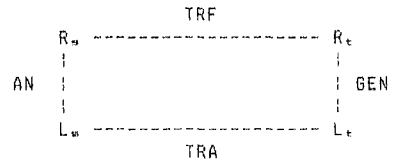
LINGUISTIC DEVELOPMENTS IN EUROTRA SINCE 1983.

Lieven Jaspaert
Katholieke Universiteit Leuven (Belgium)

I wish to put the theory and metatheory currently adopted in the Eurotra project (Arno86) 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_u be some theory of representation, inducing sets of representations R_s and R_t for languages L_s and L_t (seen as sets of texts), respectively. Transfer-based translation is described as follows:



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

(i) $AN \subseteq L_s \times R_s$, $GEN \subseteq R_t \times L_t$, $TRF \subseteq R_s \times R_t$
(ii) $TRA = AN \circ TRF \circ GEN$

We also need to introduce two parameters, viz. stratification and dimensionality, to characterise hypotheses about T_u. A theory (e.g. for the AN relation) is multistratal when it consists of a set of subtheories (t1,t2,...,tn), each characterising a set of representation R_i, such that

(iii) $AN = AN_1 \circ AN_2 \circ \ldots \circ AN_n$,
(iv) $AN_1 \subseteq L_s \times R_1$,
$AN_2 \subseteq R_1 \times R_2$,
...
$AN_n \subseteq R_{n-1} \times R_s$.

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

The observation that theoretical linguistics had been incapable of providing a practically applicable basis for translation had led Grenoble to build almost no linguistic commitment into the gds formalism. Every possible form of linguistic interpretation was to be expressed as an ordered tree with complex property lists on the nodes, which was manipulated by two basic operations, viz. tree transformations and lexical substitution. The GETA preoccupation with robustness, on the other hand, made them require that all linguistic information about texts should be merged into one single gds. On failure to compute parts of a deeper linguistic dimension, the intuition went, some clever algorithm could be used to extract from the gds an equivalent piece of representation on the next less pretentious dimension. The logical extreme of this reasoning was that, if all else failed, it should be possible to recover the original text from the gds.

Grenoble, however, had perceived the usefulness of dependency theory (DT) for TBMT. There is a sense in which DT is a lexically oriented theory of language, and, in the end, translation is a question of getting the right translation for words. Nevertheless, the marriage between DT and the gds design led to (1) procrustinated linguistics, and (2) a formalism with untractable semantics.

1.1. Monostratal.

The advocated representation theory was not stratified in any interesting sense. Rather, the whole burden of modularising the relation between text and representation was put on the translation of the relation into a procedure: discussions about clever linguistic strategies were long but were never brought to bear.

The innovation of [Arno83] was its attempt to derive requirements on T_u from a set of more abstract principles, seen as a theory of MT providing a framework within which possible substantive theories for TBMT could be devised and compared. The weakness of the framework was to seek to motivate the tools inherited from GETA a-posteriori. Its merit was to be a partial theory of TBMT, independent of the inheritance.

Its major concern was directed at elucidating the division of labour between AN, TRF and GEN, and at deriving implications on T_u from this understanding. The pivotal principles of the framework that have survived the many face lifts of the Eurotra model are isoduidy and Q-differentiation.

The principle of isoduiy allowed for a principled definition, in terms of properties of T_u, of the domain of the GEN relation of some language in terms of the codomain of the AN relation for that same language, thus indirectly defining the TRF relation. The principle of Q-differentiation required that T_u should be sufficiently expressive to ensure that all meaning aspects of text that are relevant for translation (called 'Q') 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 not unrelated

to the absence of motivation for the GETA 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 subtheory of T_u for a semantic dimension. T_u became unnecessarily complex and inconsistent. Given the absence of linguistic commitments built into the tools and the failure of the framework to answer substantive linguistic questions, debates about the relative merits of particular representational choices were inconclusive.

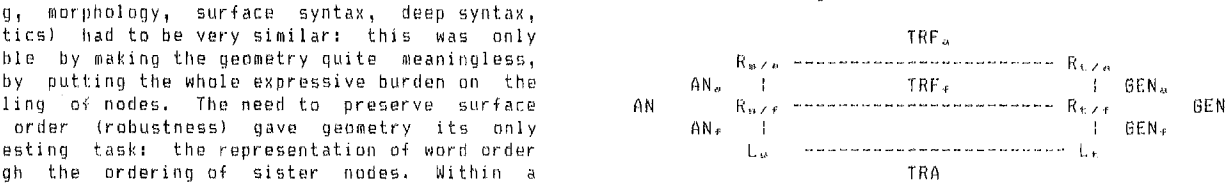
We give an example of linguistic procrustination. Surface word order being represented by the order of sister nodes in the merged tree (the gds), tree geometry was seen as ordered. The geometry of dependency representations, on the other hand, are normally unordered. The way out was a refashioning of 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: exocentric 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: [Arno84a].

The first design was, amongst other things, unable to flesh 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 multidimension-ality. It was multistratal and multidimensional.

2.1. Multistratal.

It was observed that the representations induced by T_u 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_u was split into two subtheories, T_a and T_f, were the former was 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_a representation was reached or, alternatively, translation via less-simple transfer was possible. The model that emerged was the following:



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

The claim made was that with a monolithic T_u, the formulation of safety nets is hindered by the hybridity problem: their input domain could be any unpredictable combination of feasibly 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: [Arno84b;Arno85].

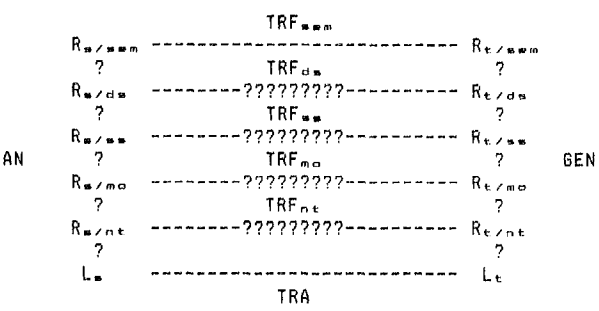
Given the rejection of theoretical modularity on the basis of considerations of reliability of computation, the only course to take seemed to be to abandon the multidimensional view itself and to let the strata themselves represent linguistic dimensions. The new model became multistratal and monodimensional.

3.1. Multistratal.

T_u was described as a set of independently defined subtheories for representing normalised text (ns), morphology (mo), surface syntax (ss), deep syntax (ds) and semantics (sem). 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):



3.2. Monodimensional.

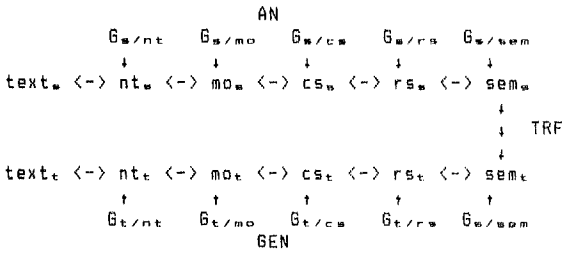
Representations reflect only one linguistic dimension: the gds approach was completely abandoned.

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:



4.1. Multistratal.

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 meaningfull 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.

4.2. Monodimensional.

The current strata envisaged are normalised text, morphology, configurational surface syntax, relational surface syntax and semantics. Morphology is based on work on word grammar as independent of phrase structure grammar. Configurational syntax draws from the X-theory literature. Relational syntax representations resemble LFG f-structures. The semantic stratum, finally, is not yet fully specified: this has to do with the very special requirements that translation by means of simple transfer puts on a semantic representation theory. The point is, however, that the non-semantic levels are claimed to be feasible (cfr. f-stratum in 2) and that they can thus provide a basis for researching a translation-oriented semantic theory.

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: Version1*, C.E.C., 1983.

[Arno84a]: Arnold, Jaspaert & des Tombe, *ETL-3 Final Report*, C.E.C., 1984.

[Arno84b]: Arnold, Jaspaert & des Tombe, *ETL-5 Final Report*, C.E.C., 1984.

[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: S. Nirenburg (ed), *Theoretical and Methodological Issues in MT*, 1986.

[desT85]: des Tombe, Arnold, Jaspaert, Johnson, Krauwer, Rosner, Varile & Warwick, *A Preliminary Linguistic Framework for EUROTRA*, In: *Proceedings of the Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, Colgate University, 1985, 283-289.