Note that if we had chosen to label subordinated

[5] Note that this coincidence of gap does not apply to indirect restrictive clauses.

[6] Note that application of the one-case-per-clause constraint is restricted to **complement** case slots.

	Overall (181)	Gapping (104)
Original	66.9%	87.7%
Revised	74.6%	91.2%

Table 3: Results of coordinated clause analysis

gaps according to their gap type in the superordinate clause, coincidence of gap type would not occur for non-passive subordinated gapping clauses.

Two verb types which do not contribute to the clause sub-type, and are hence disregarded during the resolution process, are the *excluding* and *including* types. These restrict/exemplify the set membership of the overall situation described by the relative clause construction, by identifying excluded/included elements. As such, excluding and including clauses are clause modifying constructions, accounting for their treatment as subordinated clauses and avoiding any inconsistency with hypothesis (15). Considering (12), in which the first clause is of the excluding type, the main clause is essentially treated as a single uncoordinated clause, and the subject gapping sub-type can be recovered.

One fact which is clear from the original description of conjunction types is that peripheral subordinating usages exist for all conjunctions except the ren'yō form, suggesting difficulty in correctly predicting the type of clause dependency in a given clause prior to being able to apply the restrictions proposed in section 7.2. While this is certainly the case for *te* clauses, complement analysis-based heuristics were found to be productive in correctly analysing *nagara* and *tutu* clauses. These heuristics consist of analysing the complement content of the coordinated clause to determine if all non-subject complement case slots are instantiated. If full instantiation is detected, the unit clause in question is momentarily removed from the resolution process. If analysis of the remaining clause content of that relative clause returns a non-subject case slot analysis, the original *nagara* or *tutu* clause must have been subordinated, whereas if a subject case slot analysis is produced, the original clause must have been coordinated (as an extension of (Kuno, 1973)). In this second case, the subject gap extends to the *nagara* or *tutu* clause. This process can be seen to correctly identify the subordinated *nagara* clause in (13), with the direct object gap existing only in the main clause.

7.4 Evaluation of clause coordination

The basic method outlined above was tested on a set of 181 relative clauses containing multiple clause instances marked with the *nagara*, *si* and *tutu* conjunctions, or and ren'yō inflection. As a means of comparison, the original algorithm was used to analyse the same test set, and accuracy on gapping relative clauses contained in the original test set was calculated. The results for the evaluation are given in Table 3.

Clearly, the revised method of handling interclausal dependency outperforms the original algorithm, although the disparity between the respective results is perhaps not as marked as could have been expected. One of the main sources of error was that coindexed zero subjects tended to be mistaken as subject gaps, which accounted for around 75% of the errors in both cases. Perhaps more important, though, is the fact that hypothesis (15) was upheld for all observed relative clause instances, and that the heuristic for distinguishing between coordinated and subordinated relative clauses worked successfully on all applications.

7.5 The treatment of subordinate clauses

A preliminary study of the relative clause corpus produced for the system suggested that around 6–7% of all relative clauses involve clause coordination, pointing to the significance of the above method of analysing coordinate clause complexes. Relative clauses containing subordinate clauses (excluding cases of subordinate gapping), however, seem to account for a much higher proportion, at around 20% of all relative clauses. While the clause subtype hypothesis proposed above does not apply to subordinated relative clause constructions, the more general suggestion of coincidence of clause type (gapping vs. non-gapping) is suggested to apply to all relative clauses. As is evident in all levels of evaluation, the accuracy of the system for non-gapping clauses is significantly lower than that for gapping clauses, and the application of this basic restriction presents itself as a possible tool in enhancing resolution of the clause type.

In terms of identifying gap variation between subordinate and superordinate clauses, (Okumura and Tamura, 1996, p 874) suggest that 'subject switching' occurs given a surface subject in either the subordinate or superordinate clause, although they go on to suggest that variations in gap type are largely context dependent and not predictable simply from local constraints. The application of their proposed heuristic, and further analysis of the gap switching mechanism, however, remain as outstanding issues in the system development.

8 Conclusions

We have proposed two verb scoring methods, biased according to inflectional simplicity, to make our system deterministic in output. The first of the two verb scoring methods, based on simple frequency of occurrence, proved to be slightly more effective, and comparative in performance with the original nondeterministic evaluation method.

More importantly, perhaps, we introduced the notion of subordinate gapping, and proposed an add-on algorithm which enables the original algorithm to successfully analyse this gapping type. This increased the overall accuracy of the algorithm on potentially subordinate gapping clauses from 49.0% to 90.2%. We then went on to hypothesise that the clause gapping type must be unique across coordinated clauses within a relative clause, and combined this with analysis of subordinate gapping to increase the resolution accuracy for coordinated relative clauses from 66.9% to 74.6%.

Acknowledgements

I would like to thank Hozumi Tanaka (TITech) and Takenobu Tokunaga (TITech) for their supervision in preparing this paper, and Francis Bond (NTT) for his characteristically insightful comments. Additionally, Atushi Fujii (TITech) provided valuable feedback on the effectivity and evaluation of the scoring method.

References

T. Baldwin, T. Tokunaga, and H. Tanaka. 1997. Semantic verb classes in the analysis of head gapping in Japanese relative clauses. In *Proceedings of the 4th Natural Language Processing Pacific Rim Symposium 1997 (NLPRS'97)*, pages 409–14.

EDR, 1995. *EDR Electronic Dictionary Technical Guide*. Japan Electronic Dictionary Research Institute, Ltd. (In Japanese).

C.J. Fillmore. 1968. The case for case. In E. Bach and R.T. Harms, editors, *Universals in linguistic theory*, pages 1–88. New York: Holt, Rinehart and Winston.

I.J. Good. 1965. *The Estimation of Probabilities*. MIT Press, Cambridge MA.

Y. Hasegawa. 1996. *A Study of Japanese Clause Linkage*. CSLI.

M. Kameyama. 1995. The syntax and semantics of the Japanese Language Engine. In *Japanese Sentence Processing* (Mazuka and Nagai, 1995), pages 153–76.

K. Kanzaki. 1997. Lexical semantic relations between adnominal constituents and their head nouns. *Mathematical Linguistics*, 21(2):53–68. (In Japanese).

S. Kuno and E. Kaburaki. 1977. Empathy and syntax. *Linguistic Inquiry*, 8(4):627–72.

S. Kuno. 1973. *The Structure of the Japanese Language*. MIT Press, Cambridge MA.

S. Kuno. 1978. *Danwa no Bunpō*. Tokyo: Taishukan. (In Japanese).

Y. Matsumoto. 1997. *Noun Modifying Constructions in Japanese*. John Benjamins.

R. Mazuka and N. Nagai. 1995. *Japanese Sentence Processing*. Lawrence Erlbaum Associates.

F. Minami. 1974. *Gendai Nihongo no Kōzō*. Taishukan.

M. Okumura and K. Tamura. 1996. Zero pronoun resolution in Japanese discourse based on centering theory. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING '96)*, pages 871–6.

T. Sakamoto. 1995. Transparency between parser and grammar: On the proceessing of empty subjects in Japanese. In *Japanese Sentence Processing* (Mazuka and Nagai, 1995), pages 275–94.

R. Sato. 1989. *Research relating to the semantic analysis of Japanese attributive clauses*. Master's thesis, Tokyo Institute of Technology. (In Japanese).

M. Shibatani. 1990. *The Languages of Japan*. CUP.

M. Silverstein. 1976. Hierarchy of features and ergativity. In R.M.W. Dixon, editor, *Grammatical Categories in Australian Languages*, pages 112–71. Humanities Press.

K. Tateishi. 1994. *The Syntax of 'Subjects'*. CSLI.

K. Yoshimoto. 1986. Study of Japanese zero pronouns in discourse processing. In *IPSJ SIG Notes*, volume 86, no. 52. (In Japanese).