Memory-Inductive Categorial Grammar: An Approach to Gap Resolution in Analytic-Language Translation

Prachya Boonkwan **Thepchai Supnithi**

Human Language Technology Laboratory
National Electronics and Computer Technology Center (NECTEC)
112 Thailand Science Park, Phaholyothin Road,
Khlong 1, Pathumthani 12120, Thailand
{prachya.boonkwan, thepchai.supnithi}@nectec.or.th

Abstract

This paper presents a generalized framework of syntax-based gap resolution in analytic language translation using an extended version of categorial grammar. Translating analytic languages into Indo-European languages suffers the issues of gapping, because “deletion under coordination” and “verb serialization” are necessary to be resolved beforehand. Rudimentary operations, i.e. antecedent memorization, gap induction, and gap resolution, were introduced to the categorial grammar to resolve gapping issues syntactically. Hereby, pronominal references can be generated for deletion under coordination, while sentence structures can be properly selected for verb serialization.

1 Background

Analytic language, such as Chinese, Thai, and Vietnamese, is any language whose syntax and meaning relies on particles and word orders rather than inflection. Pronouns and other grammatical information, such as tense, aspect, and number, expressed by use of adverbs and adjectives, are often omitted. In addition to *deletion under coordination* and *verb serialization*, called *gapping* (Hendriks, 1995), translation from analytic languages into Indo-European ones becomes a hard task because (1) an ordinary parser cannot parse some problematic gapping patterns and (2) these omissions are necessary to be resolved beforehand. We classify resolution of the issue into two levels: syntactic/semantic and pragmatic. Gapping, which we considered as a set of bound variables, can be resolved in syntactic/semantic level (Partee, 1975). Omission of other grammatical information is, on the contrary, to be resolved in pragmatic level because some extra-linguistic knowledge is required. Consequently, we concentrate in this paper the resolution of gapping by means of syntax and semantics.

Many proposals to gap resolution were introduced, but we classify them into two groups: non-ellipsis-based and ellipsis-based. *Non-ellipsis-based approach* is characterized by: (a) strong proof system (Lambek, 1958), and (b) functional composition and type raising that allow coordination of incomplete constituents, such as CG (Ajdukiewicz, 1935; Bar-Hillel, 1953; Moortgat, 2002), CCG (Steedman, 2000), and multimodal CCG (Baldridge and Kruijff, 2003). Proposals in this approach, such as (Hendriks, 1995; Jäger, 1998a; Jäger, 1998b), introduced specialized operators to resolve overt anaphora, while covert anaphora is left unsolved. *Ellipsis-based approach* is characterized by treating incomplete constituents as if they are of the same simple type but contain ellipsis inside (Yatabe, 2002; Cryssmann, 2003; Beavers and Sag, 2004). However, Beavers and Sag (2004) evidenced that ellipsis-based analysis possibly reduces the acceptability of language, because the resolution is *per se* completely uncontrolled.

In this paper, we introduce an integration of the two approaches that incorporates strong proof system and ellipsis-based analysis. Antecedent memorization and gap induction are introduced to imitate ellipsis-based analysis. The directions of ellipsis are also used to improve the acceptability of language.

The rest of the paper is structured as follows. Section 2 describes the formalization of our method. Section 3 evidences the coverage of the framework on coping with the gapping issues in analytic languages. Section 4 further discusses coverage and limitations of the framework comparing with CG and its descendants. Section 5 explains relevance of the proposed formalism to MT. Finally, Section 6 concludes the paper and lists up future work.

2 Memory-Inductive Categorial Grammar

Memory-Inductive Categorial Grammar, abbreviated MICG, is a version of pure categorial grammar extended by ellipsis-based analysis. On the contrary, it relies on antecedent memorization, gap induction, and gap resolution that outperform CCG's functional composition and type raising.

All grammatical expressions of MICG are, like CG, distinguished by a syntactic category identifying them as either a function from arguments of one type to result another (a.k.a. *function*), or an argument (a.k.a. *primitive category*). Let us exemplify the MICG by defining an example grammar G below.

$$\begin{array}{rcl} \textsf{John}, \textsf{Mary}, \textsf{sandwich}, \textsf{noodle} & \vdash & np \\ \textsf{eats} & \vdash & (np\backslash s)/np \\ \textsf{and} & \vdash & \& \end{array}$$

The lexicons John, Mary, sandwich, and noodle are assigned with a primitive category np. The lexicon eats is assigned with a function that forms a sentence s after taking np from the right side ($/np$) and then taking np from the left side ($np\backslash$). The lexicon and is assigned with a conjunction category (&). By means of syntactic categories assigned to each lexicon, the derivation for a simple sentence 'John eats noodle' is shown in (1).

(1)
$$\dfrac{\dfrac{\textsf{John}}{\textsf{John} \vdash np} \quad \dfrac{\dfrac{\textsf{eats}}{\textsf{eats} \vdash (np\backslash s)/np} \quad \dfrac{\textsf{noodle}}{\textsf{noodle} \vdash np}}{\textsf{eats} \circ \textsf{noodle} \vdash np\backslash s}}{\textsf{John} \circ (\textsf{eats} \circ \textsf{noodle}) \vdash s}$$

CG suffers some patterns of coordination e.g. SVO&SO as exemplified in (2).

(2) John eats noodle, and Mary, sandwich.

One should find that the second conjunct cannot be reduced into s by means of CG, because it lacks of the main verb 'eats.' The main verb in the first conjunct should be remembered and then filled up to the ellipsis of the second conjunct to accomplish the derivation. This matter of fact motivated us to develop MICG by introducing to CG the process of remembering an antecedent from a conjunct, called *memorization*, and filling up an ellipsis in the other conjunct, called *induction*. There are three mandatory operations in MICG: antecedent memorization, gap induction, and gap resolution.

One of two immediate formulae combined in the derivation can be memorized as an antecedent. The resulted syntactic category is modalized by the modality $\Box_F^D$, where D is a direction of memorization ($<$ for the left side and $>$ for the right side), and F is the memorized formula. The syntactic structure of the memorized formula is also modalized with the notation $\Box$ to denote the memorization. It is restricted in MICG that the memorized formula must be unmodalized to maintain mild context-sensitivity. For example, let us consider the derivation of the first conjunct of (2), 'John eats noodle,' with antecedent memorization at the verb 'eats' in (3). As seen, a modalized formula can combine with another unmodalized formula while all modalities are preserved.

(3)
$$\dfrac{\dfrac{\textsf{John}}{\textsf{John} \vdash np} \quad \dfrac{\dfrac{\textsf{eats}}{\Box\textsf{eats} \vdash (np\backslash s)/np} \quad \dfrac{\textsf{noodle}}{\textsf{noodle} \vdash np}}{\Box\textsf{eats} \circ \textsf{noodle} \vdash \Box^{<}_{\textsf{eats} \vdash (np\backslash s)/np}(np\backslash s)}}{\textsf{John} \circ (\Box\textsf{eats} \circ \textsf{noodle}) \vdash \Box^{<}_{\textsf{eats} \vdash (np\backslash s)/np} s}$$

Any given formula can be induced for a missing formula, or a *gap*, at any direction, and the induced gap contains a syntactic category that can be combined to that of the formula. The resulted syntactic category of combining the formula and the gap is modalized by the modality $\Diamond_F^D$, where D is a direction of induction, and F is the induced formula at the gap. The syntactic structure of F is an uninstantiated variable and also modalized with the notation $\Diamond$ to denote the induction. The induced formula is necessary to be unmodalized for mild context-sensitivity. For example, let us consider the derivation of the second conjunct of (2), 'Mary, sandwich,' with gap induction before the word 'sandwich' in (4). The variable of syntactic structure will be resolved with an appropriate antecedent containing the same syntactic category in the gap resolution process.

(4)
$$\frac{\dfrac{\text{Mary}}{\text{Mary}\vdash np}\qquad\dfrac{\dfrac{\text{sandwich}}{\text{sandwich}\vdash np}}{\Diamond X\circ\text{sandwich}\vdash\Diamond^{<}_{X\vdash(np\backslash s)/np}(np\backslash s)}}{\text{Mary}\circ(\Diamond X\circ\text{sandwich})\vdash\Diamond^{<}_{X\vdash(np\backslash s)/np}s}$$

Gap resolution matches between memorized antecedents and induced gaps to associate ellipses to their antecedents during derivation of coordination and serialization. That is, two syntactic categories $\Box^{D_1}_{F_1}C$ and $\Diamond^{D_2}_{F_2}C$ are matched up and canceled from the resulted syntactic category, if they have the same syntactic categories C, their directions D_1 and D_2 are equal, and their memorized/induced formulae F_1 and F_2 are unified. For example, let us consider the derivation of 'John eats noodle, and Mary, sandwich' in Figure 1. The modalities $\Box^{<}_{\text{eats}\vdash(np\backslash s)/np}s$ and $\Diamond^{<}_{X\vdash(np\backslash s)/np}s$ are matched up together. Their memorized/induced formulae are also unified by instantiating the variable X with 'eats'. Eventually, after combining them and the conjunction 'and,' the derivation yields out the formula $(\text{John}\circ(\Box\text{eats}\circ\text{noodle}))\circ(\text{and}\circ(\text{Mary}\circ(\Diamond\text{eats}\circ\text{sandwich})))\vdash s$.

Gap resolution could also indicate argument sharing in coordination and serialization. $\Diamond^{D_1}_{F_1}C$ and $\Diamond^{D_2}_{F_2}C$ can be also matched up, if they have the same syntactic categories C, their directions D_1 and D_2 are equal, and their memorized/induced formulae F_1 and F_2 are unified. However, they must be preserved in the resulted syntactic category. For example, let us consider the derivation in Figure 2. By means of unification of induced formulae, the variables X and Y are unified into the variable Z.

A formal definition of MICG is given in Appendix A. MICG is applied to resolve deletion under coordination and serialization in analytic languages in the next section.

3 Gap Resolution in Analytic Languages

There are two causes of gapping in analytic languages: coordination and serial verb construction. Each of which complicates the analysis module of MT to resolve such issue before transferring. In this section, problematic gapping patterns are analyzed in forms of generalized patterns by MICG. For simplification reason, syntactic structure is suppressed during derivation.

3.1 To resolve gapping under coordination

Coordination in analytic languages is more complex than that of Indo-European ones. Multi-conjunct coordination is suppressed here because biconjunct coordination can be applied. Besides SVO&VO and SV&SVO patterns already resolved by CCG (Steedman, 2000), there are also SVO&SV, SVO&V, SVO&SO (already illustrated in Figure 1), and SVO&SA patterns.

The pattern SVO&SV exhibits ellipsis at the object position of the second conjunct. The analysis of SVO&SV is illustrated in (5). It shows that the object of the first conjunct is memorized while the verb of the second conjunct is induced for the object.

(5)
$$\frac{\dfrac{\dfrac{\text{S}}{np}\quad\dfrac{\dfrac{\text{V}}{(np\backslash s)/np}\quad\dfrac{\text{O}}{np}}{\Box^{>}_{np}(np\backslash s)}}{\Box^{>}_{np}s}\quad\dfrac{\&}{\&}\quad\dfrac{\dfrac{\text{S}}{np}\quad\dfrac{\dfrac{\text{V}}{(np\backslash s)/np}}{\Diamond^{>}_{np}(np\backslash s)}}{\Diamond^{>}_{np}s}}{s}$$

Analysis of the sentence pattern SVO&V, illustrated in (6), exhibits ellipses at the subject and the object positions of the second conjunct. The subject and the object of the first conjunct are memorized, while the verb of the second conjunct is induced twice for the object and for the subject, respectively.

(6)
$$\frac{\dfrac{\dfrac{\text{S}}{np}\quad\dfrac{\dfrac{\text{V}}{(np\backslash s)/np}\quad\dfrac{\text{O}}{np}}{\Box^{>}_{np}(np\backslash s)}}{\Box^{<}_{np}\Box^{>}_{np}s}\quad\dfrac{\&}{\&}\quad\dfrac{\dfrac{\dfrac{\text{V}}{(np\backslash s)/np}}{\Diamond^{>}_{np}(np\backslash s)}}{\Diamond^{<}_{np}\Diamond^{>}_{np}s}}{s}$$

The pattern SVO&SA exhibits ellipsis at the predicate position of the second conjunct, because only the adverb (A) is left. Suppose the adverb, typed $(np\backslash s)/(np\backslash s)$, precedes the predicate. Illustrated in (7), the predicate of the first conjunct is memorized, while the adverb of the second conjunct is inducted for the predicate.

(7)
$$\frac{\dfrac{\dfrac{\text{S}}{np}\quad\dfrac{\dfrac{\text{V}}{(np\backslash s)/np}\quad\dfrac{\text{O}}{np}}{np\backslash s}}{\Box^{>}_{np\backslash s}s}\quad\dfrac{\&}{\&}\quad\dfrac{\dfrac{\text{S}}{np}\quad\dfrac{\dfrac{\text{A}}{(np\backslash s)/(np\backslash s)}}{\Diamond^{>}_{np\backslash s}(np\backslash s)}}{\Diamond^{>}_{np\backslash s}s}}{s}$$

$$\dfrac{\dfrac{\text{John eats noodle}}{\text{John}\circ(\Box\text{eats}\circ\text{noodle})\vdash\Box^{<}_{\text{eats}\vdash(np\backslash s)/np}s}\quad\dfrac{\text{and}}{\text{and}\vdash\&}\quad\dfrac{\text{Mary, sandwich}}{\text{Mary}\circ(\Diamond X\circ\text{sandwich})\vdash\Diamond^{<}_{X\vdash(np\backslash s)/np}s}}{(\text{John}\circ(\Box\text{eats}\circ\text{noodle}))\circ(\text{and}\circ(\text{Mary}\circ(\Diamond\text{eats}\circ\text{sandwich})))\vdash s}$$

Figure 1: Derivation of ‘John eats noodle, and Mary, sandwich.’

$$\dfrac{\dfrac{\text{eats noodle}}{\Diamond X\circ(\text{eats}\circ\text{noodle})\vdash\Diamond^{<}_{X\vdash np}s}\quad\dfrac{\text{and}}{\text{and}\vdash\&}\quad\dfrac{\text{drinks coke}}{\Diamond Y\circ(\text{drinks}\circ\text{coke})\vdash\Diamond^{<}_{Y\vdash np}s}}{(\Diamond Z\circ(\text{eats}\circ\text{noodle}))\circ(\text{and}\circ(\Diamond Z\circ(\text{drinks}\circ\text{coke})))\vdash\Diamond^{<}_{Z\vdash np}s}$$

Figure 2: Preservation of modalities in derivation

3.2 To resolve gapping under serial verb construction

Serial verb construction (SVC) (Baker, 1989) is construction in which a sequence of verbs appears in what seems to be a single clause. Usually, the verbs have a single structural object and share logical arguments (Baker, 1989). Following (Li and Thompson, 1981; Wang, 2007; Thepkanjana, 2006), we classify SVC into three main types: consecutive/concurrent events, purpose, and circumstance.

No operation specialized for tracing antecedent projection in consecutive/concurrent event construction has been proposed in CG or its descendants. In MICG, the serialization operation is specialized for this construction. For example, a Chinese sentence from (Wang, 2007) in (8) is analyzed as in (9).

(8) tā mǎi piào jīn qù
he buy ticket enter go
‘He buys a ticket and then goes inside.’

(9)
$$\dfrac{\dfrac{\dfrac{\text{tā}}{np}\quad\dfrac{\dfrac{\text{mǎi}}{(np\backslash s)/np}\quad\dfrac{\text{piào}}{np}}{np\backslash s}}{\Box^{<}_{np}s}\quad\dfrac{\dfrac{\dfrac{\text{jīn}}{np\backslash s}}{\Diamond^{<}_{np}s}\quad\dfrac{\dfrac{\text{qù}}{np\backslash s}}{\Diamond^{<}_{np}s}}{\Diamond^{<}_{np}s}}{s}$$

Illustrated in (9), the subject argument tā ‘he’ is projected through the verb sequence by means of memorization and induction modalities.

Purpose construction can also be handled by MICG. For example, a Thai sentence in (10) is analyzed as in (11).

(10) kʰǎu tɔ̀ː tʰɔ̂ː paj ɕʰáj naj bâːn
he attach pipe go use in house
‘He attaches pipes to use in the house.’

(11)
$$\dfrac{\dfrac{\dfrac{\dfrac{\text{kʰǎu}}{np}\quad\dfrac{\dfrac{\text{tɔ̀ː}}{(np\backslash s)/np}\quad\dfrac{\text{tʰɔ̂ː}}{np}}{\Box^{>}_{np}(np\backslash s)}}{\Box^{<}_{np}\Box^{>}_{np}s}\quad\dfrac{\text{paj}}{s\backslash s}}{\Box^{<}_{np}\Box^{>}_{np}s}\quad\dfrac{\dfrac{\dfrac{\dfrac{\text{ɕʰáj}}{(np\backslash s)/np}}{\Diamond^{>}_{np}(np\backslash s)}}{\Diamond^{<}_{np}\Diamond^{>}_{np}s}\quad\dfrac{\dfrac{\text{naj}}{(s\backslash s)/np}\quad\dfrac{\text{bâːn}}{np}}{s\backslash s}}{\Diamond^{<}_{np}\Diamond^{>}_{np}s}}{s}$$

Illustrated in (11), the two logical arguments, i.e. the subject kʰǎu ‘he’ and the object tʰɔ̂ː ‘pipe,’ are projected through the construction.

SVC expressing circumstance of action is syntactically considered much as consecutive event construction. For example, a Chinese sentence from (Wang, 2007) in (12) is analyzed as in (13).

(12) wǒ yòng kuàizi chī fàn
I use chopstick eat meal
‘I eat meal with chopsticks.’

(13)
$$\dfrac{\dfrac{\dfrac{\text{wǒ}}{np}\quad\dfrac{\dfrac{\text{yòng}}{(np\backslash s)/np}\quad\dfrac{\text{kuàizi}}{np}}{np\backslash s}}{\Box^{<}_{np}s}\quad\dfrac{\dfrac{\dfrac{\text{chī}}{(np\backslash s)/np}\quad\dfrac{\text{fàn}}{np}}{np\backslash s}}{\Diamond^{<}_{np}s}}{s}$$

4 Coverage and Limitations

Proven in Theorem 1 in Appendix A, memorized constituents and induced constituents are cross-serially associated. Controlled by order and direction, each memorized constituent is guaranteed to be cross-serially associated to its corresponding induced gap, while each gap pair is also cross-serially associated revealing argument sharing. This causes cross-serial association, illustrated in Figure 3, among memorized constituents and induced gaps. Since paired modalities are either eliminated or preserved and no modalities are left on the start symbol, it guarantees that there is eventually no modality in derivation. In conclusion, no excessive gap is over-generated in the language.

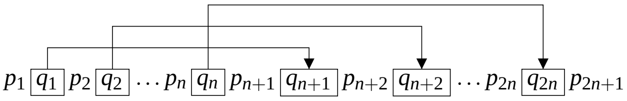
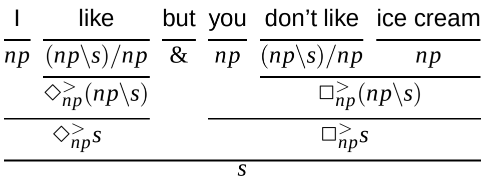
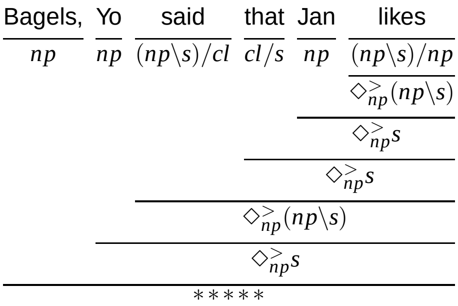

p_1 q_1 p_2 q_2 $\cdots$ p_n q_n p_{n+1} q_{n+1} p_{n+2} q_{n+2} $\cdots$ p_{2n} q_{2n} p_{2n+1}

Figure 3: Cross-serial association

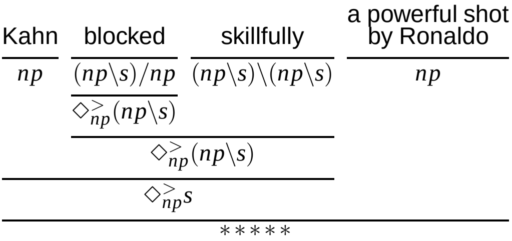
MICG's antecedent memorization and gap induction perform well in handling node raising. Node raising is analyzed in terms of MICG by memorizing the raised constituent at the conjunct it occurs and inducing a gap at the other conjunct. For example, the right node 'ice cream' is raised in the sentence 'I like but you don't like *ice cream*.' The sentence can be analyzed in terms of MICG in (14).

(14)


Topicalization and contraposition are still the issues to be concerned for coverage over CCG. For example, in an example sentence 'Bagels, Yo said that Jan likes' from (Beavers and Sag, 2004), the NP 'Bagels' is topicalized from the object position of the relative clause's complement. (15) shows unparsability of the sentence.

(15)


Furthermore, constituent shifting, such as dative shift and adjunct shift, is not supported by MICG. We found that it is also constituent extraction as consecutive constituents other than the shifted one are extracted from the sentence. For example, the adjunct 'skillfully' is shifted next to the main verb in the sentence 'Kahn blocked skillfully a powerful shot by Ronaldo' from (Baldridge, 2002) in (16).

(16)

Kahn
blocked
skillfully
a powerful shot by Ronaldo
np
$(np\backslash s)/np$
$(np\backslash s)\backslash(np\backslash s)$
np
$\Diamond^{>}_{np}(np\backslash s)$
$\Diamond^{>}_{np}(np\backslash s)$
$\Diamond^{>}_{np}s$
*****

Since MICG was inspired by reasons other than those of CCG, the coverage of MICG is therefore different from CCG. Let us compare CG, CCG, and MICG in Table 1. CCG initially attempted to handle linguistic phenomena in English and other Indo-European languages, in which topicalization and dative shift play an important role. Applied to many other languages such as German, Dutch, Japanese, and Turkish, CCG is still unsuitable for analytic languages. MICG instead was inspired by deletion under coordination and serial verb construction in analytic languages. We are in progress to develop an extension of MICG that allows topicalization and dative shift avoiding combinatoric explosion.

5 Relevance to RBMT

Major issues of MT from analytic languages into Indo-European ones include three issues: anaphora generation, semantic duplication, and sentence structuring. Both syntax and semantics are used to solve such problems by MICG's capability of gap resolution. Case studies from our RBMT are exemplified for better understanding.

Our Thai-English MT system is rule-based and consists of three modules: analysis, transfer, and generation. MICG is used to tackle sentences with deletion under coordination and SVC which cannot be parsed by ordinary parsers. For good speed efficiency, an MICG parser was implemented in GLR-based approach and used to analyze the syntactic structure of a given sentence before transferring. The parser detects zero anaphora and resolves their antecedents in coordinate structure, and reveals argument sharing in SVC. Therefore, coordinate structure and SVC can be properly translated.

No experiment has been done on our system yet, but we hope to see an improvement of translation quality. We planned to evaluate the translation accuracy by using both statistical and human methods.

Table 1: Coverage comparison among CG, CCG, and MICG (Y = supported, N = not supported)

Linguistic phenomena	CG	CCG	MICG
Basic application	Y	Y	Y
Node raising	N	Y	Y
Topicalization/contraposition	N	**Y**	N
Constituent shifting	N	**Y**	N
Deletion under coordination	N	N	**Y**
Serial verb construction	N	N	**Y**

5.1 Translation of deletion under coordination

Coordinate structures in Thai drastically differ from those of English. This is because Thai allows zero anaphora at subject and object positions while English does not. Pronouns and VP ellipses must therefore be generated in place of deletion under coordination for grammaticality of English. Moreover, semantic duplication is often made use to emphasize the meaning of sentence, but its direct translation becomes redundant.

MICG helps us detect zero anaphora and resolve their antecedents, so that appropriate pronouns and ellipses can be generated at the right positions. By tracing resolved antecedents and ellipses, argument projections are disclosed and they can be used to control verb fusion. We exemplify three cases of translation of coordinate structure.

Case 1: Pronouns are generated to maintain grammaticality of English translation if the two verbs are not postulated in the verb-fusion table. For example, a Thai sentence in (17) is translated, while pronouns ‘he’ and ‘it’ are generated from Thai NPs nákˑrian ‘student’ and kʰàˑnǒm ‘candy,’ respectively.

(17)

nákˑrian$_S$	sɯ́ː$_V$	kʰàˑnǒm$_O$	lɛ́ːʋ$_\&$	kin$_V$
student	buy	candy	then	eat

‘A student buys candy, then *he* eats *it*.’

Case 2: Two verbs V_1 and V_2 are fused togetherif they are postulated in the verb-fusion table to eliminate semantic duplication in English translation. The object form of S_2 is necessary to be generated in some cases. For example, in (18), the translation becomes ‘He reports her this matter’ instead of ‘He tells her to know this matter.’ Two verbs bɔ̀ːk ‘tell’ and sâːb ‘know’ are fused into a single verb ‘report.’ The object form of ‘she,’ ‘*her*,’ is also generated.

(18)

kʰǎʋ$_S$	bɔ̀ːk$_V$	hâj$_\&$	thəː$_S$	sâːp$_V$	rɯ̂ːəŋ níː$_O$
he	tell	TO	she	know	this matter

‘He *reports her* this matter.’

Case 3: A VP ellipsis is generated to maintain English grammaticality. For example, in (19), a VP ellipsis ‘*do*’ is generated from a Thai VP mâi ɕʰɔ̂ːb donˑtriː rɔ́k ‘not like rock music.’

(19)

ɕɔːn$_S$	ɕʰɔ̂ːp$_V$	donˑtriː rɔ́k$_O$	tɛ̀ː$_\&$	ɕʰǎn$_S$	mâi$_A$
John	like	rock music	but	I	not

‘John likes rock music, but I *do* not.’

5.2 Translation of SVC

Sentence structuring is also nontrivial for translation of Thai SVC. Thai uses SVC to describe consecutive/concurrent events, purposes, and circumstances. On the other hand, English describes each of those with different sentence structure. A series of verbs with duplicated semantics can be also clustered to emphasize the meaning of sentence in Thai, while English does not allow this phenomenon.

Because MICG reveals argument sharing in SVC, appropriate sentence structures can be selected by tracing argument sharing between two consecutive verbs. We exemplify two cases of translation of SVC.

Case 1: The second verb is participialized if the first verb is intransitive and its semantic concept is an action. For example, the present participial form of the verb ‘see,’ ‘*seeing*,’ is generated in (20) .

(20)

sǒmˑɕʰaːj$_S$	dəːn$_V$	ɕʰom$_V$	pʰâːpˑkʰiǎn$_O$
Somchai	walk	see	paintings

‘Somchai walks *seeing* paintings.’

Case 2: If the two cases above do not apply to the two verbs, they are translated directly by default. The conjunction ‘and’ is automatically added to conjoin two verb phrases. In case of multiple-conjunct coordination, the conjunction will be added only before the last conjunct. For example, in (21), a pronoun ‘*it*’ is generated from the NP kʰóːk ‘coke,’ while the conjunction ‘and’ is automatically added.

(21)	pʰîːˑsǎːʊ$_S$	sɯ́ː$_V$	kʰóːk$_O$	dɯ̀ːm$_V$
	my elder sister	buy	coke	drink
	‘My elder sister buys coke *and* drinks *it*.’			

6 Conclusion and Future Work

This paper presents Memory-Inductive Categorial Grammar (MICG), an extended version of categorial grammar, for gap resolution in analytic language translation. Antecedent memorization, gap induction, and gap resolution, are proposed to cope with deletion under coordination and serial verb construction. By means of MICG, anaphora can be generated for deletion under coordination, while sentence structure can be properly selected for serial verb construction. No experiment has been done to show improvement of translation quality by MICG.

The following future work remains. First, we will experiment on our Thai-English RBMT to measure improvement of translation quality. Second, criteria for pronominal reference generation in place of deletion under coordination will be studied. Third, once serial verb construction is analyzed, criteria of sentence structuring will further be studied based on an analysis of antecedent projection. Fourth and finally, constituent extraction and the use of extraction direction in the extraction resolution will be studied to avoid combinatoric explosion.

References

K. Ajdukiewicz. 1935. Die Syntaktische Konnexität. *Polish Logic*, pages 207–231.

M. C. Baker. 1989. Object Sharing and Projection in Serial Verb Constructions. *Linguistic Inquiry*, 20:513–553.

J. Baldridge and G. J. M. Kruijff. 2003. Multimodal combinatory categorial grammar. In *Proceedings of the 10th Conference of the European Chapter of the ACL 2003*, Budapest, Hungary.

J. Baldridge. 2002. *Lexically Specified Derivational Control in Combinatory Categorial Grammar*. Ph.D. thesis, University of Edinburgh.

Y. Bar-Hillel. 1953. A Quasi-Arithmetical Notation for Syntactic Description. *Language*, 29:47–58.

J. Beavers and I. A. Sag. 2004. Coordinate ellipsis and apparent non-constituent coordination. In *Proceedings of the HPSG04 Conference*. Center for Computational Linguistics, Katholieke Universiteit Leuven, CSLI Publications.

B. Cryssmann. 2003. An asymmetric theory of peripheral sharing in HPSG: Conjunction reduction and coordination of unlikes. In *Proceedings of Formal Grammar Conference*.

P. Hendriks. 1995. Ellipsis and multimodal categorial type logic. In *Proceedings of Formal Grammar Conference*. Barcelona, Spain.

G. Jäger. 1998a. Anaphora and ellipsis in type-logical grammar. In *Proceedings of the 1th Amsterdam Colloquium*, Amsterdam, the Netherland. ILLC, Universiteit van Amsterdam.

G. Jäger. 1998b. Anaphora and quantification in categorial grammar. In *Lecture Notes in Computer Science; Selected papers from the 3rd International Conference, on logical aspects of Computational Linguistics*, volume 2014, pages 70–89.

J. Lambek. 1958. The Mathematics of Sentence Structure. *American Mathematical Monthly*, 65:154–170.

C. N. Li and S. A. Thompson. 1981. *Mandarin Chinese: A Functional Reference Grammar*. Berkeley: University of California Press.

M. Moortgat. 2002. Categorial grammar and formal semantics. In *Encyclopedia of Cognitive Science*, volume 1, pages 435–447. Nature Publishing Group.

B. H. Partee. 1975. Bound variables and other anaphors. In *Theoretical Issues in Natural Language Processing-2 (TINLAP-2)*, pages 79–85, University of Illinois at Urbana Champaign, July.

M. Steedman. 2000. *The Syntactic Process*. The MIT Press, Cambridge, Massachusetts.

K. Thepkanjana. 2006. Properties of events expressed by serial verb constructions in Thai. In *Proceedings of the 11th Biennial Symposium: Intertheoretical Approaches to Complex Verb Constructions*, Rice University.

X. Wang. 2007. Notes about Serial Verb Constructions in Chinese. *California Linguistic Notes*, 32(1).

S. Yatabe. 2002. A linearization-based theory of summative agreement in peripheral-node raising constructions. In *Proceedings of the HPSG02 Conference*, Standford, California. CSLI Publications.

A Formal Definition of MICG

Definition 1 (Closure of MICG) *Let V_A of category symbols, a finite set V_T of terminal symbols, and a set of directions $D = \{<, >\}$.*

The set C of all category symbols is given by: (1) For all $x \in V_A$, $x \in C$. (2) If $x, y \in C$, then so are $x\backslash y$ and x/y. (3) If $x \in C$, then so are $\Box^{<}_{f}x$, $\Box^{>}_{f}x$, $\Diamond^{<}_{f}x$, and $\Diamond^{>}_{f}x$, where $f \in F$ is a formula (described below). (4) Nothing else is in C.

The set T of all grammatical structures is given by: (1) For all $x \in V_T$, $x \in T$. (2) If $x, y \in T$, then so are $x \circ y$. (3) If $x \in T$, then so are $\Box x$ and $\Diamond x$. (4) Nothing else is in T.

The set F of all formulae is a set of terms $t \vdash x$, where $t \in T$ and $x \in C$. The set Q of all modalities is a set of all terms $\Box^{<}_{f}$, $\Box^{>}_{f}$, $\Diamond^{<}_{f}$, and $\Diamond^{>}_{f}$, where $f \in F$.

Definition 2 (Modality resolution) *For any directions $d \in D$, any formulae $f \in F$, and any modality sequences $\mathbf{M}, \mathbf{M}_1, \mathbf{M}_2 \in Q^*$, the function $\oplus : Q^* \times Q^* \mapsto Q^*$ is defined as follows:*

$$\begin{array}{rcl}
\Box^{d}_{f}\mathbf{M}_1 \oplus \Diamond^{d}_{f}\mathbf{M}_2 & \equiv & \mathbf{M}_1 \oplus \mathbf{M}_2 \\
\Diamond^{d}_{f}\mathbf{M}_1 \oplus \Box^{d}_{f}\mathbf{M}_2 & \equiv & \mathbf{M}_1 \oplus \mathbf{M}_2 \\
\Box^{d}_{f}\mathbf{M}_1 \oplus \Box^{d}_{f}\mathbf{M}_2 & \equiv & \Box^{d}_{f}(\mathbf{M}_1 \oplus \mathbf{M}_2) \\
\Diamond^{d}_{f}\mathbf{M}_1 \oplus \Diamond^{d}_{f}\mathbf{M}_2 & \equiv & \Diamond^{d}_{f}(\mathbf{M}_1 \oplus \mathbf{M}_2)
\end{array}$$

$$\varepsilon \oplus \mathbf{M} \quad \equiv \quad \mathbf{M} \oplus \varepsilon \quad \equiv \quad \mathbf{M}$$

Definition 3 (MICG) *A memory-inductive categorial grammar (MICG) is defined as a quadruple $G = \langle V_T, V_A, s, R \rangle$, where: (1) V_T and V_A are as above. (2) $s \in V_A$ is the designated symbol called 'start symbol.' (3) $R : V_T \mapsto P(F)$ is a function assigning to each terminal symbol a set of formulae from F. The set of all strings generated from G is denoted as $L(G)$.*

Definition 4 (Acceptance of strings) *For any formulae $x, y \in F$, any grammatical structures $t_1, t_2, t_3 \in T$, any variables v of grammatical structures, and any modality sequences $\mathbf{M}, \mathbf{M}_1, \mathbf{M}_2 \in Q^*$, the binary relation $\models \subseteq F^* \times F$ controls combination of formulae as follows:*

$$\begin{array}{rcl}
t_1 \vdash y \quad t_2 \vdash y\backslash x & \models & t_1 \circ t_2 \vdash x \\
t_1 \vdash x/y \quad t_2 \vdash y & \models & t_1 \circ t_2 \vdash x \\
t_1 \vdash y \quad t_2 \vdash \mathbf{M}y\backslash x & \models & \Box t_1 \circ t_2 \vdash \Box^{<}_{t_1 \vdash y}\mathbf{M}x \\
t_1 \vdash \mathbf{M}y \quad t_2 \vdash y\backslash x & \models & t_1 \circ \Box t_2 \vdash \Box^{>}_{t_2 \vdash y\backslash x}\mathbf{M}x \\
t_1 \vdash x/y \quad t_2 \vdash \mathbf{M}y & \models & \Box t_1 \circ t_2 \vdash \Box^{<}_{t_1 \vdash x/y}\mathbf{M}x \\
t_1 \vdash \mathbf{M}x/y \quad t_2 \vdash y & \models & t_1 \circ \Box t_2 \vdash \Box^{>}_{t_2 \vdash y}\mathbf{M}x \\
t_2 \vdash \mathbf{M}y\backslash x & \models & \Diamond v \circ t_2 \vdash \Diamond^{<}_{v \vdash y}\mathbf{M}x \\
t_1 \vdash \mathbf{M}y & \models & t_1 \circ \Diamond v \vdash \Diamond^{>}_{v \vdash y\backslash x}\mathbf{M}x \\
t_2 \vdash \mathbf{M}y & \models & \Diamond v \circ t_2 \vdash \Diamond^{<}_{v \vdash x/y}\mathbf{M}x \\
t_1 \vdash \mathbf{M}x/y & \models & t_1 \circ \Diamond v \vdash \Diamond^{>}_{v \vdash y}\mathbf{M}x \\
t_1 \vdash \mathbf{M}_1x \quad t_3 \vdash \& \quad t_2 \vdash \mathbf{M}_2x & \models & t_1 \circ (t_3 \circ t_2) \vdash (\mathbf{M}_1 \oplus \mathbf{M}_2)x \\
t_1 \vdash \mathbf{M}_1x \quad t_2 \vdash \mathbf{M}_2x & \models & t_1 \circ t_2 \vdash (\mathbf{M}_1 \oplus \mathbf{M}_2)x
\end{array}$$

The binary relation $\Rightarrow \subseteq F^* \times F^*$ holds between two strings of formulae $\alpha X \beta$ and $\alpha Y \beta$, denoted $\alpha X \beta \Rightarrow \alpha Y \beta$, if and only if $X \models Y$, where $X, Y, \alpha, \beta \in F^*$ and $|X| \geq |Y|$. The relation $\Rightarrow^*$ is the reflexive transitive closure of $\Rightarrow$.

A string $w \in V_T^*$ is generated by G, denoted by $w \in L(G)$, if and only if $w = w_1 \ldots w_n$ and there is some sequence of formulae $f_1 \ldots f_n$ such that $f_i \in R(w_i)$ for all $1 \leq i \leq n$, and $f_1 \ldots f_n \Rightarrow^* s$. That is, $w_1 \ldots w_n$ is generated if and only if there is some choice of formula assignments by R to the symbols in $w_1 \ldots w_n$ that reduces to s.

Definition 5 *Correspondence between a grammatical structure and its syntactic category can be viewed as a tree with specialized node types. Each node is represented (m, S), where m is a node type $\{\emptyset, \Box, \Diamond\}$, and S is a modality sequence attached to the node's syntactic category.*

Definition 6 *A node that has the type m is said to be* marked *m where $m \in \{\Box, \Diamond\}$, while a node that has the type $\emptyset$ is said to be* unmarked.

Definition 7 *The function $\tau : Q \mapsto \{\Box, \Diamond\}$ maps a modality to a node modality, where $\tau(\Box^{d}_{f}) = \Box$ and $\tau(\Diamond^{d}_{f}) = \Diamond$ for all $d \in D$ and $f \in F$.*

Definition 8 *A substring generated from a node marked $\tau(\boldsymbol{M})$ beneath the node n is said to be* unpaired *under n, if and only if n has the modality sequence S and $\boldsymbol{M} \in S$.*

Definition 9 *Every string w generated from MICG can be rewritten in the form $w = p_1q_1 \ldots p_lq_lp_{l+1}q_{l+1} \ldots p_{2l}q_{2l}p_{2l+1}$, where q_i is a substring unpaired under n, p_j is a substring generated from unmarked nodes beneath n, $1 \leq i \leq l$, $1 \leq j \leq l+1$, and $l \geq 0$.*

Theorem 1 (Cross-serial association) *For every string generated from MICG $w = p_1q_1 \ldots p_lq_lp_{j(l)}q_{j(1)} \cdots p_{j(l)}q_{j(l)}p_{j(l)+1}$, every couple q_i and $q_{j(i)}$ are associated by $\oplus$ for all $1 \leq i \leq l$, where $j(i) = l + i$ and $l \geq 0$.*

Proof Let us prove this property by mathematical induction.

Basic step: Let $l = 0$. We obtain that $w_0 = p_1$. Since there is no unpaired substring, this case is trivially proven.

Hypothesis: Let $l = k$. Suppose that $w_k = p_1q_1 \cdots p_{j(k)}q_{j(k)}p_{j(k)+1}$. We rewrite $w_k = w_k^1w_k^2$, where $w_k^1 = p_1q_1 \cdots p_kq_kp'_{j(1)}$ and $w_k^2 = p''_{j(1)}q_{j(1)} \cdots p_{j(k)}q_{j(k)}p_{j(k)+1}$. Every couple q_i and $q_{j(k)}$ are associated by $\oplus$ for all $1 \leq i \leq k$.

Induction: Let $l = k+1$; $w_{k+1} = p_1q_1 \cdots p_{j(k)+2}q_{j(k)+2}$ $p_{j(k)+3}$, consequently. Let the formulae of the substrings $w_{k+1} = w_{k+1}^1w_{k+1}^2$ be $t_{k+1}^1 \vdash m_1\mathbf{M}_1$ and $t_{k+1}^2 \vdash m_2\mathbf{M}_2$, respectively. We can rewrite the substrings $w_{k+1} = w_{k+1}^1w_{k+1}^2$ in terms of $w_k = w_k^1w_k^2$ in three cases.

Case I: Suppose $w_{k+1}^1 = pqw_k^1$. It follows that the direction of q is $<$. Since w_{k+1}^1 combines w_{k+1}^2, we can conclude that $w_{k+1}^2 = p'q'w_k^2$. Therefore, q and q' are also associated by $\oplus$.

Case II: Suppose $w_{k+1}^1 = w_k^1qp$. It follows that the direction of q is $>$. Since w_{k+1}^1 combines w_{k+1}^2, we can conclude that $w_{k+1}^2 = w_k^2q'p'$. Therefore, q and q' are also associated by $\oplus$.

Case III: $w_{k+1}^1 = p_1q_1 \ldots p_mq_mpqp_{m+1}q_{m+1} \cdots p_nq_np_{k+1}$ and $w_{k+1}^2 = p_{j(1)}q_{j(1)} \cdots p_{j(m')}q_{j(m')}p'q'p_{j(m')+1}q_{j(m')+1} \cdots p_{j(k)}$ $q_{j(k)}p_{j(k)+1}$, where $1 < m, m' < k$. Since w_{k+1}^1 and w_{k+1}^2 combine and every q_i and $q_{j(i)}$ are associated, we can conclude that $m = m'$. Therefore, q and q' are also associated by $\oplus$.

From Case I, Case II, and Case III, we can rewrite $w_{k+1}^1 = p'_1q'_1p'_2q'_2 \cdots p'_{k+1}$ and $w_{k+1}^2 = p'_{j(1)}q'_{j(1)}p'_{j(2)}q'_{j(2)} \cdots p'_{j(k+1)}$. Since each q_i in w_k^1 and $q_{j(i)}$ in w_k^2 are already associated by $\oplus$, it follows that all q_i and $q_{j(i)+1}$ are also associated. ■