[From: L.Grähs et al.,eds.*Theory and practice of translation* (Bern: Lang, 1978)

Machine Translation — State of the Art

BENGT SIGURD (Stockholm)

O. INTRODUCTION

I want to make it clear from the outset that machine translation (mechanical translation, MT, automatic translation; I will use the terms interchangeably) has its place only in industrial translation - translation for special purposes, where speed is more important than quality, or when a rough idea of the content is sufficient and better than nothing. Machine translation is not being developed for the translation of literary texts, poetry, Shakespeare or the Bible.

Typical sponsors of machine translation projects now or earlier are NASA, the Foreign Office, the Pentagon, the Foreign Broadcast Information Service in the USA, Euratom and Nato in Europe. It has been reported (Hays & Mathias, 1976, p. 43) that NASA used automatic translation routinely for working papers used in the Apollo-Soyuz space project.

Automatic translation of languages was one of the first applications of computers suggested. The dreams of the early pioneers have not come true, however. Development was fast in the beginning and large dictionaries were built, but the quality of automatic translation has improved only little during the last decades. As a consequence researchers have been looking for new points of attack, based on recent developments in syntax, semantics, artificial intelligence and computational linguistics .

Automatic translation faces all the well-known problems of human translation, such as the problem of finding perfect equivalents in other languages. In addition machine translation faces the problem of simulating human verbal behaviour: text comprehension and text production. That is why automatic translation has its special problems. In my view automatic translation projects are interesting from theoretical points of view as they deepen our insights into the translation process. I might even add that the complexities of translating can be understood fully only by those who have tried to simulate the process by computer.

I will base many of my conclusions below on reports at a conference organized by the Foreign Broadcast Information Service (FBIS) in the USA in 1976. I will relate some of the work at my institute on machine translation. (I am grateful for further information about projects in machine translation or projects where large dictionaries or terminology are stored on computers for the benefit of human translators.)

1. SOME CURRENT PROJECTS

One of *the* few systems used routinely is a Canadian system for translating weather forecasts from English into French. The system is reported (in Hays & Mathias, 1976) to be in operation since the beginning of 1976.

The system operates at a speed of 1000 words per minute - which is fast - and the cost is one third of human translation. The system has a load of 30 000 words a day - which would mean that it is in operation only half an hour a day. The success of the project is mainly due to the fact that weather forecasts deal with a restricted semantic area (micro world). Only certain things happen, only certain objects appear or change their states. The dictionary can be restricted and the syntactic parser does not encounter a rich variation of constructions. Translation is mostly mapping English idioms on French ones in this case. The system has to rely on human translators when a sentence is rejected as untranslatable. Less than 20 %, however, of the sentences are rejected (due to misspelling, words not found in the dictionary or ungrammatical English). I think this is a fairly high figure, discrediting the whole system substantially. If a sentence is rejected, it is sent to the local translation bureau, where it is displayed on a screen and translated immediately by a human translator - who thus has to be available. I will not discuss the project any more - only scant information is given in the report - but I will show an example of the kind of translation that is carried out. Note changes in word order and the use of passager for periods .

> Mostly clear and cold with periods of very light snow today and Wednesday. Highs near minus 10 both days. Lows tonight minus 20 to minus 22.
>
> Aujourd'hui et mercredi généralement clair et froid avec très faibles chutes de neige passagères. Maximum pour les deux jours environ moins 10 Minimum ce jour moins 20 à moins 22.

The Georgetown University System has been developed over a long period of time. Georgetown University was one of the first institutions to embark on an automatic translation project. The project translates from Russian into English. It is reported to be used at Euratom, Italy and Oak Ridge National Laboratories, Tennessee. It is run by the library staff as a routine service. The system gives a rather crude output and a willingness to accept such style is required. According to estimates based on interviews, about 90 % of the users called the result good or acceptable, and about 90 % of the users had recommended or would recommend machine translation to their colleagues.

The great advantage of the system is speed. The translation of a 10-page article can be carried out in minutes. Key punching, however, requires 3 to 4 hours, and pre- and/or postediting maybe half an hour. A requester can have the translation back the day after he asks for it. Costs are said to be comparable to human translation.

Translation from Russian into English has always been a favourite aim in the USA. Priority has also been given to translation from Chinese into English. A project with that purpose is reported from University of California, Berkeley. There are a few commercially available systems, e. g. SYSTRAN and XONICS which focus on translation from Russian into English. The reports from such projects are difficult to judge as they seem to be over-enthusiast! c and tend to over-estimate their quality and potentials.

The Soviet Union had several projects in the 1950s, notably in Leningrad. Soviet researchers nave met with the same problems as other researchers and I have not heard much about Soviet projects recently. In Europe, Grenoble is known to have one of the few research groups still existing in the field. I think most of the activity that is of relevance for automatic translation today can be found in projects dealing with comprehension, text generation or artificial intelligence. In Europe, Edinburgh is known as a center of artificial intelligence research. In the USA, Stanford University is a center of artificial intelligence research and the two researchers Schank and Wilks have both taken an interest in automatic translation as well. But it is rather as a possible application of approaches suggested in artificial intelligence research: question-answering systems, inference systems and other information handling systems.

A critical review of automatic translation systems and related projects was undertaken by Hood Roberts some ten years ago (the ALPAC-report) and funding institutions and agencies have been restrictive to enthusiastic applicants since then. I think that report and its criticism was justified, but I do not think we should give up thinking in terms of automatic translation.

As a related goal on-line computer aids for human translators have been suggested. Such systems should make work simpler for translators, terminologists, lexicographers, editors, and typists involved in translation. The systems suggested would give the following advantages:

1. Increased productivity through accelerated dictionary lookup, rapid and con venient revision of successive translation drafts, high-speed layout and printing.

2. Easily activated text-related glossaries with statistical information and concor dance information would promote terminological consistency.

3. Reduced handling and consumption of paper through emphasis on visual display during all but the final phase.

Few computerized translation aids have been tested in large scale by translators. Most computerized translation facilities offered so far seem expensive and impracti - cal. Members of this conference may have better suggestions about computerized dictionaries or other equipment that really would be helpful.

I will mention one idea that has occurred to me: a computerized pocket sized school dictionary. Considering the success of the pocket calculators and their present use in class-rooms pocket dictionaries may not be completely utopian.

2. PROCEDURES IN THE ANALYSIS OF THE SOURCE TEXT

Before discussing translation proper I want to mention that automatic translation in a restricted sense can be achieved with few problems. Such cases are translating 1900 Swedish into 1950 Swedish by changing the spelling (e.g. spelling hvem as vem , vidt as vi tt etc). A program for this purpose has been developed by Benny Brodda at the Institute of Linguistics, University of Stockholm. It has been used for normalizing texts to be investigated subsequently by concordance and frequency programs. Similarly programs for the translation of numerals from and into different languages have been developed. Translation from a programming language such as ALGOL, FORTRAN, LISP into the internal computer language is a routine operation. Such translation systems are called compilers.

Automatic translation systems used in translation proper must simulate human translators and part of the work has consisted of finding out what human translators actually do when they translate. Automatic translation systems display the processes in flow charts. It is by no means clear how human translation (or interpretation) is achieved, or for that matter, how human beings understand and produce utterances. Fig. 1 lists some procedures that seem to be necessary in automatic translation systems. We will discuss the different processes below. The discussion is meant to show the complexity of the problems any automatic translation project sets out to solve.

2.1. Morpho-syntactic procedures

Automatic dictionary lookup procedures have been developed for many languages - and the problems are fairly well known (Sågvall, 1973). Given the input Flickornas in Swedish the result should be something like Girl + Plur + Def + Gen and whatever lexical information will be needed in the following procedures. The neces-sity of including irregular forms (such as ran) and idiomatic expressions is obvious, No dictionary can include all potential words, including all compounds and metaphori-cal constructions. The translation of Swedish or German compounds into other languages must often involve an interpretation of the meaning of the compound. A potential metaphorical compound such as bilstim cannot occur in the lexicon. As metaphores it is difficult to translate: car shoal is hardly acceptable, shoal

Fig. 1: Procedures distinguished in source text understanding

Procedures	Results	Problems
Morphological analysis and dictionary lookup	Morphemes and words identified, stylistic information, foreign equivalents	Alternative interpretations of ambiguous words (e.g. mean (adj.)/ mean (verb))
Syntactic analysis (parsing)	Sentence structure identified, stylistic information, foreign sentence equivalents Word class ambiguities solved	Alternative solutions (in Swedish "Osten åt katten", the cat or the cheese may both be the subject)
Case frame (semantic template) matching	Case structure (agent, object, instrument etc.) identification Identification of topic	Meaning of deictic expressions not clear
Discourse	Identification of the referents of deictic expressions (pronouns, definite articles etc.) Identification of plan and thematic progression	Difficulties in finding referents, which may be implicit. Problems of semantic representation of text
General knowledge (encyclopädic) consultation	Identification of implicit referents, inferences, solving ambiguities where "common sense" is needed	Impossible to store all (human) knowledge in the computer. Search procedures must be fast and "human"

of cars better, "group of cars considered as fish in a shoal" would be an explanation rather than a translation. (The problem of understanding metaphorical and other compounds is treated in Sigurd, 1976a).

All dictionary lookup programs leave alternative interpretations, and texts where the alternatives have not been eliminated are indeed difficult to read. The following is an example of an English text produced by a system translating from Russian. Reading is too much guesswork to be convenient.

> Construction/design of standard, his/its physical properties and method of realization are defined/determined/assigned by nature of magnitude/quantity, unit/one (of) (to/for) (by/with/as) which is reproduced, and/even/too by state/fortune of measuring/ dimensional techniques/practice in/on/of given area of measurements.

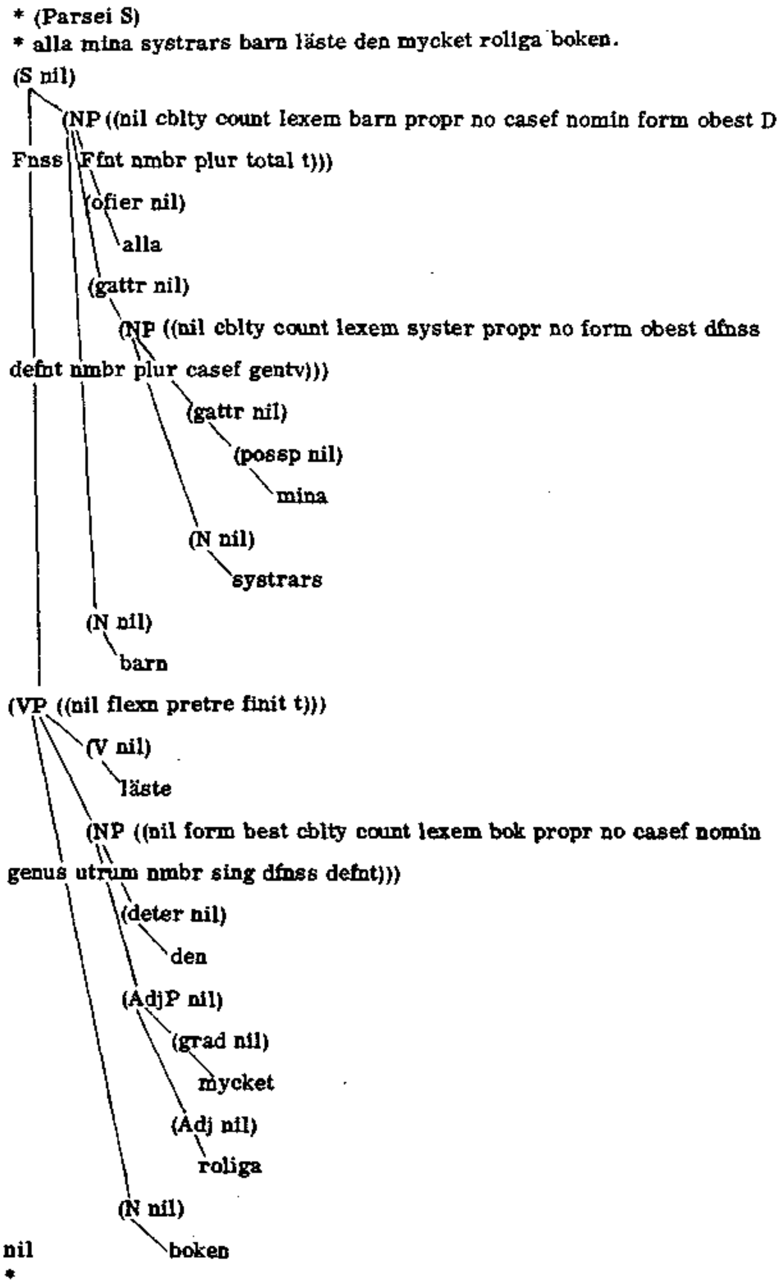
Word-for-word translation leaves us with bulky texts and patching up such texts by disambiguation programs based on the neighbouring two or three words is not enough. The proper way out must include a syntactic parsing program which identifies the syntactic structure of the sentence to be translated. Such a program also solves ambiguities, where word classes or parts of speech are at issue. In the sentence His talk was mean both talk and mean are ambiguous when considered in isolation: talk could be a verb or a noun, mean could be a verb or an adjective. A syntactic parser would make the proper choice on the basis information offered by the dictionary.

Parsing programs have been much studied - they are of interest to all comprehension or translation systems including the compilers translating from one programming language into another. Several parsers have been developed including one for Esperanto.

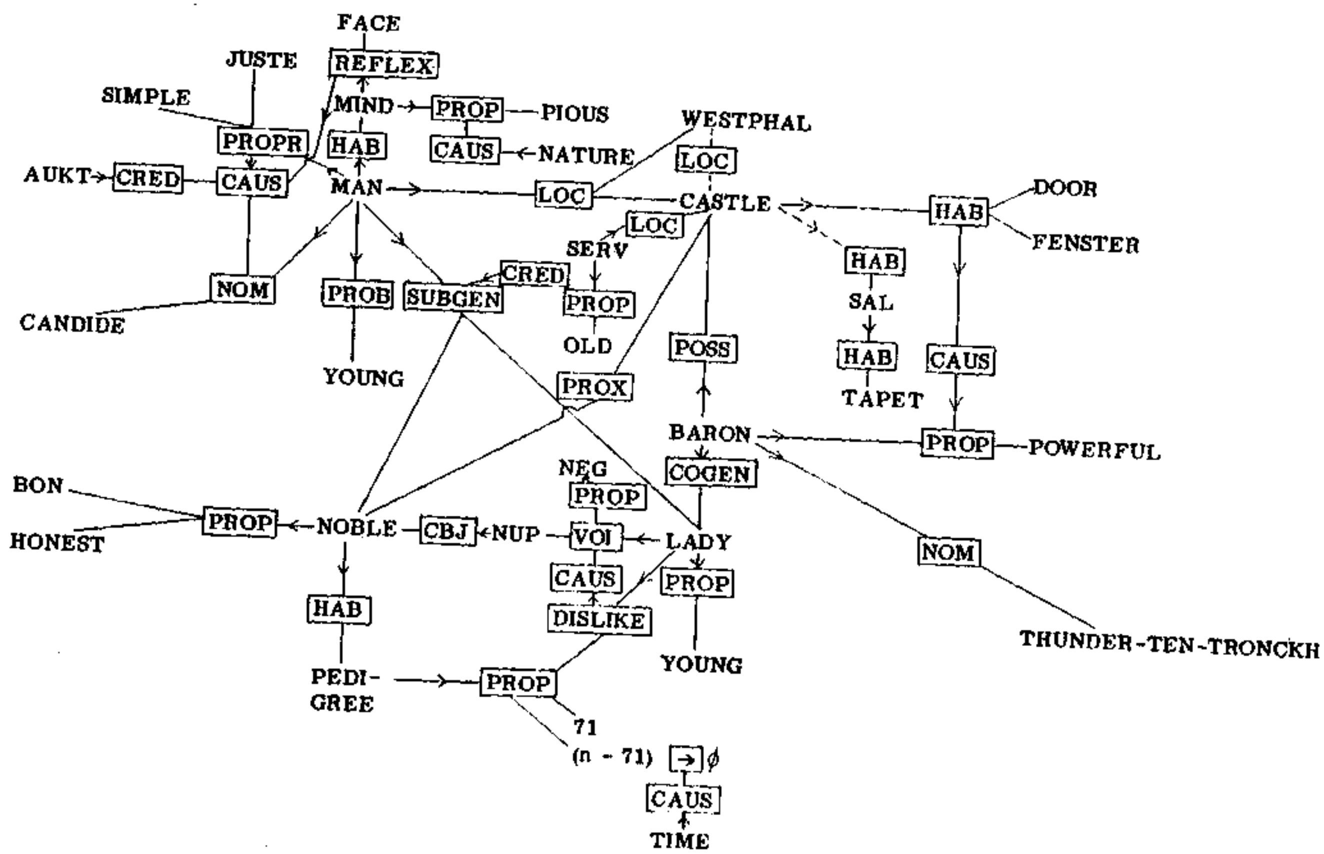
A parser for fragments of Swedish has been developed by C. W. Welin within the project Automatic Text Comprehension, at the Institute of Linguistics, University of Stockholm. Figure 2 shows the result of an analysis of "Alla mina systrars barn läste den mycket roliga boken", "All the children of my sisters read the very funny book". The parser delivers a tree with labeled nodes and additional information at the nodes. In a translating system foreign equivalents such as special position, genitive endings, prepositions etc. can be associated with the tree at proper points.

A parser tries to fit the words and morphemes of the sentence into a tree which conforms to the rules of a (generative) grammar. In dealing with programming languages or formal languages this grammar is fixed and well-defined. Natural language texts, however, are much less normalized, full of fragments, deletions, mistakes etc. A parser for natural language texts must handle distorted sentences without breaking down. At present most parsers require adherence to certain grammatical patterns, but many parsers are quite successful and impressive.

Fig. 2: The result of automatic analysis (parsing) of the Swedish sentence Alla mina systrars barn läste den mycket roliga boken (All my sisters children read the very funny book) (after Welin, 1976).



... represented as a semantic net-work. The first sentences of a French version are also given.



Il y avait en Westphalie, dans le château de M. le baron T-T-T, un jeune garçon à qui la nature avait donné les moeurs les plus douces. Sa physionomie annonçait son âme. Il avait le jugement assez droit, avec 1'esprit le plus simple; c'est, je crois, pour cette raison qu'on le nommait Candide. Les anciens domestiques de la maison soupçonnaient qu'il était fils de la soeur de M. le Baron, et d'un bon et honnête gentilhomme du voisinage, que cette demoiselle ne voulut jamais épouser, parce qu'il n'avait pu prouver que soixante et once quartiers, et que le reste de son arbre généalogique avait été perdu par l'injure du temps. M. le Baron était un des plus puissants seigneurs de la Westphalie, car son château avait une porte et des fenêtres. Sa grande salle même était ornée d'une tapisserie.

There lived in Westphalia, at the country seat of Baron Thunder-Ten-Tronckh, a young lad blessed by nature with the most agreeable manners. You could read his character in his face. He combined sound judgment with unaffected simplicity; and that, I suppose, was why he was called Candide. The old family servants suspected that he was the son of the Baron's sister by a worthy gentleman of that neighbourhood, whom the young lady would never agree to marry because he could only claim seventy-one quarterings, the rest of his family tree having suffered from the ravages of time. The Baron was one of the most influential noblemen in Westphalia, for his house had a door and several windows and his hall was actually draped with tapestry.

Det levde en gäng i Westfalen, i ett slott tillhörigt Herr Baron von Thunder-Ten-Tronckh, en ung man, åt vilken naturen förlänat en den mest fromma karaktär, Hans ansikte speglade hans själ. Hans tunga var lika rättfram som hans hjärta var okonstlat: sannolikt var detta grunden till att man gav honom namnet Candide. De gamla tjänarna i huset antogo, att han var son till herr Baronens syster och en hederlig och aktningsvärd adelsman från grannskapet, vilken denna fröken aldrig hade velat äkta, därför att han inte kunde uppvisa mer än sjuttioen anor; resten av hans stamtavla hade gätt förlorad genom tidernas ogunst. Herr Baronen var en av de allra mest lysande grandseigneurerna i Westfalen; hans slott hade port och fönster. Riddarsalen var till och med smyckad med tapeter.

A parser generally make predictions as to what is to come next as it goes along. Having processed a possessive pronoun in Swedish it expects to find another attribute or the bead noun. Having processed a noun it expects a relative clause of the finite verb if we are in the beginning of the sentence etc.

2.2. Semantic procedures

Lexical identification and syntactic parsing would still leave us with ambiguities in a case such as Swedish "Osten åt råttan", which could be a sentence where the cheese (osten) is the subject or a sentence where the cheese is the proposed (topicalized) object of eating. One way to solve such ambiguities is to use the knowledge of the relations between participants in actions: the "selectional restrictions" of Chomsky, the "case frames" of Fillmore or the "semantic templates" discussed by Wilks. The semantic templates would include knowledge of what is normally eaten, what objects can float, what objects can act and think intentionally, what objects can have colours etc. Finding out that the rat (rattan) must be the subject in this case also gives information about the topic-comment structure of the text. This piece of information is of interest in a translation system as most languages have ways of marking the topic.

Comprehension systems also have to keep track of the discourse (running text) as it may be referred to by deictic expressions, such as pronouns, definite nouns, adjectives such as other, same, previous. The information introduced gradually during the text has to be stored in some way before it is integrated completely with the general knowledge of (long time) memory.

How linguistic information should be stored if we are to simulate human behaviour 5 by no means clear, but something like semantic or cognitive networks may be needed. Semantic networks are interesting from several points of view - including translation - and I will show an example, the beginning of Voltaire's Candide, and make some comments (Fig. 3).

A semantic network is, roughly speaking, a set of related facts (relations, predications, propositions, atomic expressions) which may also be called a data base. A graphic display like the network of fig. 3 is more suggestive than a list because it avoids repetition of the same name (or index) and it presents the information as a map with no linear order imposed. In a diagram information about the same item is represented as arms (arcs) by the same item. Semantic networks have a long history as intuitive maps of knowledge, and they have been much discussed lately in psychology, computational linguistics, artificial intelligence, information retrieval. How should networks be constructed to allow operations of the same kind human beings master, when they search their memories, draw conclusions, answer questions, produce sentences or think?

A semantic network built gradually by the discourse must often be supplemented with knowledge previously stored. Some problems of ambiguity may be solved in this way. If we want to translate the sentence "The soldiers fired at the women and they fell" into a language such as French where the gender of the referent of they is relevant (ils or elles) syntactic parsing does not help (Wilks 1973). The knowledge necessary is rather knowledge of what generally happens to somebody who is fired at. A similar problem is the translation of Swedish "Bladen gulnade om hösten" (the leaves / pages grew yellow in the fall). Knowledge of natural processes during fall makes leaves the proper choice - the pages of a book would rarely change colours in the fall.

In the section of Candide presented the inference system is needed if we are to understand the irony of the words about the baron "was one of the most influential noblemen in Westphalia, for his house had a door and several windows and his hall was actually draped with tapestry". The knowledge that houses generally have doors and several windows gives this comment its full meaning.

The network Figure 3. only includes denotative (logical) meaning or information - not stylistic information. A fuller representation could include this information (for each relation) and the order of the presentation of the relations in the original text - one network could represent several texts with varying order of presentation. As a map of the beginning of the narrative the network diagram suggests characteristic features: important nodes with several arms (or arcs) are the young man Candide, his father and bis mother, the castle and the baron.

The information is stored as 2-place predicates or relations (equivalent to the property lists of LISP). The names of the predicates and their argument should be self-explanatory. Well-known names from grammar or logic are: CAUS (causing), POSS (possessing, "genitive"), AG (Agent), OB (object). The choice of primitive relations is only suggestive. The abbreviations are supposed to give associations to different languages. The network can be used for asking questions, e.g. "Is Candide related to the baron?" Looking for an answer is looking for a path along certain kinship relations. In this case there is a path MAN SUBGEN LADY COGEN BARON which gives the answer "Candide is the nephew of the baron" if translated into English (Sigurd, 1976b).

3. TARGET TEXT GENERATION

As mentioned above the crude and simplistic approach to machine translation is a word-for-word substitution process. A text with many alternatives is, however, unacceptable. The most sophisticated procedure would be a regeneration of the source text (recording) on the basis of a complete understanding of the source text in its cultural and social setting, keeping stylistic values etc.

A semantic network of the kind under discussion may be too poor or too rich as a basis of a target text - different languages require different semantic information. We will discuss the network presented as a basis for the English, French, and Swedish versions.

A translation is required to meet certain conditions, which we will touch on briefly in this connection.

1. The same informational (propositional, cognitive etc.) content as the source text (fidelity)

In our example this means that all the relations of the net and no others should occur in the target text. Relations which were inferred (and marked by dotted lines in the diagram) might be mentioned as well, and repetition of the same relation might also be allowed. But such changes may have stylistic effects.

2. Grammaticality and coherence of the target text (naturalness, beauty)

This requirement is often difficult to meet in automatic systems. Their mistakes are often horrible and most "unhuman". Human translators who are native speakers of the target language will generally fulfill this condition. Sentence grammaticality is perhaps easier to achieve than text coherence where various subtle means are needed (cf. Enkvist, 1973, Enkvist & Kohonen, 1976).

3. The same order of presentation as in the source text (thematic progression)

In our model this means traversing the relations in the same order as in the source text. Clearly the order of presentation must vary between languages - at least locally in the net - due to word order restrictions. The PROP relation would generally occur after the head in French surface structure, but before in Germanic languages. The POSS relation would occur as a de phrase after the head in French, before as an inflectional s or after as an of phrase in English etc. Differences in the order of presentation must be permitted and possible psychological and perceptional effects of this cannot be avoided (cf. Enkvist & Kohonen, 1976).

The relation BARON POSS CASTLE is placed after castle in the surface structure all the three versions. The Swedish version uses tillhörigt instead of the ordinary genitive construction with inflectional s - this allows the baron to be introduced as an indefinite noun.

The relations between the semantic relations of the net and surface structure may be complex and it is not always possible to determine the order of presentation. The negation in French n e ... ja ma i s circles the exponent of the VOL relation (v o u - lut). Relations may also mix into words - word-building processes may operate on the net. The relation LADY PROP YOUNG is represented by one word only in French (demoiselle) and Swedish (fröken).

The order of presentation is not followed in free translations - they are generally also characterized by differences in the use of pronouns and definite expressions, the division into clauses etc. It is fairly easy to generate several texts from the same network and computer programs for that purpose have been developed. If taken at random the string of relations will make up an incoherent text, but a computer program can easily connect relations dealing with the same thing and produce coherent texts. It is, however, difficult to control the output in detail and the computer text may easily be very boring or monotonous.

The original order of presentation in the three texts is: time and location of the hero, explanation of his name, his family relations, the importance of the baron. The beginning of Candide would be retold e.g. in the following ways:

> Thunder-ten-tronckh
> was the name of a baron living at a castle in Westphalia. He had a sister, whom the old servants ...

or starting at Candide's face:

> Candide's face reflected
> the pious mind that nature had bestowed on him ...

or starting at Candide's father:

> A good and honest nobleman lived close to a castle owned by ...

If continued such versions could very well meet the requirement that the informational content be the same and also the grammatical requirement. It is, however, dubious whether it would be easier to succeed in automatic translation if the requirement of identical order of presentation is slackened.

4. The same distribution of the relations of the net on syntactic modules such as clauses, phrases, words (verbalization).

The three versions under study almost meet this requirement. We nave noted some differences and we will observe that the relation AUKT CRED is rendered as one word (sannolikt, probably) in the Swedish version. If the text is started at another node than the source text it is often impossible to avoid differences in the verbalization.

The relations of a network are not equally well suited to occur as sentences and it is more natural to traverse a relation in one direction than in another (the converse) (cf. Sigurd, 1976b). The PROP relations generally reduce to attributes or adverbials, VOL relations reduce to auxiliaries, POSS relations reduce to prepositional or inflectional phrases (the genitive) in the languages under discussion. The content is thus packed and compressed differently in the verbal modules offered by the language.

There are many choices to be made and many automatic consequences to control when a text is produced. Stylistic values depend heavily on the choices and their consequences. Computer programs are by no means able to mirror the selective verbalization processes of a versatile human translator.

5. Further requirements of a translation could be: the same number of words, the same number of syllables, the same number of stressed syllables, the same last stressed vowel and following consonants (rhyme). The search procedures necessary to find equivalents that meet such requirements are not available to computers and literary or poetic translation is consequently not possible.

The greatest variation in the three texts under consideration occurs in the rendering of the moral qualities of Candide. The Swedish version is the most elaborate. If rendered in English it is: "His tongue was as straightforward as his heart was unsophisticated". (The French version is by René Louis Doyon, it is not Voltaire's original version.) Automatic systems will never be able to mirror the creative selection of expressions shown in the three versions.

4. CONCLUSION

We will certainly experience many new computer systems handling information and language in the future. Even now we are being told by computers about our salaries, tax deductions, bank accounts, driver's licences. But I think translation requires simulation of mental processes too sophisticated and too little understood to be handled by computers at least of this generation. High quality translation will probably remain an activity uniquely human for ever.

REFERENCES

Borko, H. (ed.), 1967: Automated Language Processing. New York: John Wiley and Sons, Inc.

Enkvist, N. E. ,1973: Linguistic Stylistics. The Hague: Mouton (Janua Ling.)

Enkvist, N.E. and Kohonen, V., (eds.) 1976: Reports on Text Linguistics: Approaches to Word Order . Åbo: Åbo Academy

Hays, D. & Mathias, J. (eds.), 1976: "FBIS Seminar on Machine translation. American Journal of Computational Linguistics, microfiche 46

Herskovits, A., 1973: "The Generation of French from a Semantic Representation". In: Advanced Research Projects Agency No 457, Standford University Artificial Intelligence Laboratory Memo AIM 212

Locke, W.N., & Booth, A.D. (eds.), 1955. Machine Translation of Languages . New York: Technology Press of M.I.T.

Sågvall, A. -L., 1973: "A System for Automatic Inflectional Analysis, implemented for Russian". In: Data Linguistica 8. Uppsala

Schank, R. and Colby, K.M. (eds.), 1973: Computer Models of Thought and Language. San Francisco: W.H. Freeman and Company

Sigurd, B., 1976a: "Understanding compounds and Salvador Dali". In: Hovdhaugen (ed.), Proc. of 2nd Conf. Scand. Linguistics. Oslo

Sigurd, B., 1976b: "Converse terms in Swedish", In: Enkvist & Kohonen, 1976

Vauquois, B., 1967: "Le système de Traduction Automatique du CETA. " Document C.E.T.A.

Welin, C.W., 1976: "En parser för parallell syntaktisk och semantisk analys". In: Papers from the Institute of Linguistics University of Stockholm 31

Wilks, Y., 1973: "An Artificial Intelligence Approach to Machine Translation". In: Schank and Colby, 1973

Woods, W.A., 1969: "Augmented Translation Networks for Natural Language Analysis." In: Report Cs-1, Comput. Lab. Harvard University, Cambridge Mass.