MACHINE TRANSLATION, LINGUISTICS, AND INTERLINGUA

Petr Sgall and Jarmila Panevová
Faculty of Mathematics and Physics,
Charles University
Malostranské n.25, 118 00 Praha 1, Czechoslovakia

ABSTRACT

An adequate, complete, and economical linguistic theory is necessary for MT and the question is whether a consistent use of the often unduly neglected dependency syntax, including a systematic description of topic and focus, cannot serve as a reliable base for the grammar of an interlingua, or of a set of interrelated interface structures.

1. As Slocum (1985) convincingly shows, the attitude towards translation in general, and therefore also towards automatic translation in the U.S.A. never has been based on urgent wide-spread needs of translating technical texts, and mostly has not been connected with broad interest in theoretical background. Outside the U.S.A., with the exception of G.E.T.A., Grenoble, the research had the character of scattered projects carried out by relatively small groups; only in the recent years the EUROTRA project and, especially, the two Japanese projects bring some hope as for the possibility of sufficiently concentrated research.

The question whether linguistics is able to offer a reliable theoretical basis for MT cannot be answered in a qualified way without examining such linguistically based systems as Garvin's 'fulcrum' approach (which was abolished on external grounds, after the unfortunate ALPAC report) or the systems formulated by Kulagina and Apresyan. Certain features of their frameworks, as well as of Vauquois' (1975; Vauquois and Boitet, 1985) are more closely connected to classical structural linguistics than is the case with other MT systems. Also in Prague, the research group of MT and formal linguistics at Charles University has devoted much effort (starting at the end of the 1950's) to identify the positive results of classical European linguistics and to reformulate them in a metalanguage that would make them usable in the context of Chomskyan (and Montaguean) methodology and of automated language processing (now see Sgall et al., 1986).

2. The requirements on linguistic theory as a background for MT can be summarized as follows:

(a) Adequacy: The theory should underlie relatively complete descriptions reflecting the structure of language. Since humans differ from computers (in freely combining factual knowledge and other mental capacities with their knowledge and use of language, and in being able to develop their language while using it), the correspondence between theory and its MT application cannot be immediate. The open-endedness of language makes it necessary to restrict the completeness of the description to a reasonably estimated core of language, leaving the (possibly not too large) periphery in the application to postediting, etc.

(b) Testability: It follows from (a) that testability also is limited: not every counterexample disqualifies a theory.

(c) Economy: To be applicable, the theory cannot be too complex. It seems necessary to draw a boundary line between the system of language on the one side, and its use or its semantic interpretation on the other, although in several respects it may be useful for the applications not to follow this boundary quite exactly.

(d) Modularity: Since huge programs are extremely difficult to be handled (debugged, updated, etc.), priority will be given to such a theory that not only allows for a division of labour between the description of linguistic and communicative competence - see (c), - but makes also a cooperation between specialists in the different layers of language itself possible.

The comparison of different approaches to linguistic theory as to point (a) is a matter of the theory itself; let us only note that many theories seem not to be sufficiently adequate in that they do not properly distinguish between the three dimensions of the sentence structure (valency or theta roles, coordination and apposition, and also topic and focus, which often is almost altogether neglected[1]) and the morphological categories (tense, aspect, number, definiteness, and so on); the latter occupy no immediate positions in the structure of the sentence with its recursive properties, and thus it is not adequate to denote e.g. prepositions as if they permitted for unlimited complementation, as verbs do.

For point (b) it is important that the theory uses operational criteria in delimiting its units and oppositions, thus representing a suitable starting point for implementable application systems.

With regard to (c), the relative generality of the formulations used by the theory is relevant; thus e.g. Chomsky´s universal principles are relatively economical. On the other hand, the abundance of nodes in the P-markers (cf. what was just said on point (a)) brings along the necessity to use tree pruning and to introduce devices making it possible to find an orientation in the unnecessarily large trees.

As for (d), it seems preferable to work with two levels of sentence structure and with a separate level of morphemic representations in the theory, although in the applications this pattern may be simplified (we are then aware what we have left out e.g. in our parser, and are able to restore a missing subpart, if this proves to be necessary, e.g. when the system is to be generalized to handle new kinds of texts).

A systematic investigation into comparing different linguistic theories from these viewpoints has resulted in our preference for dependency grammar, based on valency or theta roles (see Sgall et al., 1986, for a detailed discussion). A dependency based linguistic description is adequate in the quoted respect (e.g. the morphological values are denoted here by parts of complex node labels); the theory is fully testable and uses operational criteria, and it ensures both economy (no non-terminal symbols are present in the representations although as many as necessary can be used during the derivation procedure) and modularity (the underlying representations contain all the semantically relevant in-

[1] The relevance of the topic-focus articulation for translation and for other aims of language comprehension can be illustrated by the following examples: In the hallway one smokes should be distinguished from One smokes in the hallway similarly as Few books are read by many men from Many men read few books.

formation, since also the topic-focus articulation is denoted here).

Bloomfield´s ´exocentric´ constructions are often mistakenly understood as an obstacle for dependency syntax; however, as Robinson (1970) showed, they can be handled without serious difficulties within dependency trees. Let us add that, if constructions are analyzed in the terms of word classes (parts of speech), rather than in those of individual words, than the distributional properties clearly show that e.g. your sister is a noun group (since e.g. Mother or syntax occur without a determiner), to hit the ball is a verb group (due to to read,...), and also a sentence has a verb as its governor, since in It rains no subject (Actor/Bearer) is present at the level of meaning (or in the underlying structure).

A formal treatment combining dependency syntax with a description of coordination and apposition, allowing for an indefinite number of sister nodes, was presented by Plátek et al. (1984).

As one of the referrres of our papers has duly recalled, the number of publications concerning dependency grammar is much smaller than that on constituent structures, but the popularity of the model is not directly relevant for its evaluation. Therefore it seems highly useful to notice the advantages of the less known model, a more intensive use of which might be of importance for the further development of the field.

3. An interlingua for MT can well be based on such a theory. Since the 1960´s - see e.g.Mel´čuk (1962), Vauquois (1962), Sgall (1963) - the research in this direction has been connected with theoretical investigations. It has been clear that the formulation of an interlingua is a practical task, for the underlying units differ from one language to the other, so that the structure of interlingua is based rather on the structural similarities (formal universals) of languages than on an assumed identity of their underlying structures, or their patternings of meaning.

As for the known difficulties concerning e.g. the formulation of fail-soft rules or the presence of surface clues (see Slocum, 1985,5; Vauquois and Boitet, 1985), it appears that for a multilingual system of MT these disadvantages have to be compared with those present in the large number of binary systems which are otherwise necessary. The difference between the use of an interlingua and of a smaller number of "interfaces" (one for each language)appears not to be crucial. If, for a system including n languages, m among them display a certain opposition (that of dual versus plural number, or of gender with personal pronouns, etc.), then the degree of importance of this opposition for the system depends on the difference between n and m and on the importance of the languages displaying the opposition. Extremely marginal oppositions will probably be ignored in a system using interfaces as well as in one with an interlingua. In this case, a translation between two languages exhibiting the marginal opposition will be faced with a similar problem as a translation from a "prototypical" language into a "marginal" one (e.g. the use of dual number will be determined - perhaps only for some cases - by contextual clues, rather than by the presence of dual in the input text).

If the relative weight of such surplus difficulties (and resulting mistakes) is considerable, then it may be useful to formulate interfaces, perhaps not always in a one-to-one correspondence to the processed languages, but relating to certain groups of them. Certain "dialectal" differences in the interlingua would then be useful, each of which would share some opposition(s) with a group of the processed languages. This may concern the differences between languages having and not having articles, verbal aspects, various moods, and so on. The substitution of a single interlingua by a set of closely related interface structures (see Vauquois and Boitet, 1985,32; as for its application in the EUROTRA project, Johnson et al., 1985,164) perhaps is also important with regard to handling the semantic relationships between the lexical units of the languages concerned.

This schematical view can be systematically elaborated only on the basis of experience with multilingual MT systems.

4. A not quite negligible experience with MT systems based (at least to a great part) on dependency syntax has been gained already. The Grenoble group has used a graph grammar based on this approach within a system that is multilingual, though centred aroundFrench (see Vauquois, 1975; Boitet and Nedobejkine, 1981; Vauquois and Boitet, 1985,28f); although in this system the dependency relations are used along with a kind of phrase structure, the importance of complex node labels and of the syntactic relations (valency) has always been fully recognized. Also Nagao et al. (1985,esp.98) point out that dependency tree structures are used in their project (which certainly belongs to those with the best traditions and results); in the Eurotra system the dependency relations and the notion of ´gov (ernor)´ play an important role (see e.g. Johnson et al., 1985). In Prague, especially the English-to-Czech translation project, the main author of which is Kirschner (1982; 1984), is based on a dependency description.

5. The perspectives of MT seem to be connected with two major conditions, in addition to the choice of an appropriate underlying linguistic theory, which we discussed above:

(a) As is known, for the resolution of many lexical ambiguities and also for the identification of grammatically obligatory values of the target language not present in the input text, a MT system has to include not only a purely linguistic description. It has to be found out to what degree the practical purposes of MT can be achieved by systems "modelling the world" by such elementary means as sets of semantic features. Where means of this kind will be found to be insufficient, it is probable that neither data bases of the common types will do. It is then necessary to look for suitable kinds of knowledge representation systems.

(b) The main perspective appears to be connected with the hope that a wider practical application of MT will lead to a new situation, in which the construction of MT systems will no longer be a matter of small research groups scattered and more or less isolated in different countries, but there will emerge large-scale and well--coordinated international projects based on the best results achieved and verified by widespread practical application. Under such new circumstances it will be possible not only to compile grammatically well founded data on tens of thousands of lexical units from different languages, but also to connect translation systems in an effective way with broadly based nets of knowledge representation. Effective ways of human-machine interaction can then be found, and the formulation of appropriate intermediate languages will meet good conditions. Post-editing will certainly remain necessary, the main condition being that it should not be much more difficult than it is with human translations of technical texts (although other kinds of mistakes will prevail).

REFERENCES

Boitet Ch. and N. Nedobejkine (1981), Recent developments in Russian-French machine translation at Grenoble, Linguistics 19, 199-271

Johnson R., King M. and L.des Tombe (1985), EUROTRA: A multilingual system under development, Computational Linguistics 11, 155-169

Kirschner Z. (1982), A dependency-based analysis of English for the purpose of Machine Translation, Explizite Beschreibung der Sprache und automatische Textbearbeitung IX, Prague

Kirschner Z. (1984), On a dependency analysis of English for machine translation. In Contributions to Functional Syntax, Semantics and Language Comprehension, ed. by P.Sgall, Prague - Amsterdam, 335-358

Mel´čuk I. (1962), K voprosu o "grammatičeskom" v jazyke-posrednike. In: Mašinnyj perevod i prikladnaja lingvistika 4, Moscow, 25-45

Nagao M., Tsujii J. and J.Nakamura (1985), The Japanese government project for Machine Translation, Computational Linguistics 11, 91-110

Plátek M., Sgall J. and P.Sgall (1984), A dependency base for a linguistic description. In: Contributions to Functional Syntax, Semantics, and Language Comprehension, ed.by P.Sgall, Prague - Amsterdam, 63-98

Robinson J. (1970), Dependency structures and transformational rules, Language 46, 259-285

Sgall P. (1963), The intermediate language in Machine Translation and the theory of grammar. In: American Documentation Institute, 26th Annual Meeting, Chicago, Ill., 41-42

Sgall P., Hajičová E. and J.Panevová (1986), The Meaning of the Sentence in Its Semantic and Pragmatic Aspects, Prague - Dordrecht, Holland

Slocum J. (1985), A survey of Machine Translation: its history, current status, and future prospects, Computational Linguistics 11, 1-17

Vauquois B. (1962), Langages artificiels, systèmes formels et Traduction Automatique. In: Cours à l´Eté de l´OTAN, ed. by A.Ghizetti, Oxford, 1966, 211-236

Vauquois B. (1975), La traduction automatique à Grenoble. Documents de linguistique quantitative 24, Paris

Vauquois B. and Ch.Boitet (1985), Automated translation at Grenoble University, Computational Linguistics 11, 18-36