

How Close Can We Get to the Ideal of Simple Transfer in Multi-lingual Machine Translation (MT)?

Abstract

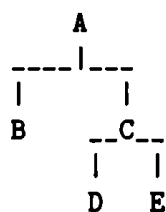
The ideal of simple transfer aims at restricting transfer rules to the exchange of unstructured lexical entities—the terminal leaves in the tree structure that is output from monolingual analysis. All information that is not lexicalised in the source language is represented as features to be transferred unchanged to the target language. In EUROTRA this ideal is approached through a centrally coordinated research within various phenomena which are supposed to be of translational relevance, i.e. having language specific surface manifestations. The outcome of this research ideally is to agree on a uniform treatment of these phenomena across languages, thus leading to simple transfer.

The paper makes a non-exhaustive overview over problems solved, problems under investigation, known but outstanding problems, and on this basis introduces a discussion of what will remain as unsolvable problems within an essentially sentence-based MT-system.

1 Introduction

MT-systems traditionally are classified into transfer-based systems and interlingual systems, as illustrated by figure 1 and figure 2, resp., on the next pages. The Interface Structure (IS) for some language is an annotated tree structure, where information is encoded as structure + features:

structure



features

```
B = {attribute 1 = value X,
      attribute 2 = value Y,
      ... }
```

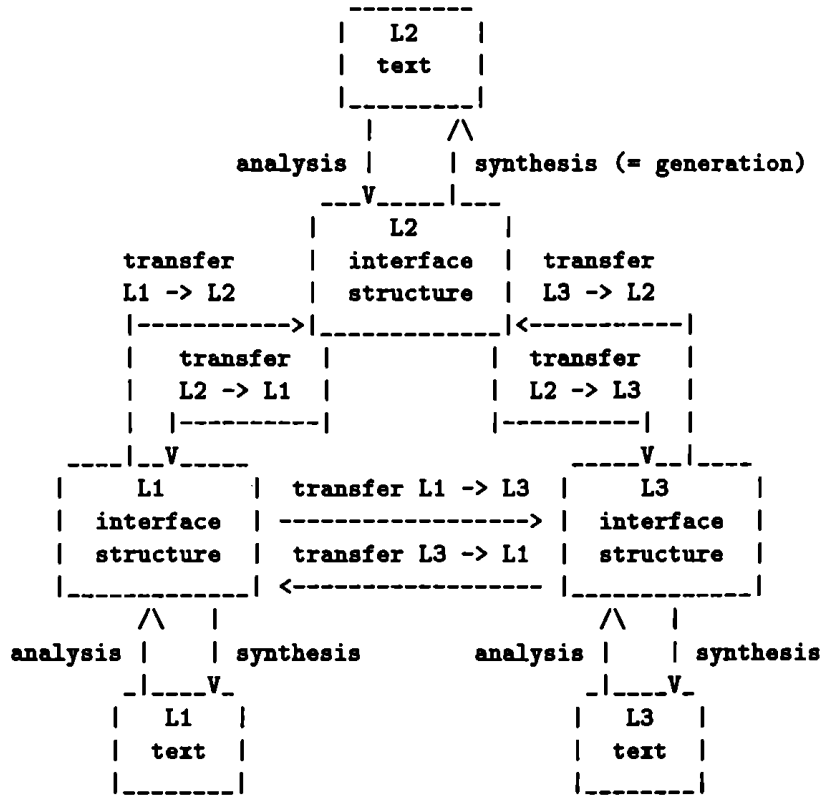


Figure 1: Schematic representation of transfer-based multi-lingual MT

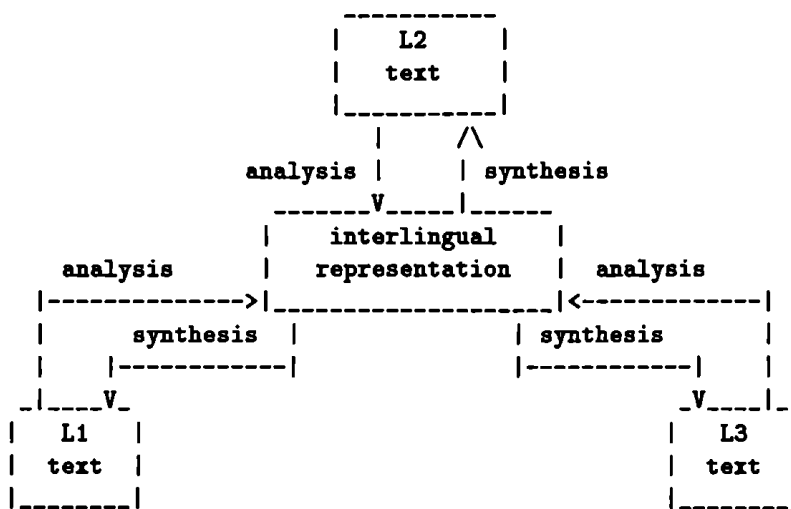


Figure 2: Schematic representation of interlingual multi-lingual MT

EUROTRA is conceived as a transfer-based system, which may seem less appropriate for an MT system comprising 9 languages in all combinations, thus leading to the construction of 72 transfer modules on top of 9 analysis and 9 synthesis modules, instead of just having one interlingua, 9 analysis and 9 synthesis modules.

What we want to show, is that the distinction between transfer- and interlingua-based systems should not be pushed too hard, especially if an interlingua is not perceived as a natural language-like representation but as any kind of information encoding that is neutral with respect to a source language and a target language.

The ideal in transfer is sometimes described as simple lexical transfer, which means that the lexical values are the only information in the interface structure that is not shared by source and target language and which consequently has to be changed by a transfer component, whereas all other information is represented language-independently in an interlingua. Actually, the greater part of lexical transfer may also be dispensed with through the inclusion of a comprehensive terminological component that is treated interlingually.

As the IS representation may be split up into structure information and feature information, we shall treat these independently and distinguish between

1. Transfer of structure

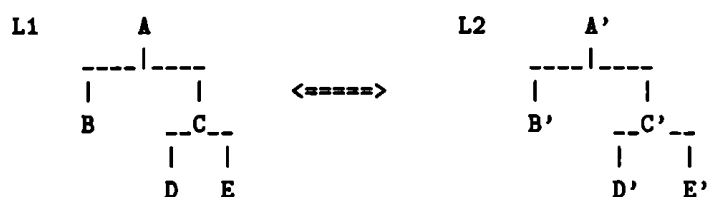
and

2. Transfer of features

2 Transfer of Structure

Here we distinguish between three possibilities:

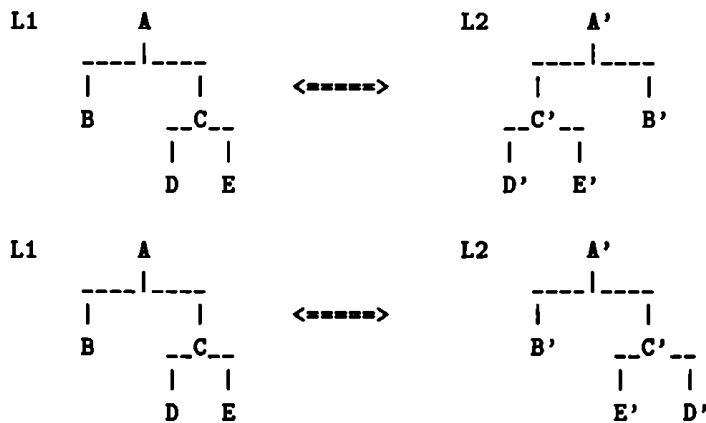
2.1 Simple transfer = interlingua (= no explicit transfer)



2.2 Deletion/insertion of node



2.3 Reordering of elements



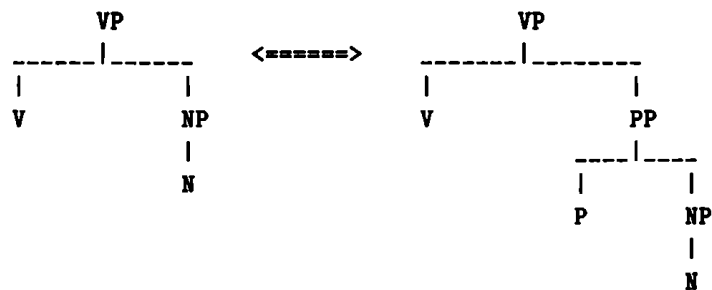
2.1 Simple Transfer

This is the unproblematic case where there is isomorphy between source language and target language or where this isomorphy is achieved between output from source language analysis and input to target language synthesis. How this isomorphy is achieved, is described in 2.2 and 2.3 below.

2.2 Deletion/Insertion of Node

2.2.1 Direct/Indirect Government

We have to delete, respectively insert, a node in cases where we have direct government by a verb of a noun phrase in one language corresponding to indirect government through a preposition in another language, e.g.



EN : (He) trusted her DA : (Han) stolede på hende
 DA : (Han) betragtede hende EN : (He) looked at her

The solution is to featurise all valency bound prepositions, without regard to whether they have a correspondence or not in one or more other languages, and delete the preposition and the PP-node from the IS representation:

a node in a structure than to modify information that has been featurised. As two examples of surface expressions that might be featurised—and actually were featurised, but now must be present as nodes in the IS representation—we may mention modal verbs and demonstrative pronouns.

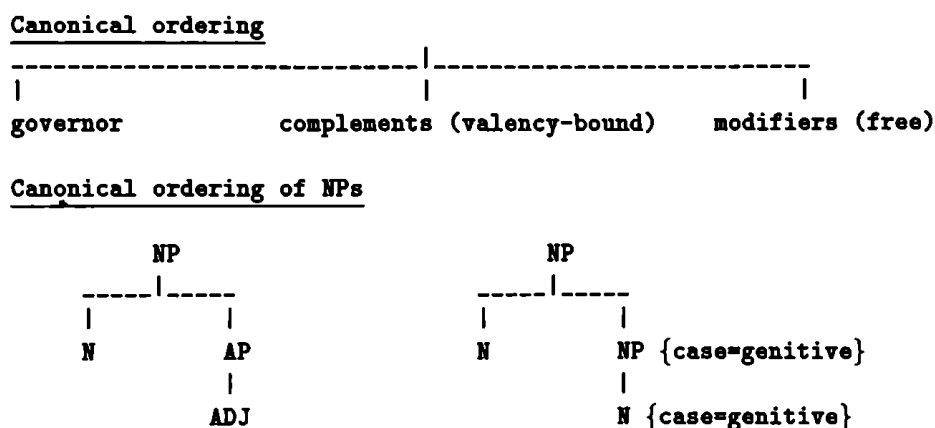
2.3 Reordering of Elements

2.3.1 Reordering at NP Level

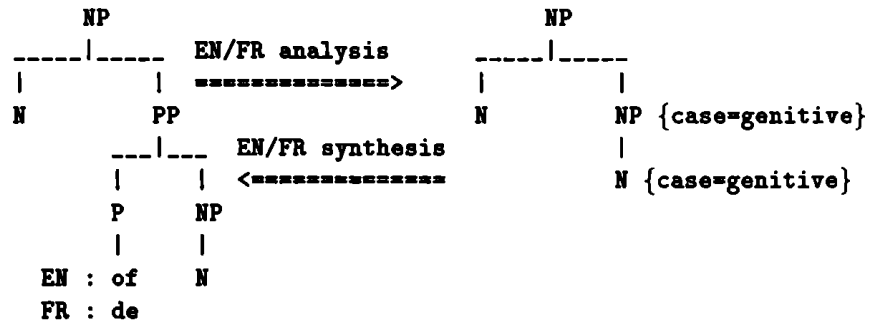
Reordering of elements occurs at NP level, where a modifier may precede the noun or follow after the noun, and where the ordering in different languages also differ according to the category of the modifier:

<u>Adjective + Noun</u>	<====>	<u>Noun + Adjective</u>
DA : den blå himmel		
EN : the blue sky		FR : le ciel bleu
DE : der blaue Himmel		
but		
FR : la petite fille		
<u>NP modifier + Noun</u>	<=====>	<u>Noun + NP/PP modifier</u>
DA : landets indbyggere		EN : the inhabitants of the country
		DE : die Einwohner des Landes
		FR : les habitants du pays

The solution is to have a common, language-independent ordering (referred to as 'canonical ordering') of the elements in the IS representation, and do the necessary reordering in analysis and synthesis:



+ featurisation of prepositions:



2.4 Reordering at Sentence Level

Two examples of reordering at sentence level:

sentence 1

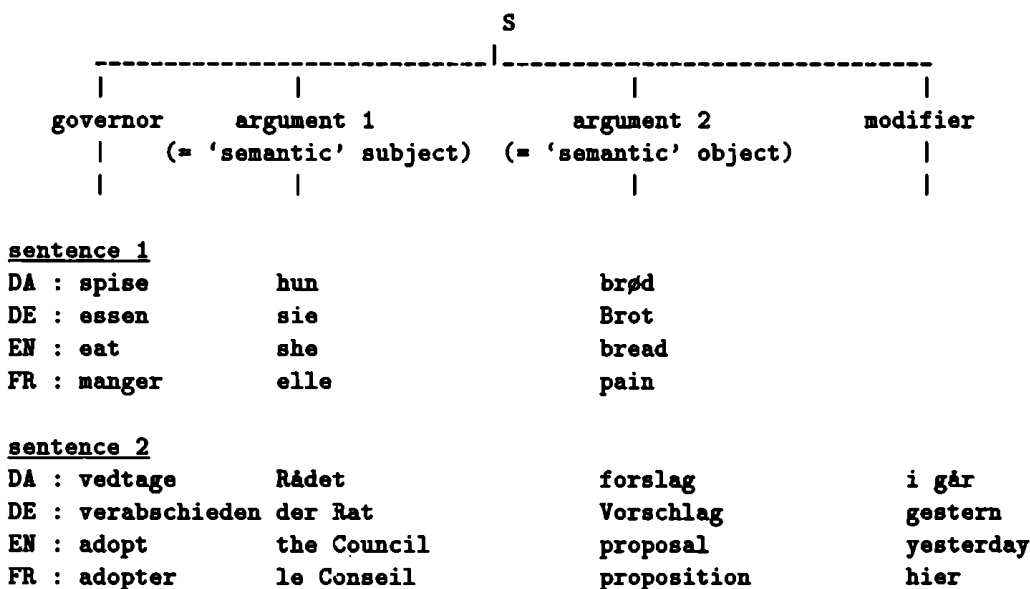
<u>NP + Vaux + Vmain + NP</u>	<u>NP + Vaux + NP + Vmain</u>
DA : Hun har spist brødet	DE : Sie hat das Brot gegessen
EN : She has eaten the bread	
FR : Elle a mangé le pain	

sentence 2

<u>AdvP + Vaux + NP + Vmain + PP</u>
DA : I går blev forslaget vedtaget af Rådet
<u>AdvP + Vaux + NP + PP + Vmain</u>
DE : Gestern wurde der Vorschlag vom Rat verabschiedet
<u>AdvP + NP + Vaux + Vmain + PP</u>
EN : Yesterday the proposal was adopted by the Council
FR : Hier la proposition a été adoptée par le Conseil

In analysis, articles and auxiliary verbs are featurised and removed from the structure, and the fact that the sentence is in passive voice is marked as a feature at the top node. At present, we do not use a refined set of semantic case roles but restrict ourselves to a numbering of arguments, where i.a. the subject of a sentence in active voice is labelled 'arg1' and the object is labelled 'arg2'. The maximum number of arguments in a sentence is 4.

Somewhat simplified, and without feature information, the IS representation of the two sentences looks like this:



The canonical ordering of the elements is in itself fairly straightforward and poses no major problems. What creates problems may be differences between languages and differences between language groups in analysis of some constituent, e.g. as complement or modifier. This is the reason why we are very wary of introducing a too ambitious approach in assigning case roles, as this would give rise to inconsistencies between assignment carried out in different language groups.

3 Transfer of Features

Here again we distinguish between three possibilities:

- 3.1 Features which are transferred unchanged.
- 3.2 Features which are not transferred but calculated again in the target language or found in the target dictionary.
- 3.3 Features with an explicit translation in the transfer component.

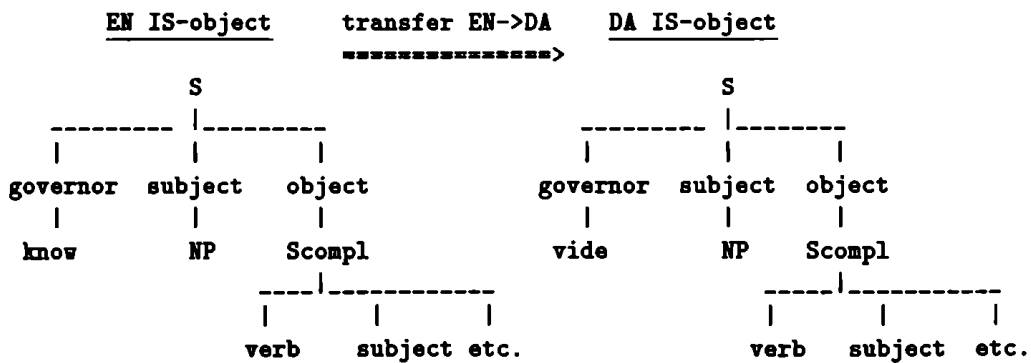
A feature has the form {attribute=value}, and what is transferred unchanged, calculated or translated explicitly is only the value of the feature.

In the first two cases no explicit transfer is needed, which means that we have simple transfer or interlingual treatment. In 2.2.2 above we mentioned definiteness as an example of a surface phenomenon that gives rise to both the first two types of transfer of features. In many cases morpho-syntactic definiteness may express semantic definiteness in a consistent way across languages, and in these cases we may transfer the value for the 'definiteness' attribute unchanged. Some (sub)categories, however, allow only of one of the paradigmatic set of values for definiteness, and this value may not be the same for different languages. In these cases the value is not transferred but found in the target dictionary — or in the target grammar if it is possible to generalise over a class of words, cf. the example

mentioned in 2.2.2 with country names, e.g. 'la France' {definiteness=definite} versus 'Frankrig' {definiteness=absent}.

In general, feature values which are not transferred, are typically bound to a lexical value, e.g. gender and semantic features on nouns and pforms on verbs (i.e. the preposition used in a valency bound PP), as well as other valency frame information, including restrictions on the semantic features of valency bound complements. These values are looked up in the dictionary. The value for 'gender' is then used to generate the correct form of modifying adjectives and determiners, and the valency information in the dictionary entry for a verb is matched with the available information on the complements. It is also used to connect the complement by means of the correct preposition in the target language.

Where there is more than one translation of a verb, the valency information in the target dictionary is used to decide which translation matches the structure of the IS object, which may be transferred unchanged, e.g.



lexical transfer rules

```

{en_lu = know} => {da_lu = kende}
{en_lu = know} => {da_lu = vide}
    
```

Danish target dictionary

```

{da_lu = kende, da_isframe = np_subject_np_object}
{da_lu = vide, da_isframe = np_subject_scompl_object}
    
```

We do not need to transfer the valency information of 'know', as only 'vide' matches the transferred IS-object, due to the restriction on 'da.isframe' in the Danish target dictionary (the description here is somewhat simplified).

We want to restrict explicit translation in the transfer component to lexical values in the narrow sense as uninflected wordforms. But this lexical transfer may also be reduced through an interlingual approach to certain categories of words. We have already mentioned function words such as noun determiners and auxiliary verbs, which are featurised and given an interlingual description.

But where we really hope to save a lot of explicit transfer rules is in the treatment of terms. The implementation of terminology is just being started, but we hope to treat the greater part of the planned 20.000 entries dictionaries

The only difference in the IS representation for the two languages are the lexical values for the two non-terms, so the only explicit transfer rules needed are the following:

FR-DA transfer dictionary

```
{fr_lu = adopter} ==> {da_lu = vedtage}
{fr_lu = hier}    ==> {da_lu = i_går}
```

All other information is contained in the two monolingual dictionaries:

<u>FR dictionary</u>	<u>DA dictionary</u>
{term = 184, fr_lu = 'France', fr_definiteness = definite, fr_gender = feminin, fr_number = singular}	{term = 184, da_lu = 'Frankrig', da_definiteness = absent, da_gender = neuter, da_number = singular}
{term = 200, fr_lu = 'proposition', fr_gender = feminin}	{term = 200, da_lu = 'forslag', da_gender = neuter}
{term = 237, fr_lu = 'Conseil', fr_definiteness = definite, fr_gender = masculin, fr_number = singular}	{term = 237, da_lu = 'Rådet', da_definiteness = definite, da_gender = neuter, da_number = singular}
{fr_lu = 'adopter', fr_isframe = subject_object}	{da_lu = 'vedtage', da_isframe = subject_object}
{fr_lu = 'hier'}	{da_lu = 'i_går'}

For clarity of exposition, only information relevant to our example is included here.

In this example we distinguish between 'definiteness' and 'fr_definiteness'/'da_definiteness'. The idea is that a feature may have a language-independent attribute name in cases when it expresses semantic information to be carried over, and the same attribute name with a language prefix in cases when the value is not semantically significant but concerns monolingual wellformedness. The distinction between universal features and monolingual features is currently made in EUROTRA by means of uniform attribute names + prefixes, which enables/disables matching, but this way of using the same attribute name with or without prefix is not implemented.

EUROTRA-DK
Njalsgade 80
DK-2300 København S.
poul@eurotra.dk