LINGUISTIC DEVELOPMENTS IN EUROTRA SINCE 1983.

Lieven Janssens
Katholieke Universiteit Leuven (Belgium)

I wish to put the theory and metatheory currently adopted in the Eurotra project (Arno83) into a historical perspective, indicating where and why changes to its basic design for a transfer-based MT (TBMT) system have been made.

1. A basic model for comparing TBMT theories.

Let $T$ be some theory of representation, inducing sets of representations $R_n$ and $R_l$ for languages $L_n$ and $L_l$, respectively. Transfer-based translation is described as follows:

$$ TRF : R_n \rightarrow \rightarrow \rightarrow R_l $$

$$ AN : \rightarrow \rightarrow \rightarrow GEN $$

$$ L_n \rightarrow \rightarrow \rightarrow L_l $$

where $AN$, $GEN$ and $TRF$ are binary relations, and $TRA$ is the composition of $AN$, $TRF$ and $GEN$, i.e.

$$ (i) \ AN \in L_l x R_l, \ GEN \in R_l x L_l, \ TRF \in R_l x R_l $$

$$ (ii) \ TRA = AN \ast TRF \ast GEN $$

We also need to introduce two parameters, viz. stratification and dimensionality, to characterise hypotheses about $T$. A theory $T$ is stratified when it consists of a set of subtheories $\{T_1, T_2, \ldots, T_n\}$, each characterising a set of representation $R_i$, such that

$$ (iii) \ AN = AN_1 \ast AN_2 \ast \ldots \ast AN_n $$

$$ (iv) \ AN_1 \in L_1 \ast R_1, \ AN_2 \in R_1 \ast R_2 $$

Otherwise, a theory is monostratal.

A theory $T$ is multidimensional when descriptions of linguistic objects along several linguistic dimensions are merged into one single representational object. The notion of linguistic dimension is meant to correspond to some organising principle for a theory of representation (e.g. constituency, grammatical relations, logical semantics, etc.). Otherwise, a theory is monodimensional.

In what follows we describe the various Eurotra approaches to TBMT in terms of this basic model.

1. The first Eurotra design (Arno83).

Initially, due to its BETA inheritance, Eurotra adhered to a monostratal multidimensional model for TBMT. Computationally, it was based on the Grenoble formalism of the générant de structures (gds). Linguistically, it advocated a diluted form of dependency theory as a basis for TBMT.

The major concern was directed at elucidating the division of labour between AN, TRF and GEN, and at deriving implications. The principles of Q-differentiation required that $T$ be sufficiently expressive to ensure that all meaning aspects of text that are relevant for translation be represented in members of $R$. The two principles together provided a basis for designing a transfer device that was (1) developmentally simple, and (2) Q-preserving. These are necessary features of any multilingual TBMT system striving for good-quality translation.

1.2. Multidimensional.

Despite its success in providing an initial framework for Eurotra, Arno83 failed dismally when it came to deriving from it a substantive linguistic representation theory. The failure was unrelated
to the absence of motivation for the GEN-META vestiges.

The gds comprised a flat geometry and a rich decoration on the nodes. Given the requirement of merging, the geometry for all dimensions (text string, morphology, surface syntax, deep syntax, semantics) had to be very similar; this was only possible by making the geometry quite meaningless, and by putting the whole expressive burden on the labelling of nodes. The need to preserve surface word order (robustness) gave geometry its only interesting task: the representation of word order through the ordering of sister nodes. Within a merged approach, this requirement led to the arbitrary interdependence of the subtheories for the various linguistic dimensions. The problem was most tangible in the design of a substratum of T, for a semantic dimension. T, became unnecessarily complex and inconsistent. Given the absence of linguistic constituents built into the tools and the failure of the framework to answer substantive linguistic questions, debates about the merits of particular representational choices were inconclusive.

We give an example of linguistic procrustization. Surface word order being represented by the order of sister nodes in the merged tree (the gds), tree geometry was seen as ordered. The geometry of dependency representations, on the other hand, are normally unordered. The way out was a relabelling of DT as a compromise between DT and X-theory with a single bar: a subset of the information about the governing node was lowered into the subtree representing its dependents and to require that the subtree be ordered conforming to the position of elements in the input text. This worked badly with all sorts of difficult linguistic phenomena: excentric constructions (e.g., conjunction), gapping, discontinuity, long-distance dependencies, etc. Much of the linguistic research, then, was aimed at overcoming these problems in a principled way by means of a theory of empty elements. Although the latter was intuitively consistent, it caused such an increase in the complexity of the formalism that the latter defied any coherent formal characterisation.

2. The second design: [Ann084a,b].

The first design was, amongst other things, unable to clash out the problem of robustness. Combining a multidimensional representation with a basically all paths combinatorial algorithm led to the inability to rely on the actual computation of combinations of information required by the safety net algorithm. The second design (which was never formally accepted by the project) purported to solve this problem, without eliminating multidimensionality. It was multistratal and multidimensional.

2.1. Multistratal.

It was observed that the representations induced by T, had to meet two (possibly conflicting) requirements: (1) they had to have sufficient expressive power to allow for adequate translation via simple transfer, and (2) their computation had to be feasible. As a consequence, T, was split into two subtheories, T, and T,. T, was then directed at the needs of adequacy for simple transfer and the latter to the reliability of presence of a consistent representation from which either the more pretentious T, representation was reached or, alternatively, translation via less simple transfer was possible. The model that emerged was the following:

```
 TRF
 Rn/r ------------ Rn/r
 ANA | TRF |
 GEN |
 ANA | L ------------ Rn/r |
 GEN |
 TRA |
```

The motivation for this design hinged on (1) the fact that the f-stratum could make use of knowledge in computational linguistics, (2) the f-stratum was a good starting point for innovative research on what T, should be for multilingual TBM, (3) the model gave content to the notion of safety nets (robustness), (4) developmental issues.

The claim made was that with a monolithic T,, the formulation of safety nets is hindered by the hybridity problem: their input domain could be any unpredictable combination of feasible computable and adequate information on several dimensions in the gds. The new design provided the f-stratum as a more reliable basis for safe safety nets.

2.2. Multidimensional.

This feature of the design did not change, Instead of one multidimensional representation, we now had two. No further attempt was made, however, to justify the use of multidimensional representations.

3. The third design: [Ann084b,a].

Given the rejection of theoretical modularity on the basis of considerations of reliability of computation, the only course to take seemed to be to abandon the multidimensional view itself and to let the strata themselves represent linguistic dimensions. The new model became multistratal and monodimensional.

3.1. Multistratal.

T, was described as a set of independently defined subtheories for representing normalised text (ns), morphology (m), surface syntax (ss), deep syntax (ds) and semantics (sema). They were conceptually related to each other, however, by being based on a common central notion of dependency defined in terms of slotfilling and modification. A strength of this move is that linguistics in Eurotra could now profit from linguistic work in the outside world.

The proposal suffered, however, from the absence of a clear view on what sorts of dedicated operations were needed to actually map between arbitrarily different dependency trees. Nor were considerations of the computational complexity of arbitrary tree-transformation formalisms taken into account in the definition of the levels. A proposal to relate all these levels to each other by giving them all a lexicalist underpinning was rejected by the C.E.C. Finally, a stratificational strategy was imposed on the makers of the design, with the (unjustified) intuition that it would provide a basis for the incorporation of safety nets into the model.
The model now roughly looked as follows (with question marks indicating undefined parts):

```
AN  | TRF   | GEN
--|-------|---
R_{text} | R_{nt} | R_{mo} | R_{cs} | R_{rs} | R_{sem} | R_{gen}
```

Each stratum corresponds to an autonomous generating device for a representation language. Each generator consists of a set of atoms and a set of constructors that together allow for the generation of L(G), a set of formally well-formed derivation trees. The latter are then evaluated (by unification) to a set of meaningful representations, R(G).

The intuition underlying this model is that translation between natural language texts can be split up into a sequence of more primitive translations between elements of adjacent generators. Adjacent generators must be devised so that the primitive translations that obtain are also simple. This is taken to mean that primitive translations must be (1) compositional and (2) one-shot. The justification for compositionality is the intuition that the translation of some expression E is a straightforward function of the translation of E's parts and of the way these parts are put together. The latter is required to restrain the complexity of this function; the codomain of a primitive translation must always be well-formed in terms of the target generator. This forbids internal strategy inside translators.

The project is examining various hypotheses about particular instantiations of this core model: e.g. translators could perform any one of the following four mappings: (i) derivation to derivation, (ii) derivation to representation, (iii) representation to derivation and (iv) representation to representation. Possibility (i) was found to be too restrictive. We now study possibility (iii). Note the similarity between (iv) and the structural correspondence approach adopted in LFG for mapping between information structures of a different nature.

3.2. Monodimensional.

Representations reflect only one linguistic dimension: the gds approach was completely abandoned.

The theories identified described the representation of normalised text strings, the internal structure of words, the surface dependency, the canonical dependency and the semantic dependency of the input texts.

4. The present design [desT85|Arno86].

The properties of the current Eurotra design constitute the topic of Arnold & des Tombe's paper in this volume. Here, I merely relate it to previous hypotheses about the Eurotra translation model.

The design is multistratal and monodimensional and can be depicted as follows:

```
AN  | TRF   | GEN
--|-------|---
G_{text} | G_{nt} | G_{mo} | G_{cs} | G_{rs} | G_{sem} | G_{gen}
```

5. Conclusion.

I hope to have slightly lifted the veil that has hidden the Eurotra project from the scientific community for a number of years. It has become clear, hopefully, that the Eurotra design has become more homogeneous and that it constitutes a valuable step towards a better understanding of the problem of machine translation.

REFERENCES.

[Arno83]: Arnold, Jaspaert & des Tombe, Linguistic Specifications, Version 1, C.E.C., 1983.
[Arno84a]: Arnold, Jaspaert & des Tombe, EFL-3 Final Report, C.E.C., 1984.
[Arno84b]: Arnold, Jaspaert & des Tombe, EFL-5 Final Report, C.E.C., 1986.
[Arno85a]: Arnold, Jaspaert & des Tombe, Eurotra Linguistic Specifications: Version 3, C.E.C., 1985.
[Arno86]: Arnold & des Tombe, Basic Theory and Methodology in Eurotra, to appear in: B. Nirenburg (ed.), Theoretical and Methodological Issues in MT, 1988.
[desT85]: des Tombe, Arnold, Jaspaert, Johnson, Krauwer, Rosner, Vartie & Warwick, A Preliminary Linguistic Framework for EUROTA, In: Proceedings of the Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages, Colgate University, 1985, 263-269.