KOD – Intermediate Representation for MT

(new-revised)

Feng Zhiwei

(Institute of Applied Linguistics, Ministry of Education of China)

KOD system is a communication system for information exchange. So KOD is a means of communication based on all existing languages, it is a means of communication that respects and preserves each language in its uniqueness, and it is a means of communication that sees all languages participating in the system on an equal level in the process of global communication. So KOD is very useful for social communication. In the same time, I believe, KOD can also be used as intermediate representation for Machine Translation (MT) in order to overcome the language barrier.

1. Intermediate Language in MT

The MT pioneer, Russian engineer P. P. Troyansky pointed out (1933) that we can use the intermediate language to favorite the machine translation. For example,

German: Das Bild der Welt zeigt, wie die Materie sich bewegt, wie die Materie denkt.
         o   de as   --   o   as   --   o   as

French: Le tableau du monde montre, comment la matière se meut, comment la matière pense.
         o   de   as   --   o   as   --   o   as

Russian: Картина мира показывает, как материя движется, как материя мыслит.
         o   de   as   --   o   as   --   o   as

The logical symbols "o, de, as, --" can be regarded as intermediate language, its grammatical functions respectively as follows:

   o: subject, noun, singular, nominative case.
   as: predicate, verb, present tense, declarative mood.
   de: noun with genitive case, or preposition representing the possessive relation.
   --: relative words.

All languages share these common logical structures, so we can use them as the intermediate language in machine translation.

The American scientist W. Weaver pointed out (1946) in his <*memorandum of machine translation*>: The common basis of human communication is perhaps an universal language, but now we have not yet find it.

In the first machine translation conference, the American scientist E. Reifler pointed out (1952): It is necessary to use intermediate language in machine translation.

KOD is a means of communication based on all existing languages, so it is suitable also for machine translation. However, KOD is not a language, because of the fact that KOD does not have its own grammar, KOD bases on the grammar of the different source languages. Therefore it would be more appropriate to name the KOD communication system 'the intermediate representation' as a phenomenon in between. So it can be used as intermediate representation in MT.

2. KOD as intermediate representation in MT

Let me give an example of Chinese-English MT to explain my idea.

In the case of using KOD as intermediate representation, the process of Chinese-English machine translation will be as follows:

-- Analysis of Chinese:

The Chinese sentence is: "人类语言是一种无比丰富的现象。" (renlei yuyan shi yi zhong wubi fengfu de xianxiang.).

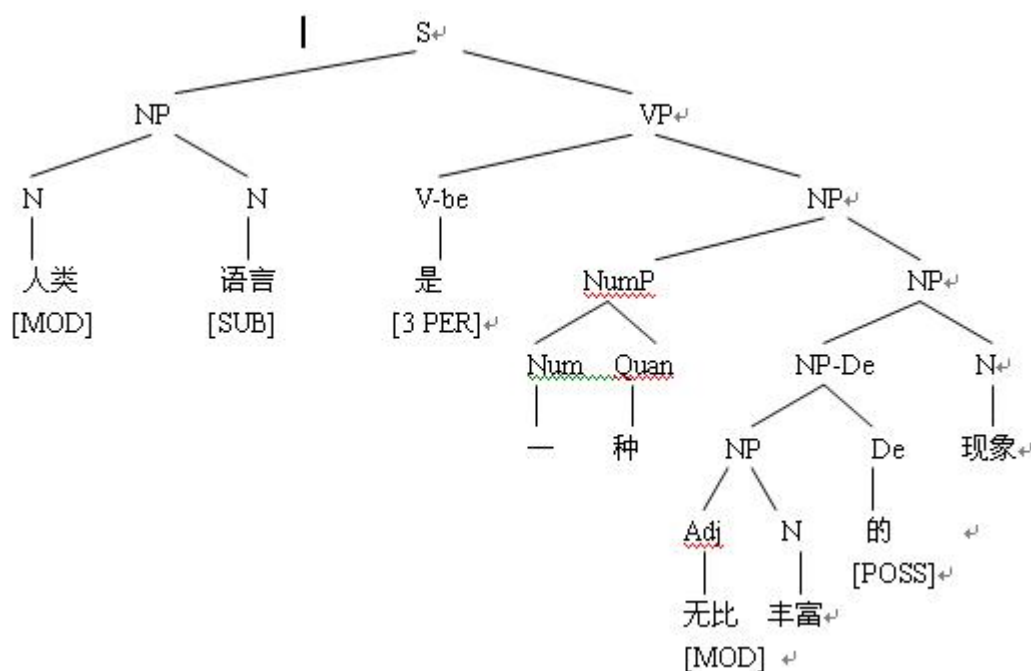After the parsing of this Chinese sentence, we can get following tree graph structure

Fig 1. Chinese Analysis

-- Lexical transfer: We set up a Chinese-KOD dictionary:

Chinese	KOD
人类	ninik
语言	tokim
是	put
一	tes
种	ado

无比	hipermeg
丰富	nemim
的	[POSS]
现象	fenomen

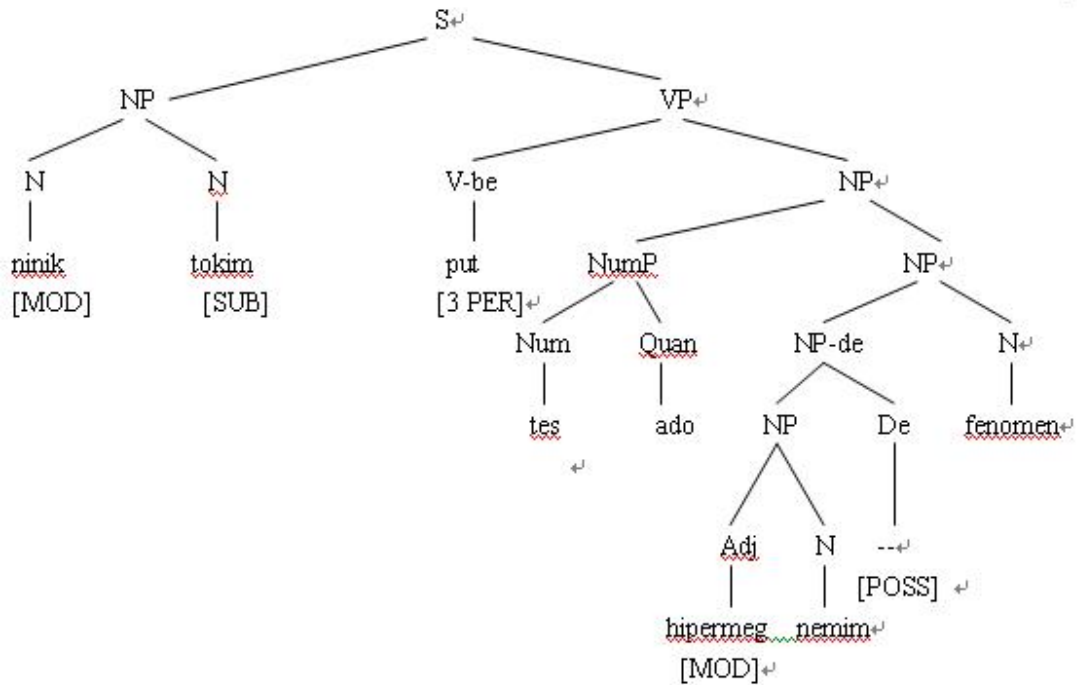Then, we can get the tree graph with KOD words and MT grammatical feature tags:

Fig 2. Lexical Transfer to KOD words

-- Grammar marker transfer from MT grammatical feature tag to KOD grammar marker:

MT tag	KOD marker
[MOD]	[ube]
[SUB]	[ora]
[3 PER]	[ane]
[POSS]	[oro]

-- KOD- lingua-Chinese Intermediate representation is as follows:

```
                                    S
                   ┌────────────────┴────────────────┐
                  NP                                  VP
             ┌─────┴─────┐                    ┌───────┴───────┐
             N           N                  V-be              NP
             │           │                    │        ┌──────┴──────┐
          ninik        tokim                 put     NumP            NP
          [ube]        [ora]                 [ane] ┌──┴──┐      ┌─────┴─────┐
                                                  Num   Quan  NP-de         N
                                                   │     │   ┌──┴──┐        │
                                                  tes   ado NP    De      fenomen
                                                        ┌──┴──┐    │
                                                       Adj    N   --
                                                        │     │  [oro]
                                                     hipermeg nemim
                                                       [ube]
```
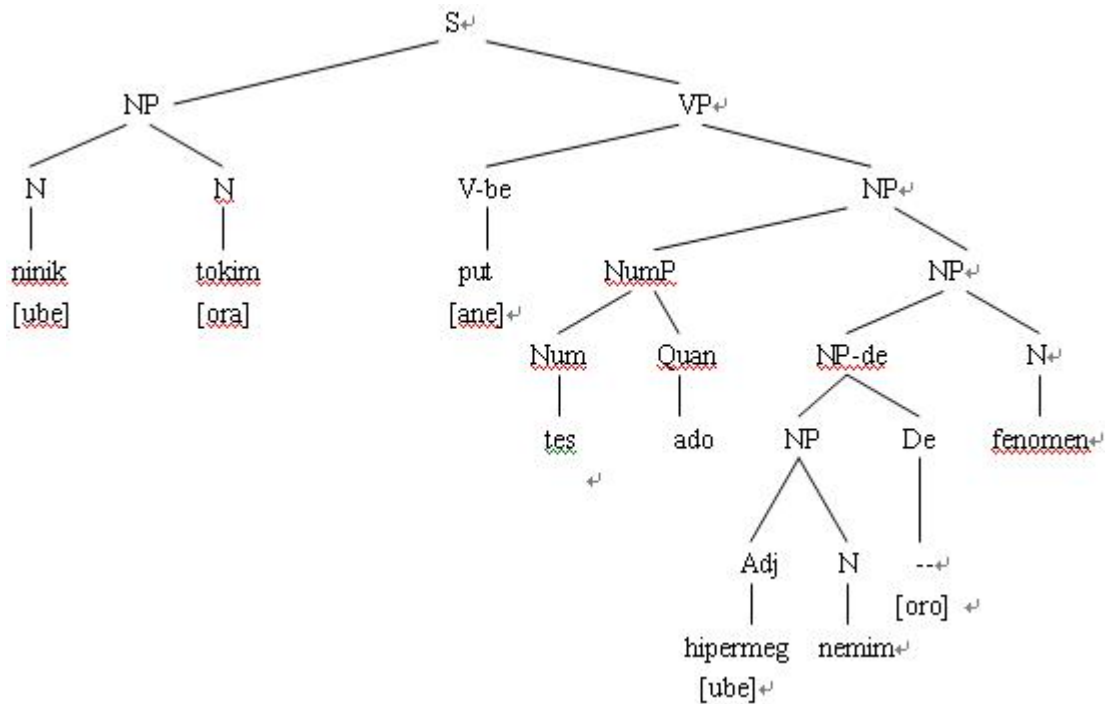
Fig 3. Grammar marker transfer

We see clear, this tree graph is the structure representation of following KOD-lingua Chinese sentence:

"ninikube tokimora putane tes ado hipermegube nemim oro fenomen".

So the structure of KOD-lingua Chinese sentence can be regarded as the intermediate representation of MT, it can be easily transformed to the equivalent structure of KOD-lingua English sentence. The transfer operation will be reduced to the minimum. In this case, we can say that KOD is the intermediate representation for Chinese-English machine translation. The KOD-lingua English structure is the representation of English in KOD. By this reason, we can transfer the KOD-lingua English structure to authentic English structure and then generate equivalent English surface sentence.

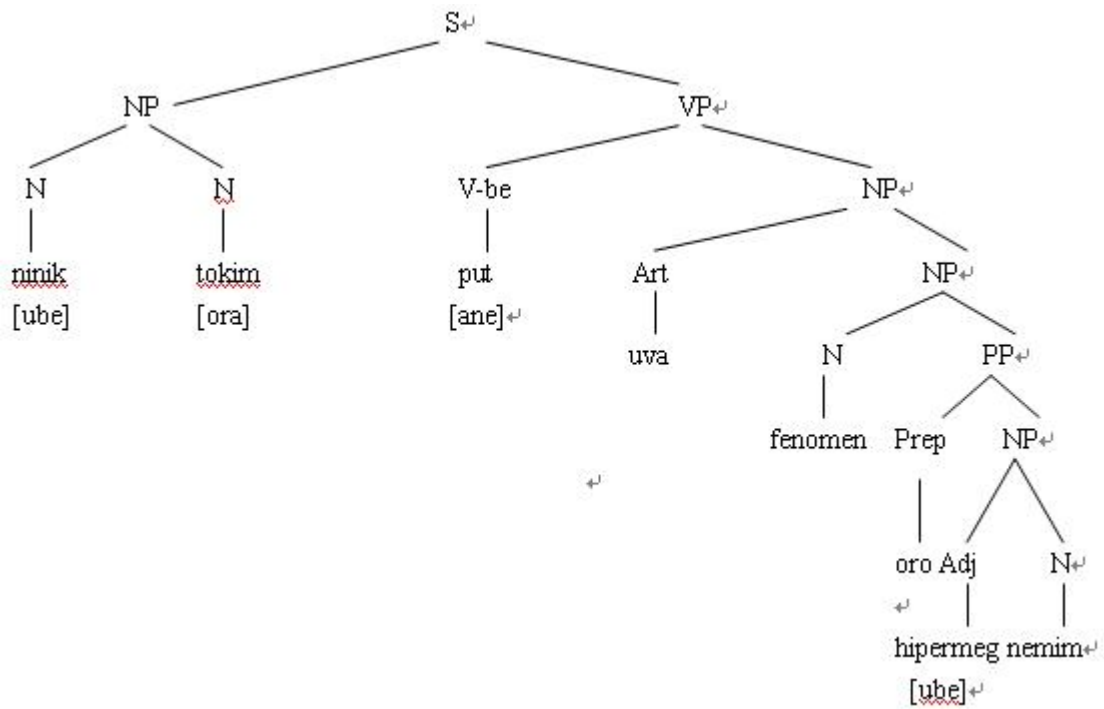-- Transfer from KOD-lingua Chinese structure to KOD-lingua English structure::

```
                                  S
                 ┌────────────────┴──────────────────┐
                NP                                   VP
            ┌────┴────┐                      ┌────────┴────────┐
            N         N                     V-be              NP
            │         │                      │          ┌──────┴──────┐
          ninik     tokim                   put        Art           NP
          [ube]     [ora]                  [ane]        │       ┌─────┴─────┐
                                                       uva      N          PP
                                                             fenomen   ┌───┴───┐
                                                                      Prep    NP
                                                                       │    ┌──┴──┐
                                                                      oro  Adj    N
                                                                        hipermeg nemim
                                                                          [ube]
```

Fig 4. KOD-lingua Chinese to KOD-lingua English

-- Lexical generation of English:

We set up a KOD-English Dictionary:

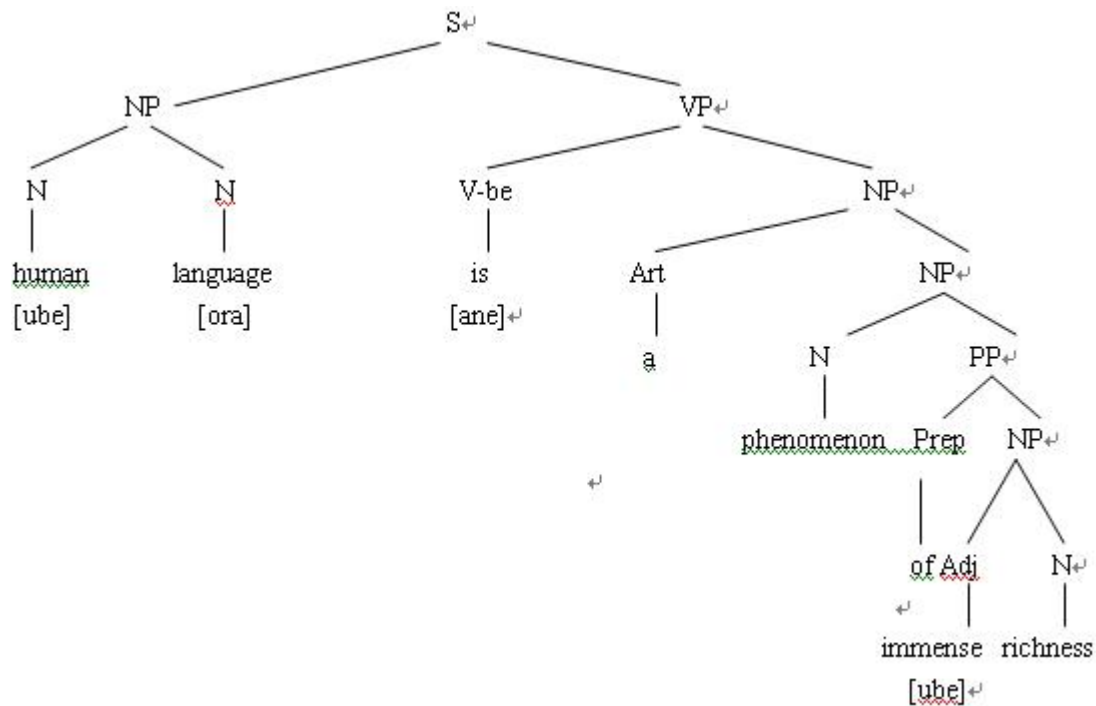| KOD	English
ninik[ube]	human
tokim[ora]	language
put[ane]	is
uva	a
fenomen	phenomenon
oro	of
hiopermeg[ube]	immense
nemim	richness

-- Structure generation of English:

Fig 5. Structure generation

Then we can get the English translation as follows:

"Human language is a phenomenon of immense richness."

Obviously, if we adapt KOD as the intermediate representation for MT, Chinese-English MT procedure will be as follows:
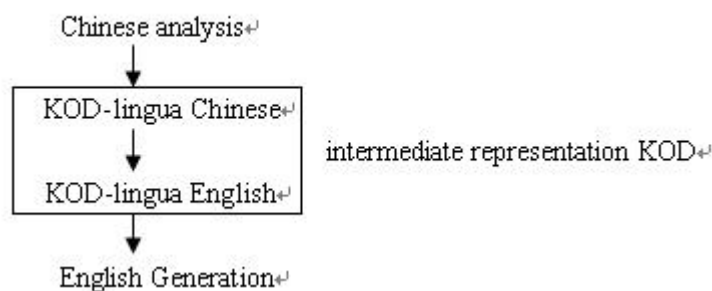

Fig 6. Chinese-English MT procedure

In the level of intermediate representation KOD, the transfer from KOD-lingua Chinese to KOD-lingua English is necessary, but it is very easy. So KOD will be a favorite representation tool for MT.

In order to use KOD as the intermediate representation, I suggest Dr. Johann Vielberth to add the "KOD-transfer" into KOD communication system. The function of KOD-transfer is to transfer KOD-lingua source language to KOD-lingua target language. In this case, the KOD communication shall include three components: KOD-global, KOD-lingua and KOD-transfer.

3. Advantage of KOD for MT

In the condition of application of KOD as the intermediate representation, the process of machine translation will be as follows:

Source language → intermediate representation KOD

Intermediate representation KOD → target language

In the intermediate representation KOD, the transfer from KOD-lingua source language to KOD-lingua target language is very easy, if we apply the KOD communication technique.

Due to the application of KOD as the intermediate representation, the number of MT systems will be reduced.

Many MT systems adopt the strategy of dependent analysis and independent generation. It means that the analysis of source language is dependent to target language, but the generation of target language is independent to source language.

For MT systems which adopt the strategy of dependent analysis and independent generation, if we have N languages in MT, then we shall set up N(N-1) analysis systems and N generation systems, the total number of systems α will be:

$$\alpha = N(N-1)+N = N^2-N+N = N^2$$

For example, if we have 4 languages in the MT system, then the total number of systems will be:

$$\alpha = N^2 = 4^2 = 16$$

If we have 4 languages A, B, C and D, the systems will be as follows:
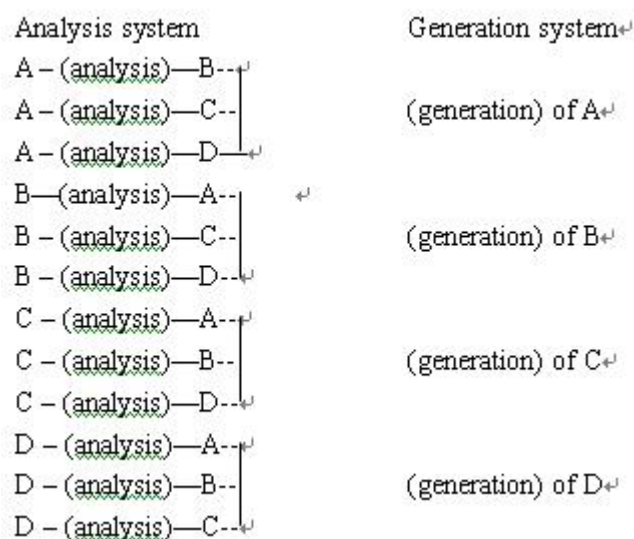

Fig 7. system number of MT for 4 languages

If we adopt KOD as the intermediate representation, for every language, we need only set up two systems (one for analysis, another one for generation), so the total number of systems β will be:

$$\beta = 2N$$

For example, if we have 4 languages A, B, C and D, the number of systems will be:

$$\beta = 2N = 2 \times 4 = 8,$$
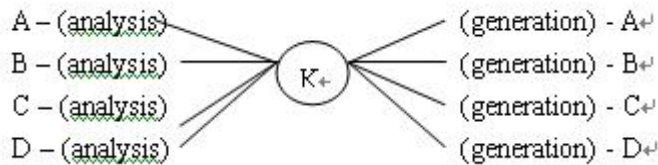
It means that we need only 8 systems:

Fig 8. system number of MT used KOD

Here "K" means intermediate representation KOD.
The ratio R of α and β is:

$$R = \frac{\alpha}{\beta} = \frac{N^2}{2N} = \frac{N}{2}$$

If N = 4, then

$$R = \frac{N}{2} = \frac{4}{2} = 2$$

If N = 2, then R = 1. If N>2, then R>1. The more is N, the more becomes R It means the number of MT systems will be largely reduced. Let us read following table:

N	2	3	4	5	6	7	8	9	10	20	30	
A	4	9	16	25	36	49	64	81	100	400	900	
B	4	6	8	10	12	14	16	18	20	40	60	
R	1	1.5	2	2.5	3	3.5	4	4.5	5	10	15	

Fig 9.   comparison table

Obviously, if we adopt KOD as intermediate representation of MT, we shall save a lot of labors and funds in MT development. It is the advantage of KOD.

References

[1] Feng Zhiwei, Research of Machine Translation, China Translation and Publication Company, 2004, Beijing.

[2] Feng Zhiwei, PAR-based Japanese-Chinese MT, China-Japan NLP Symposium, 2002, October, p. 104-117, Beijing.

[3] Trujillo Artuno, Translation Engines: Technique for Machine Translation, Springer, 1999.