TRANSFER IN A MULTILINGUAL MT SYSTEM*

Steven Krauwer & Louis des Tombe
Institute for General Linguistics
Utrecht State University
Trans 14, 3512 JK Utrecht, The Netherlands

ABSTRACT

In the context of transferbased MT systems, the nature of the intermediate represenations, and particularly their 'depth', is an important question. This paper explores the notions of 'independence of languages' and 'simple transfer', and provides some principles that may enable linguists to study this problem in a systematic way.

1. Background

This paper is relevant for a class of MT systems with the following characteristics:

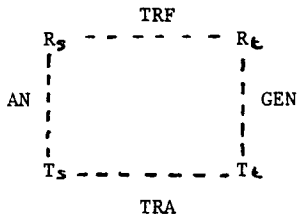
(i)
The translation process is broken down into three stages: source text analysis, transfer, and target text synthesis.
(ii)
The text that serves as unit of translation is at least a sentence.
(iii)
The system is multilingual, at least in principle.

These characteristics are not uncommon; however, Eurotra may be the only project in the world that applies (iii) not only as a matter of principle but as actual practice.

We will regard a natural language as a set of texts. A translation pair is a pair of texts (T_s, T_t) from the source and target language, respectively. One sometimes wonders whether for every T_s there is at least one translation T_t, but we will ignore that kind of issue here.

For translation systems of the analysis-transfer-synthesis family, the following diagram is a useful description:

(i)



TRA, AN, TRF, and GEN are all binary relations. Given the sets of texts SL (source language) and TL (target language), and the set of representations RR, we can say:

$TRA \subseteq SL \times TL$, $AN \subseteq SL \times RR$
$TRF \subseteq RR \times RR$, and $GEN \subseteq RR \times TL$

The subsystems analysis, transfer, and synthesis are implementations of AN, TRF, and GEN. In this paper, we are not interested in the implementations, but in the relations to be implemented.
Especially, we try to find a principled basis for the study of the represenations R and R . Such a basis can only be established in the context of some fundamental philosophy of the translation system. We will assume the following two basic ideas:

(i)
Simple transfer:
Transfer should be kept as simple as possible.

(ii)
Independence of languages:
The construction of analysis and synthesis for each of the languages should be entirely independent of knowledge about the other languages covered.

These two ideas are certainly not trivial, and especially (ii) may be a bit exceptional compared to other MT projects; however, they are quite reasonable given a project that really tries to develop a multilingual translation system. In any case, they are both held in the Eurotra project.

---

*The research described here was done in the context of the Eurotra project; we are grateful to all the Eurotrans for their stimulation and help.

The reason for (i) is simply the number of transfer systems that must be developed for k languages, which is

$$k(k-1).$$

From this, it follows that 'simple' here means 'simple to construct', not 'simple to execute'.

The reason for principle (ii) also follows for multilinguality; while developing analysis and synthesis for some language, one may be able to take into account two or three other languages, but this does not hold in a case like Eurotra, where one not only has seven languages to deal with, but also the possibility of adding languages must be kept open.

Principles (i) and (ii) together constitute a philosophy that can serve as a basis for the development of a theory about the nature of the representations R_s and R_t in (1). The remainder of this paper is devoted to a clearer and more useful formulation of them.

2. Division of labour.

Suppose that simple transfer is taken to mean that transfer will only substitute lexical elements, and that the theory of representation says that the representations are something in the way of syntactic structures. We now have a problem in cases where translation pairs consist of texts with different syntactic structures. Two well-known examples are:

(i) the graag-like case;

Example: Dutch 'Tom zwemt graag' translates as English 'Tom likes to swim', with syntactic structures:

(2) Dutch:
[$_S$ Tom [$_{VP}$ zwem [$_{Adv}$ graag]]]

(3) English:
[$_S$ Tom [$_{VP}$ like [$_S$ empty [$_{VP}$ swim]]]]

In the case of Dutch-English transfer, lexical substitution would result in an R_t like the following:

(4) Possible R :
[$_S$ Tom [$_{VP}$ swim [$_{AdvP}$ like-to]]]

In this way, the pair ⟨(4), 'Tom likes to swim'⟩ becomes a member of the relation GEN for English. However, it is hard to believe that English linguists will be able to accomodate such pairs without knowing a lot about the other languages that belong to the project.

(ii) The kenner - somebody who knows case

Dutch and English both have agentive derivation, like

talk ⇒ talker, swim ⇒ swimmer.

However, as usually, derivational processes are not entirely regular, and so, for example though Dutch has 'kenner', English does not have the corresponding 'knower'. So we have the following translation pair:

(5) Dutch: 'kenner van het Turks'
English: 'somebody who knows Turkish'

Again, the English generation writer is in trouble if he has to know that the R_t may contain a construction like '[[know]+er]', because this implies knowledge about all the other languages that participate.

The general idea is that we want to have a strictly monolingual basis for the development of the implementations of AN and GEN. Therefore, so, we have the following principle:

(6) Division of labour (simple version):
For each language L in the system,
$\langle R,T\rangle \in GEN_L$ iff $\langle T,R\rangle \in AN_L$

Principle (6) makes AN and GEN each others 'mirror image', and so it becomes more probable (though it is not guaranteed) that the linguists knowing L will understand the class of R_ts they can expect.

However, (6) is too strong, and may be in conflict with the idea of simple transfer. For example, if surface syntactic structure is taken as a theory of representation, then (6) implies that TRF relates source language surface word order to target language word order, which clearly involves a lot more than substitution of lexical elements.

Therefore, the notion of isoduidy has been developed. Isoduidy is an equivalence relation between representations that belong to the same language. Literally, the word 'isoduid' (from Greek and Dutch stems) means 'same interpretation'; but the meaning should be generalized to something like 'equivalent with respect to the essence of translation'.

To give an example, suppose that representations are surface trees with various labelings, including semantic ones like thematic relations and semantic markers. Isoduidy might then be defined loosely as follows:

two representations are isoduid if they have the same vertical geometry, and the same lexical elements and semantic labels in the corresponding positions.

Obviously, the definition of the contents of the isoduidy relation depends on the contents of the representation theory. However, we think that the general idea must be clear: isoduidy defines in some general way which aspects of representations are taken to be essential for translation.

Given isoduidy, one can give a more sophisticated version of the principle of division of labour as follows:

(7) Division of labour (final version):
For each language L in the system,
⟨R',T⟩ ∈ GEN_L
iff
⟨T,R⟩ ∈ AN_L and R' is isoduid to R

As a consequence, TRF has not to take responsibility for target language specific aspects like word order anymore.

3. Simple and complex transfer.

Given the principle of division of labour, we can relate to each other the following three things:

- the notion of simple transfer
- the representation theory, especially, the 'depth' of representation;
- the contents of the relation isoduidy

Given some definition of what counts as simple transfer, we can now see whether the representation theory is compatible with it.

It is easy to see that some popular theories of simple transfer, including the one saying that transfer is just substitution of lexical elements, will now give rise to a rather 'deep' theory of representation. This follows from cases like 'graag-like' and 'kenner-knower', where some language happens to lack lexical elements that others happen to have. In such cases, the language lacking the element usually circumscribes the meaning in some way. If one excludes transfer other than lexical substitution, such examples give rise to a theory of representation where similar circumscriptions must be assigned as representations in the language that does have the lexical element. So, in Dutch we get pairs in AN like

⟨'kenner', [somebody [who knows]]⟩
⟨'Tom zwemt graag', [Tom graag [empty zwem]]⟩

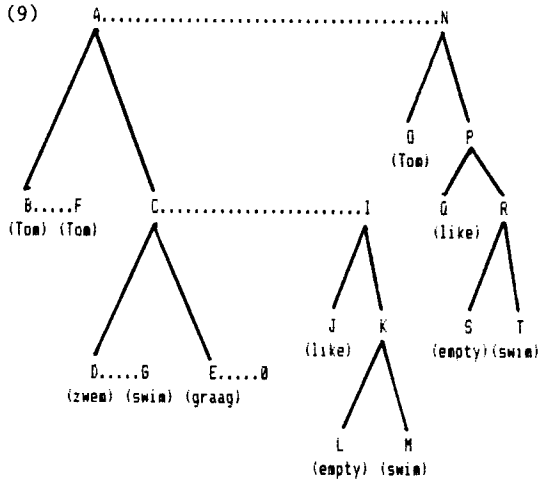
Instead of having deep representations like these, one may consider the possibility that transfer is complicated sometimes. So, one may still desire that transfer consists of just lexical substitution most of the time, but allow exceptions. The question then arises as to how simple and complex transfer interact.

As a basis for that, one may observe that the relation TRF now holds between representations, while in practice just lexical elements are translated most of the time. A straightfoward generalization is possible for the case where a representation is some hierarchical object, say some tree. We can then introduce a new relation, called translates-as. This is a binary relation, probably many-to-many; its left-hand term is a subtree of R_s, and its righthand term is a tree. Clearly, TRF is a subset of translates-as.

We then have the following principle:
(8) Transfer translates a tree node-by-node.

Note that, obviously, this only makes sense as long as we have representations that are trees.The following example may clarify the idea. Dotted lines indicate instantiations of the relation.

(9)



Note that Dutch 'graag' is not translated at all; it only serves as a basis for the complex transfer element ⟨C,I⟩.

The principle of simple transfer can now be formulated as follows:

If A translates-as A', then we will call A' a TN of A. We now call an element s,t of the set defined by translates-as a simple iff.

either
s and t are both terminal nodes,
or
(i) s is a subtree, dominated by the nonterminal node A, and
(ii) t is a tree, dominated by A', and
(iii) A' is a copy of A', and
(iv) the immediate daughters of A' are copies of the TNs of the immediate daughters of A.

The principle of simple transfer then says that the proportion of simple elements in translates-as must be maximal.

The generalised relation translates-as makes it possible to put some order into complex transfer. It localises it in a natural way, based on a tree structure.
In (9), only the pair ⟨C, I⟩ is complex; all the others are simple. This view on transfer is easily implemented by means of an inbuilt strategy that simulates recursion.

4. Conclusion.

The principle of division of labour, together with the principle of node-by-node transfer constitute a framework in which it is possible to study 'depth of representation' in a systematic way.