[*AMTA/SIG-IL First Workshop on Interlinguas*, San Diego, CA, October 28, 1997]

Interlingua Developed and Utilized in Real Multilingual MT Product Systems

Shin-ichiro KAMEI and Kazunori MURAKI

C&C Media Research Laboratories, NEC Corporation
4-1-1, Miyazaki, Miyamae-ku, Kawasaki, 216 JAPAN
kamei@ccm.cl.nec.co.jp, k-muraki@ccm.cl.nec.co.jp

abstract

This paper describes characteristics of an interlingua we have developed. It contains a large lexicon and has been tested on actual MT systems in the translation of large volumes of actual documents. The main characteristics of the interlingua are as follows: (1) Conceptual primitives, elements of the interlingua, can be linked to any parts of speech in English or Japanese. (2) Positions of the top node on the interlingua correspond to differences in syntactic structures. (3) Two or more conceptual graphs can be used for expressing the same concept, and can be converted to another by conceptual transformation rules which are independent of any specific language. (4) Conceptual primitives are divided into two classes; (a) functional conceptual primitives, which are finite and manageable and constitute, along with rules for interpreting conceptual graphs, the grammar of the interlingua, and (b) general conceptual primitives, which correspond to specific words in actual languages and which, depending on the direction of translation, may or may not be used. Our commercial MT products using the interlingua produce results of roughly the same or higher quality than systems using the syntactic transfer method, which fact indicates the feasibility of the interlingua approach.

1 Introduction

Machine Translation (MT) systems generally have intermediate structures between source and target languages. Such structures must, at the very least, be able to cope with the following two situations:

(1) The same concept is expressed with significantly different syntactic structures in source and target languages.
(2) A concept existing in the source language is not easily expressible in the target language.

Syntactic Transfer (ST) systems utilize two intermediate structures to cope with these situations, one representing the sentence structure of the source language and the other representing that of the target language. Syntactic transformation is used to transform the source sentence structure to that of the target. This method is widely used in many actual MT systems because it is easy to apply to the creation of an MT system that is to be dedicated to single direction translation between one specific source and one specific target language.

The ST method is far less suitable, however, for multilingual MT systems, i.e. those not limited in number of languages or in the direction of translation. Further, the intermediate structures used in ST systems do not express the semantic structure of an input sentence, which makes the method unsuited to future extension to treatments of discourse, i.e. of the overall meaning of whole documents.

For such a purposes, an intermediate structure must be capable of expressing the concepts contained in sentences, while it itself remains structurally independent of both source and target language.

An interlingua is one intermediate structure that satisfies this condition and is also able to cope with the two previously noted "situations." The interlingua approach is well-suited to multilingual systems, and has been the subject of much study for such purposes (Muraki 84, 86, Farwell and Wilks 91, CICC 95, EDR 95).

The interlingua approach does, however, present some difficulties. It is difficult, for example, to establish the set of conceptual primitives the interlingua is to use, and it is also difficult to determine whether or not the interlingua approach is ultimately capable of handling bi-directional translation. Such determinations cannot be made without building an actual Interlingua method MT system and using it to treat large vocabularies in a wide variety of and different types of documents.

That is why almost all current commercial MT products use the ST method; as far as we know, the only interlingua applicable to commercial MT systems operating in two directions is that which we report here, which is used in both our English-to-Japanese and Japanese-to-English systems.

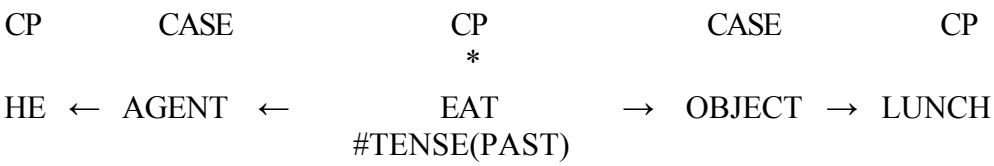
We first developed a preliminary interlingua for the purpose of handling English and Japanese, these the two languages seeming suitable to interlingua development since they are differ so fundamentally in syntactic structure and word sense coverage. Using this as a base, we built large dictionaries (about a hundred thousand (100,000) entries for each language), developed machine translation systems to handle as wide range of sentence types as possible, and translated many different types of actual documents.

The focus of this paper is not a comparison of various MT methods (interlingua, transfer, example-base), but the interlingua itself that we have developed and used in our MT systems. This paper describes its characteristics.

2 Basic Structure of Interlingua

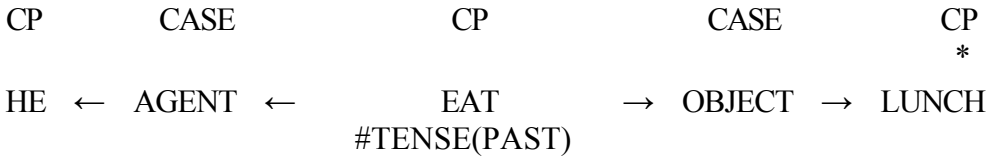
2.1 Structure of the Interlingua

Our interlingua is basically a directed, acyclic graph composed of two types of conceptual primitives. The conceptual graph consists of two types of conceptual primitives. Those of the first type, which we refer to here as "content primitives (CPs)," express the meanings of content words, primarily nouns and verbs, adjectives and adverbs. The second type of conceptual primitive, which we refer to as "CASE," represents the relationship between CPs and often expresses the meaning of such function words as prepositions and conjunctions. In the interlingua, CP nodes and CASE nodes are arranged alternately, with CASE nodes connecting CP nodes. The following shows an interlingua graph corresponding to the English sentence 'He ate lunch.'



We should note here that the CASE nodes have directions; the arrows between EAT and AGENT, and between AGENT and HE, for example, indicate that the AGENT of the EAT is HE. TENSE, ASPECT, etc. are expressed as features on nodes, e.g. #TENSE(PAST) indicating the past tense of the verb 'eat.'

Also significant is the placement of an asterisk (*) over the "locus node;" in the previous example, it is over EAT. By the way of contrast, while the structure below expresses the same node relations as the previous example, the position of the asterisk over LUNCH indicates that this graph is expressing the English phrase 'the lunch (which) he ate.'



2.2 Structure of Dictionary

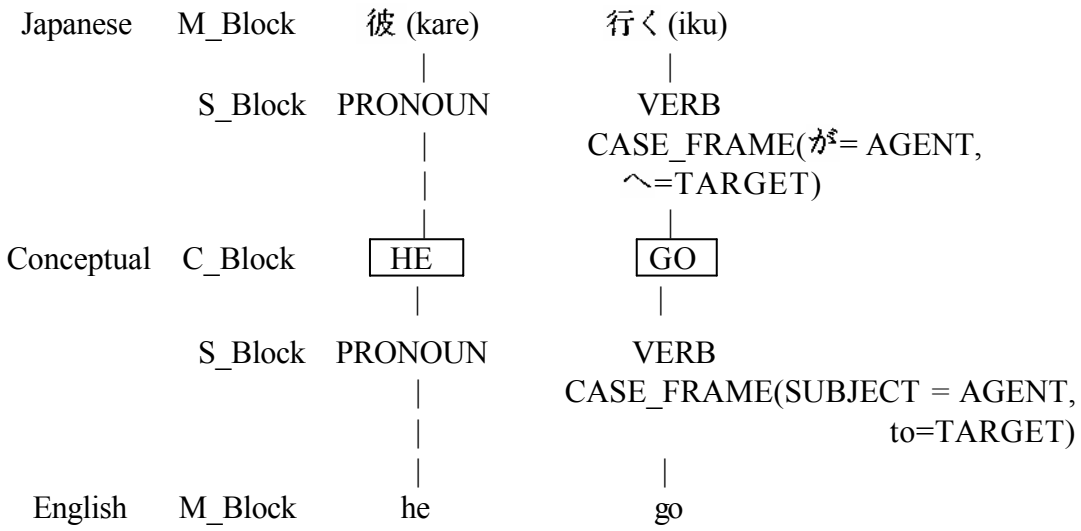
Our Interlingua dictionary is composed of the following three types of blocks:

(a) Morpheme Blocks (M_Blocks):
This block contains morphological information regarding each entry (spelling, etc.). Each source or target language has its own Morpheme Block.

(b) Syntax Blocks (S_Blocks):
This block contains syntactic information regarding each entry (part of speech, etc.). Each source or target language has its own Syntax Block.

(c) Concept Block (C_Block):
This block contains conceptual information regarding each entry, particularly its semantic categories.
Source and target languages share this block in common.
All of the conceptual primitives used in the interlingua are contained in this block.

The following figure illustrates the dictionary structure for a the Japanese-English dictionary.

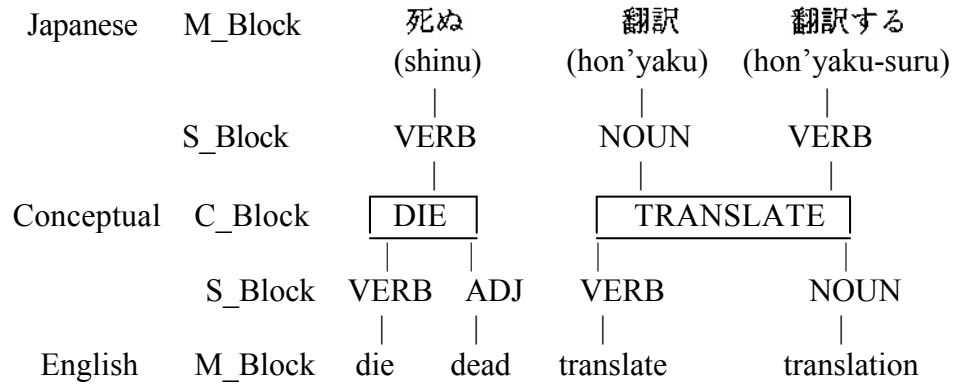


While in many cases Japanese nouns may correspond to English nouns and Japanese verbs to English verbs, actual parts of speech used may vary considerably between the two languages. For example, the parts of speech used in Japanese and English are different in the following expressions.

Japanese	English
死んでいる (shinde-iru)	dead
verb phrase	adjective phrase

Additionally, in many cases, a language may offer a choice of more than one part of speech as a possible translation, as in the case of a translation of the Japanese "彼の本の翻訳を行なった," for which reasonably natural English might be "translated his book" or "executed a translation of his book."

In order to be able to cope with such situations, all parts of speech in one language must be linkable (by way of the C_Block) to all parts of speech in the other. For the above examples, the dictionary might exhibit the following structure:



3 Interlingua Characteristics

3.1 Equivalent Interlingua Structures

The same concepts expressed in the following two sentences are all most the same:

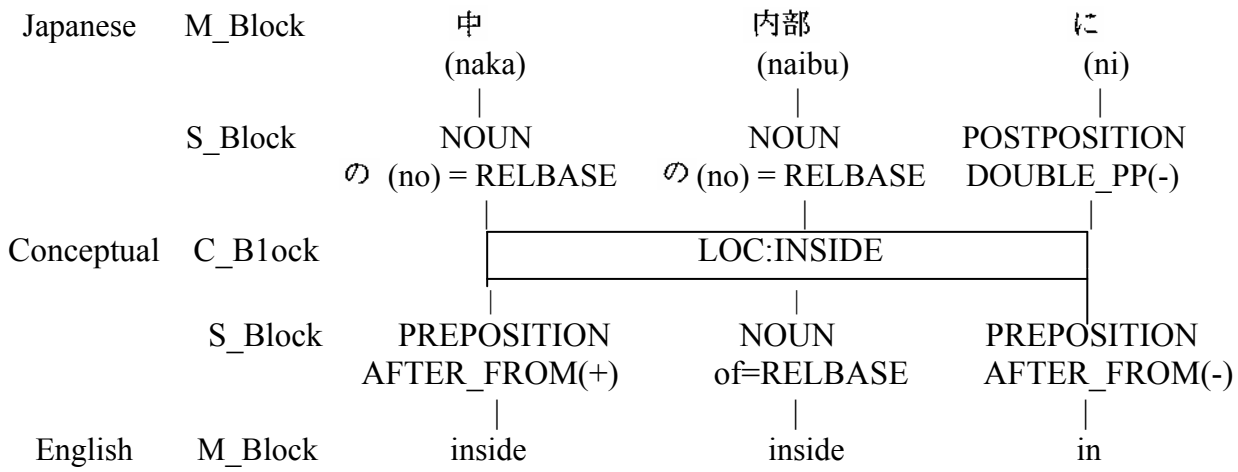
The sound came from inside the room.

The sound came from the inside of the room.

In this case, the interlingua for 'from inside' and 'from the inside of' is as follows.

CASE CP CASE
→ SOURCE → LOC:INSIDE → RELBASE →

In this graph, LOC:INSIDE expresses the concept that corresponds to the meaning of the English word 'inside.' Here, RELBASE is a CASE that expresses bases of relative concepts. The dictionary structure for LOC:INSIDE is as follows.



When the English noun 'inside' is selected for the English expression of the concept LOC:INSIDE, the English expression 'from the inside of' is obtained. That is because the English noun S_Block has information that shows this noun takes the article 'the' before it and the preposition 'of' after it. When the English preposition 'inside' is selected, the expression 'from inside' is obtained. That is because this English preposition S_Block has information that shows this preposition can be located just after the English preposition 'from.'

In the case of Japanese, only nouns 'naka' or 'naibu' can be selected for the Japanese expression for the concept LOC:INSIDE. These nouns can be connected to the Japanese postposition 'kara' and the expression 'naka kara' and 'naibu kara' are obtained. Since two postpositions cannot be connected in Japanese, the postposition 'ni' can not be selected for the expression of the concept LOC:INSIDE.

Let us take the following example.

He is standing outside the room.

In this case, the following two equivalent conceptual graphs are produced for the interlingua.

```
CP          CASE          CP                CASE          CP
STAND  →    LOC     →   LOC:OUTSIDE   →   RELBASE   →   ROOM

            CP          CASE          CP
STAND   →   LOC:OUTSIDE   →   ROOM
```

In other words, the following relation is established between the two equivalent conceptual graphs.

```
   CASE          CP              CASE
→  LOC  → LOC:SUBFEA → RELBASE →
             |   |
           CASE
       → LOC:SUBFEA →
```

Here, SUBFEA indicates the subcategory of CASE. Here, INSIDE and OUTSIDE are subcategories of LOC, which means Location. Let us think about the English sentence 'The baby is in the room.' The conceptual structure of this sentence may be expressed as follows:

```
                          *
BABY  ←  OBJ  ←  EXIST  →  LOC  →  LOC:INSIDE → RELBASE → ROOM
```

On the other hand, the conceptual structure for the English noun phrase 'the baby in the room' may be expressed as:

```
      *
BABY  →  LOC:INSIDE  →  ROOM
```

That is to say, with regard to the relationship between the sentence 'The baby is in the room' and the noun phrase 'the baby in the room,' the following relationship may be said to hold:

```
    CASE      CP       CASE          CP            CASE
←   OBJ  ← EXIST  →   LOC   →  LOC:SUBFEA → RELBASE →
                        |   |
                      CASE
                →   LOC:SUBFEA  →
```

Whether in a sentence or in a noun phrase, the two conceptual elements (here BABY and ROOM) will bear the same relationship to one another: the difference that exists

between the sentence and the noun phrase is simply indicated by the differing positions of their respective top nodes.

With this kind of situation in mind, we have made our interlingua capable of producing equivalent conceptual graph structures for the same concept and of executing transformations among such equivalent conceptual structures at the concept level, independent of natural languages.

3.2 Top-node transformation of Interlingua

3.3 Expression of Propositional Attitudes

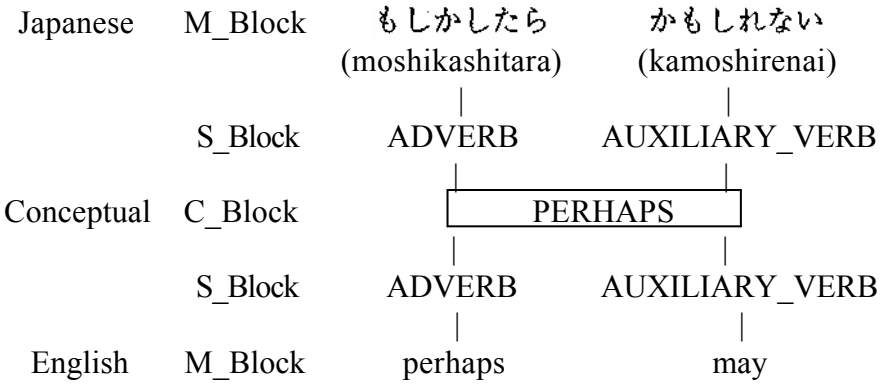
Sentences in general include propositions and propositional attitudes. For example, in the case of the sentence 'He may have come,' the proposition is 'he came,' and with the auxiliary verb 'may,' the speaker is expressing a judgment or propositional attitude, regarding a degree of reliability for the proposition. In the interlingua, propositional attitudes are expressed by features added to predicative nodes, as shown below:

```
HE ← OBJECT ← COME
                #TENSE(PAST)
                #ATTITUDE(PERHAPS)
```

Almost the same meaning as the sentence above can be expressed in a different way. In the sentence 'Perhaps he came,' the propositional attitude is expressed by the adverb. The linguistic phenomenon that propositional attitudes are expressed by the auxiliary verbs and adverbs is also observed in Japanese. Therefore interlingua and dictionary have to be defined in order to enable the following four translations in both directions.

Perhaps he came.	He may have come.
↑ ↓	
もしかしたら彼は来た。	彼は来たかも知れない。
moshikashitara(=perhaps)	
kare(= he) wa(TOPIC) kita(= came)	kare wa kita kamoshirenai(= may)

In order to handle these translations, we have built the following dictionary for the concept 'PERHAPS'



In addition, we define the following two equivalent conceptual graphs.

```
                          *
HE ← OBJECT ←     COME
                  #TENSE(PAST)
                  #ATTITUDE(PERHAPS)

                          *
HE ← OBJECT ←     COME                    ← OBJECT ← PERHAPS
                  #TENSE(PAST)
```

Here, then, propositional attitudes are expressed in two ways: (1) as features on conceptual nodes, and (2) as separate conceptual nodes. In addition, these two expressions are equivalent and may be transformed from one to the other.

What we would like to emphasize here is that such transformations among equivalent conceptual structures are independent of source and target languages. The transformations are defined for sets of conceptual graphs. This is a fundamental difference between the interlingua method the ST method. Conceptual transformation in interlingua eases the translation of varying styles of sentences.

4 Conceptual Primitives

4.1 Grammar of Interlingua

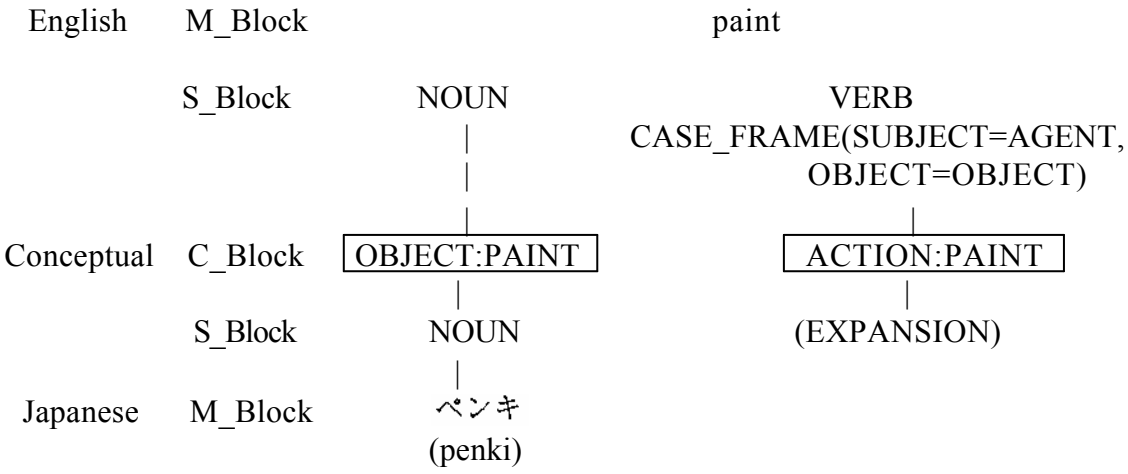
We have divided conceptual primitives into two classes; (1) functional conceptual primitives, and (2) general conceptual primitives. Functional conceptual primitives consist of CASE, TENSE/ASPECT, propositional attitudes, etc., and may be expressed in the form of prepositions, conjunctions, auxiliary verbs, and so on. Functional conceptual primitives are held in common by both languages being worked with and are independent of the direction of translation. Along with rules for interpreting conceptual graphs, they constitute the 'grammar' of the interlingua. While the number of possible functional conceptual primitives is potentially limitless, we have confined their number to a reasonably workable size. At the end of this paper, we give a list of, for example, the thirty-nine (39) CASE primitives that we have defined. The total number of their functional conceptual primitives we have defined, i.e. primitives for TENSE/ASPECT, propositional attitudes, modality, etc., is and seventy-three (73).

The difficulty of defining basic conceptual primitives is well known. We developed our set of functional conceptual primitives in a series of steps, gradually refining it as we built large size dictionaries (about a hundred thousand (100,000) entries each for English and Japanese) and English-to-Japanese and Japanese-to-English MT systems for handling as wide a range of sentence types as possible, and then actually translating many different types of documents. We also tested this same set of functional conceptual primitives in the translation of basic sentences in French, Spanish, and Korean, and found it to be generally valid for these languages as well.

4.2 Vocabulary of Interlingua

By way of contrast, general conceptual primitives, the 'Vocabulary' of the interlingua, may be specific to particular languages and dependent on the direction of translation. New general conceptual primitives can be added according to words in specific languages. In order to treat a concept that a specific language has and that does not exist in other languages, we introduce paraphrasing rules in the conceptual block in the dictionary.

For example, the English word 'paint'is used as a verb, but there is no verb in Japanese which directly correspond to it. Hence we define the following structure in order to translate the English verb into a Japanese verb phrase.



Here,

```
EXPANSION:    ACTION:PAINT = [ACTION:NURU] → OBJECT → [OBJECT:PAINT]
```

Consequently the following translation is obtained.

English He will paint a wall.
Japanese 彼は壁にペンキをぬる
Kare(= he) wa(TOPIC) kabe(= wall) ni(INDIRECT OBJECT)
penki(= paint) wo(OBJECT) nuru (= put, paint)

Our interlingua utilizes functional conceptual primitives, which are finite, and have general conceptual primitive, which are infinite. This means this approach is different from other approaches such as lexical decomposition and lexical exhaustive listing (Schank and Rieger 74, Okada and Tamachi 73a, 73b).

5 Conclusion

In this paper we described characteristics of the interlingua we developed for and utilized in actual Japanese-to-English and English-to-Japanese MT commercial systems. The structure of our interlingua is basically a directed and acyclic graph in which two types of conceptual primitives, CP and CASE, are arranged alternatively. The main characteristics of the interlingua are as follows.

(1) Conceptual primitives may be linked to any parts of speech in specific languages.
(2) Positions of the top node on the graph correspond to the differences of syntactic

structures. (3) The same concept can be expressed by two or more equivalent conceptual graphs. These graphs can be converted to one another by conceptual transformation rules, which are independent of specific languages. (4) Conceptual primitives are divided into two classes; (a) functional conceptual primitives, which are finite and manageable and independent of any specific languages, and (b) general conceptual primitives, which can be added according to words in specific languages, and sometimes depend on the translation directions. Functional conceptual primitives and interpreting rules of the conceptual graphs constitute the grammar of the interlingua.

It has been said that the difference between the ST method and interlingua method is that the former has syntactic transfer rules that connect varieties of sentence styles in a source language to appropriate sentence styles in a target language, and the latter does not have such rules. However, the conceptual devices described above of our interlingua enables us to achieve the same quality as the ST method, while holding the grammar of interlingua which are common in any languages.

We improved and fixed the interlingua by developing large lexicon and actual MT systems, translating large volumes of actual documents. Our commercial MT products using the interlingua produce results of roughly the same or higher quality than systems using the ST method. We have also confirmed the validity of the interlingua for the basic sentence translations in French, Spanish, and Korean (Okumura 91). These facts indicates the feasibility of the interlingua approach.

References

[1] Machine Translation System Laboratory, Center of the International Cooperation for Computeization (CICC). 1995. *The CICC Interlingua.* Final Edition.

[2] Japan Electronic Dictionary Research Institute (EDR). 1995. *Specification of EDR Electronic Dictionary.* Second Edition.

[3] Roger C. Schank and Charles J. Rieger III. 1974. Inference and the Computer Understanding of Natural Language. *Artificial Intelligence,* 5.

[4] Naoyuki Okada and Tsuneo Tamachi 1973. An Analysis and Classification of "Simple Matter Concepts" for Natural Language and Picture Interpretation, *the Transactions of the Institute of Electronics and Communication Engineers of Japan* D.

[5] Naoyuki Okada and Tsuneo Tamachi 1973. An Analysis and Classification of "Non-Simple Matter Concepts" for Natural Language and Picture Interpretation, *the Transactions of the Institute of Electronics and Communication Engineers of Japan* D.

[6] Kazunori Muraki. 1984. A Japanese to English Machine Translation System using Knowledgebase and Language-independent Interlingua, (in Japanese) *Nikkei Electronics.*

[7] Kazunori Muraki. 1986. Augmented Dependency Grammar for Language Comprehension. In *Proceedings of the Second European AI Conference.*

[8] Akitoshi Okumura, Kazunori Muraki, and Susumu Akamine. 1991. Multi-lingual Sentence Generation from the PIVOT Interlingua. In *Proceedings of Machine Translation Summit III.*

[9] David Farwell and Yorick Wilks. 1991. ULTRA: A Multilingual Machine Translator. In *Proceedings of Machine Translation Summit III.*

Table 1: List of Relational Conceptual Primitives (CASEs)

No.	Symbol	Brief Description	Example Sentence
1	OBJ	object of a predicate	John hit [the desk] with his fist.
2	AGT	agent of an action	[He] gave me his books.
3	CAU	causer	[She] let him leave.
4	EXP	experiancer of feeling and sense	[They] suspect that he is the murderer.
5	INS	instrument, tool	[The computer] solved the problem.
6	MEA	means, method	She persuaded him to stay [with a kiss].
7	BEN	beneficiary	They are working [for me].
8	LOC	location	I saw it [under the table].
9	TIM	time	We haven't seen land [in 20 days].
10	SOR	source, starting point	John left [town].
11	TAR	target, destination point	John spent all his money [on clothes].
12	PRT	accompanied thing	They married Taro [to Hanako].
13	CAP	role, function	He attended meeting [as a leader].
14	FCS	focus	He wrote a book [about Japan].
15	MAT	material	Their house is built of [wood].
16	ELM	element	Japan consists of [four islands].
17	POS	possessor	[This] is my house.
18	POF	part of the whole	[The front] of the car was destroyed.
19	NUM	number	[five] meters
20	NAM	name	The [Yamanote] line
21	NMOD	succession of nouns	a [magnetic] disc
22	ATT	attribute	What is your shoe [size] ?
23	VAL	value of attribute	The length of the axis is [30 meters].
24	MOD	adverbial modifier	People struggle to live [better].
25	QUNT	quantity	She bought [two] bedding sets.
26	EQ	equal relationship	That must be a whale.
27	APP	apposition	memory devices, [such as magnetic discs]
28	REF	determination of reference	[Which] man did John say saw him?
29	CPQT	quantity of comparative	He is [5 cm] taller than I.
30	MDQT	quantity of change and continuation	He walks [5 km].
31	RLBS	base of relative concept	He looks at the outside of [the window].
32	DCMP	'than' of comparative	He is taller [than] I
33	MODS	sentential modifier	He is, [in a word], an idiot.
34	PAR	parallel	John [and Mary] heard the news.
35	GOA	goal, aim, purposive	They stopped [in order to rest].
36	CON	connection of affairs	[Returning] to my office, I slept.
37	REA	reason	[The rain] made us seek shelter.
38	CAS	establishment of case	[In case of the accident], call me.
39	CASA	assumption	[If he comes], call me.