[From: The analysis of meaning: *Informatics* 5 (Aslib 1979)]

Theoretical options and practical limitations of using semantics to solve problems of natural language analysis and machine translation

Ian M. PIGOTT, Commission of the
European Communities,
Luxembourg

Introduction

Within the framework of its Multilingual Action Plan, the Commission of the European Communities has, for the past three years, been involved in the practical development of a machine translation system (Systran, designed by Peter Toma, World Translation Center, La Jolla, California). Of the language couples covered to date, the English-French pair is certainly the most highly developed, yet it may well be that ultimately the quality of translation obtained from the other systems under development (French-English and English-Italian) will be more acceptable.

Prom the evaluations conducted on the English-French system, it is clear that MT is indeed a viable alternative to human translation for certain applications although in most cases the raw machine output requires a certain amount of post-editing. The evaluations have also shown that substantial improvements in quality have been achieved and indicate that even higher standards will be reached as development work continues.

Encouraged by these results, the Commission has recently been working in close cooperation with experts from the various Member States on plans for an even more efficient MT system, Eurotra. There is reason to believe that by making full use of the experience gained from Systran and the know-how which already exists in the many linguistic research centres across Europe, output from this new system - if approved - will be substantially better that what we are now able to produce.

However, as the aim of this paper is to present the results of practical experience rather than to expound on the relative merits of new approaches, I shall attempt to describe how the Commission's MT team, working hand-in-hand with the system's designers in California, has adapted and developed the semantic features of Systran to provide a suitable basis for dealing with many of the problems of natural language analysis and machine translation. The English-French system will be used for purposes of illustration, not only in view of today's English-speaking audience but rather because English, in view of its very limited syntax and tremendous flexibility is perhaps among the most difficult of all languages to analyse by means of logical, computerized techniques.

Finally, it is to be hoped that the findings contained in this paper will help to provide a better understanding of some of the linguistic aspects of language analysis while pinpointing a number of major problems experienced in MT development work. Many of these are certainly not specific to the actual system used (in our case Systran) and will thus doubtless require careful attention in future developments.

1. Brief description of the system

This is hardly the place to give a full description of the Systran translation system, particularly as several articles have already been devoted to the matter. However, in order to provide adequate details of the various levels at which semantics play a role in the overall process, it would appear essential to give a general overview of the various stages in the translation process.

The system itself may reasonably be seen as consisting of two fairly independent components, on the one hand the dictionaries which contain both information about meanings and data concerning the "behaviour" of lexical items at the source and target levels, and on the other a set of programs which draw on the information contained in the dictionaries to carry out a whole series of operations at the various stages of analysis (source), transfer and synthesis (target) processing.

1.1 Dictionaries

At the risk of generalization, the dictionaries may be considered to fall into two distinct groups:

a) one-word dictionaries which give all applicable details of the morphology, part of speech, gender, number, person, time, homography and semanto-syntactic characteristics of each word in the source language as well as a basic translation in the target together with any supplementary information required about the behaviour of the target equivalent (e.g. part of speech, morphology, requirements when governed by or governing other words);

b) multi-word dictionaries, the purposes of which are two-fold:

   i) to limit the possible functions of a word in context (e.g. by specifying that in a given expression a homograph is to be resolved as a noun rather than a verb);

   ii) to provide the specific meaning of words in context (either when forming part of a set expression or when in grammatical and/or contextual relationships with other word(s) or word types).

1.2 Programs

On the basis of information obtained from dictionary look-up, a series of analysis programs is used to parse the source language sentence. These may be resumed as follows:

a) homograph resolution based on a series of contextual tests;

b) clause boundary definition aimed at dividing the sentence into main and subordinate clauses on the basis of punctuation, conjunctions, relatives and semantic information;

c) establishment of primary syntactic relationships between nouns and their modifiers, verbs and their objects, words governing infinitives or gerunds, etc.;

d) establishment of secondary syntactic relationships such as enumerations (nouns, adjectives or verbs functioning in parallel), frequently on the basis of semanto-syntactic information;

e) identification of the subject(s) and predicate(s) of finite verbs.

The programs situated at the transfer stage aim primarily at using the results of analysis to resolve various linguistic peculiarities of the target language, either by means of routines written on or around specific words or word classes or simply by making use of complex dictionary information. In many cases, semantic categorization is used to call programs at this level. Target meanings are supplied where appropriate.

Finally, at the target level a synthesis program serves to generate the appropriate inflected form of each word on the basis of the requirements of the target language while a rearrangement program establishes the correct sequence of words and/or phrases in the sentence.

2. Semantic characteristics of the system

It would be unrealistic to claim that the semantic component of the Systran system is independent of the basically syntactic approach to analysis. However, while in many cases surprisingly good results can be obtained from analysis based on the fundamental characteristics of each word (part of speech, homograph type, gender, number, person, tense, etc.), more often than not additional information of a semanto-syntactic nature is required to indicate the probable way in which a word behaves in a given environment.

This semanto-syntactic information, which is widely used in the entire translation process from analysis onwards, is contained in markers of two inherently different types, those which provide information about grammatical government and those which give ah idea of the basic semantic characteristics of a word.

Typical examples of the grammatical government markers are those which concern the transitivity of a verb (e.g. usually transitive, always intransitive), the possibility for an adjective to be used impersonally (as in "It is possible that ...") or the ability of a noun to govern an infinitive ("Any attempt to deal with this problem").

In regard to basic semantic characteristics, there are markers which indicate whether a noun is abstract or concrete, whether a verb normally requires a human subject or whether an adverb relates to time, manner or place.

All in all, there are 70 markers of this type, a complete list of which will be found in Annex I. The following table shows how the two types (grammatical government and basic semantic characteristics) are distributed among the various parts of speech to which they may be applied:

Marker Type	Verb	Noun	Adj.	Adv.	Conj.	Total
Gram govt.	18	3	5	2	17	45
Bas. sem.	6	10	1	8		25

These statistics clearly indicate that a variety of markers are required to describe the potential government requirements of verbs, adjectives and conjunctions whereas in the case of nouns and adverbs, information about the basic semantic characteristics predominates. Suffice it to say, at this stage, that the main reason for this is that nouns and adverbs rarely have any obvious government requirements whereas verbs and conjunctions nearly always do.

In addition to the basic markers described above, many of which are essential to basic analysis as will be illustrated later, the system also contains some 450 semantic primitives (and the capacity to accommodate over 1000 more) which were originally designed to provide information about subject fields or sectors. Of these, some 20 have been found to be particularly useful either as an aid to analysis (particularly for resolving enumerations) or as a basis on which to introduce routines at the transfer stage. (A typical alphabetical sample from the full list as well as those frequently used by the Commission will be found in Annex II).

3. Theoretical possibilities vs. practical limitations

3.1 Semanto-syntactic markers

From the above it can be seen that a very wide variety of semantic codes of various types are available to the Systran lexicographer for dealing with situations which cannot be handled by reference to pure syntax. Indeed, as the coding manuals provide little or no guidance on the relative usefulness of the various markers which

may be attached to any lexical item, the lexicographer or dictionary coder must decide himself which information he feels he should use to document the semanto-syntactic behaviour of a given word or expression.

Experience has shown that initially a coder will either be over-conscientious and add a whole series of (often conflicting) information to a given term or, recognizing the fact that it is extremely difficult to envisage all possible situations in which a word is liable to occur, he will tend to be excessively cautious and add only such codes he feels are essential.

In the first instance, when required to code the noun FISH the conscientious coder might well produce an entry containing the following markers:

HU (human) as in "he's a strange fish".
AN (animate) "the fish swims quickly".
AMB (animate/inanimate ambiguity)
CON (concrete)
CT (countable) "several fish(es)"
MS (mass) "a lot of fish was eaten".
GI (govern infinitive) "I have some fish to sell".
NAP (noun clause in apposition) "The fish that I bought yesterday ..."
GG (noun + prep, can govern gerund) "Fish for stocking reservoirs"

At the semantic primitive level, he may well go on to add:

AGRIC (agriculture)
BIO (biology)
CONG (concrete)
IMPERS (impersonal)
FPROD (food product)
NUTRI (nutrition)
PRDCT (product)
SUBST (substance)
ZOOL (zoology)

The more wary coder might either decide to add no information at all (in which case he will be sure he has not introduced any errors) or might choose CON (concrete) as the only really reliable marker he can use.

Unfortunately, both these approaches would be equally unsatisfactory (although the second would seem to be preferable to the first).

In the first case, obvious errors of judgment were made:

HU - the idiomatic use is extremely rare
GI, NAP, GG - the coder's understanding of basic grammar leaves much to be desired

Furthermore, some of the information seems superfluous (AN, AMB). In regard to the semantic primitives, however, the coder can hardly be said to have made any obvious errors.

In the second case, the omission of certain types of information could well create problems, particularly - as will be explained later - as various programs can, and have been designed to look for the presence (or indeed absence) of certain markers.

Finally, and perhaps rather obviously, the fact that two different coders are liable to enter two quite different sets of data raises the problem of lack of compatibility between markers on items which behave in identical ways. Indeed, unless data attached to similar lexical items are compatible, it is extremely difficult to develop programs to deal with those situations which cannot be handled by pure syntax alone, since such programs can only function if certain criteria are satisfied.

A straightforward method of guiding the coder in his selection therefore had to be developed, both to ensure the inclusion of essential markers and to avoid the use of superfluous data which could do more harm than good. After a certain amount of trial and error, it was found that in nearly all cases coders could be trained to use the same basic markers by considering what answers they would give to a series of questions.

When coding nouns, for instance, in order to choose the best combination of concrete/abstract and countable/mass codes, the coder was always required to select one and only one of each pair on the basis of the following criteria:

- Can you touch it? Yes → Concrete
  No → Abstract

- Would you say "much xyz" rather than "many xyz's"
  Yes → Mass
  No → Countable

- If in doubt, does the plural form of the word (xyz's) have exactly the same meaning as the singular (xyz).
  Yes → Countable
  No → Mass

Use of these criteria enabled the coder to settle any doubts he might have had about which codes to use.

The word PAPER could be considered to have all four attributes, but in answer to these questions only two would be chosen (concrete and mass). Compatibility at this level was thus assured.

At some levels, however, it proved much more difficult to provide clear selection criteria which would be used consistently by all. In the case of verbs, although it was usually possible to train coders to select the appropriate transitivity code (usually transitive, usually intransitive, always intransitive) - and here it might be added that the likelihood of a verb being "always transitive" was so slight that use of this code was strongly discouraged - it was almost completely impossible to draft any reliable guidelines in regard to the subject/object codes (weightings in favour of animate, human or inanimate subjects or animate, abstract or concrete objects). There were two major difficulties here. Firstly, sheer experience showed that most verbs could and did have subjects and objects of all the various types. Secondly, even verbs which might be considered to require human or animate subjects (read, write, look, etc.) very frequently occurred with inanimate subjects:

"The paper reads ...."
"The pen writes....."
"The situation looks promising".

Finally, it was found that even in the absence of these codes, there was seldom any problem of subject/object identification at the analysis level and that variations in meaning could be handled at other levels such as by using semantic primitives. It was therefore decided to discontinue the use of these codes except in extremely specific cases.

There is no point is giving details of all the other reasons why certain semanto-syntactic codes were found to be more useful than others. Annex I does however differentiate between three types: those considered extremely useful, those which are essential for solving particular problems but which are not generally applicable and those which have been found to be of little practical use.

3.2 Semantic primitive markers

Discovering the usefulness of semantic primitives proved to be an even longer process of trial and error than in the case of the semanto-syntactic markers. Again, very little explanation was given by the system designers on the way in which the hundreds of semantic codes in the system were to be used. It was not clear even whether the codes were intended for use with any part of speech simply to provide information about subject field or whether they had been designed to solve those specific problems of analysis or translation which had been identified in the development of the original Systran Russian-English system.

The result of this situation was that initially very little use was made of semantic primitives, time and effort being devoted almost

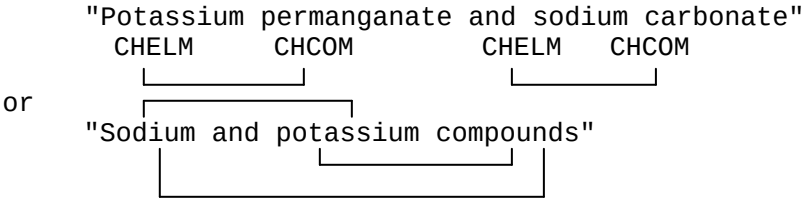
exclusively to attaining consistency in the use of semanto-syntactic markers. This was probably just as well since it enabled us to see to what extent it was really necessary to introduce this second level of semantic marking.

3.2.1 Generalized markers

The first problem to be identified was the failure of the system to recognize enumerations in certain cases, particularly with nouns carrying differing semanto-syntactic markers. In the field with which we were concerned (food technology), this problem occurred frequently in regard to chemicals and foods. A structure of the type

"Production of bread and cakes in rural areas"

would be analysed incorrectly simply because of the different markers on BREAD (mass) and CAKE (countable) with the result that PRODUCTION would only govern BREAD and not CAKES. By using the same semantic primitive code FPROD (food product) for both terms (and indeed all other foods) correct enumeration could be achieved at the analysis stage. Similarly, the codes CHCOM (chemical compound) and CHELM (chemical element) were used to establish enumerations between chemicals:



i.e. compounds of potassium and of sodium.

It was found that these codes could be introduced consistently by the various coders either when entering a dictionary item for the first time or by modifying existing data where necessary.

While certain markers of this type were used exclusively for nouns, an attempt was also made to solve verbal enumeration problems by attaching primitives such as AGPRO (agricultural processing) to verbs and verbal nouns. Here, though, it was a little more difficult to define exactly what constituted a verb of agricultural processing for whereas there could be little doubt that HARVEST, PLANT, FERTILIZE, etc., were specifically agricultural, there was a tendency to include verbs of a very general type in this group such as PROCESS, PRODUCE, etc., simply because they happened to occur in enumeration with agricultural verbs. Success was therefore rather limited although some specific problems could be solved by using this primitive on verbs.

In the case of verbal nouns it was found that enumeration could frequently be established by using this semantic primitive (AGPRO). However, simply because many of the verbal nouns encountered were not of the agricultural type, there was a tendency to use markers

which seemed more suited to the field in question. For example ANALYSIS might be coded ANTEC (analytical technique), ACIDIFICATION, PRCH (chemical process), PREFILTRATION, PRTECH (technical process), etc.; indeed, any one of the following "process type" primitives contained in the system could be used depending on the coder's subjective choice:

> AGPRO (agriculture), ANTEC (analysis),
> PRAVIA (aviation), PRBIO (biology), PRCH (chemistry),
> PRCR (creative), PREL (electrical), PRELR (electronic)
> and some 20 more (see Annex II).

Yet, the very fact that verbal nouns from (subjectively) different fields happened to occur in parallel relationships (e.g. "the production, sterilization and atomic absorption analysis of organic inpurities") caused breaks rather than links in enumerations if the primitives used were not the same.

Finally, it was realized that the true need for having a primitive at this level was simply to establish the fact that a word happened to be a verbal noun and that the specific technical field(s) in question had little, if any, effect on analysis or translation. For this reason, it was decided that the code PRGEN (general process) should be attached to all verbal nouns and that the specific subject field codes should only be used in conjunction with this for cases requiring special treatment.

A similar approach evolved in connection with nouns of property: it was found quite sufficient to use one primitive to cover any type of property irrespective of subject field. Thus WEIGHT, APPLICABILITY; TEMPERATURE, SOLUBILITY, etc., would all receive the same code. Generalization of this type had several advantages: speed, in that the coder did not have to waste time wondering which particular subject field he was dealing with, consistency owing to the more extensive coverage offered by each code, economy of effort resulting from the possibility of writing routines applicable to all words carrying a given general marker rather than duplicating these for various more specific markers and, last but not least, more accurate analysis facilitated by improved establishment of parallel structures.

Other primitives found to be particularly useful at this general level were DEV (device - any tool, instrument or piece of equipment used to facilitate production or operation, CONTNR (container), MATER (material or substance used for production or operation), SCINO (for the name of any subject field or sector) and PROF (professions).

Quite apart from the usefulness of these markers at the analysis level, it was found that they could be used to great advantage at the transfer level as a basis on which to build dictionary routines. Information from the analysis could be used in conjunction with a marker to provide special meanings or translations. A typical example of this would be preposition requirements of a given semantic category. IN when governing the name of a subject field (chemistry, geo-

logy) could be rendered EN (rather than DANS) in French for all words in the category (en chimie). In certain cases, WITH governing a device (hammer, switch, etc.) could be translated A L'AIDE DE rather than AVEC in the interests of elegance. Similarly EMPLOY would require the translation EMPLOYER rather that UTILISER when its object is a profession (engineer, secretary, etc.) There is practically no limit to the amount of analytical information which can be combined with semantic primitives to bring about general changes in meaning. Several examples are given in Annex III.

3.2.2 Specific markers

The arguments expressed above in favour of generalizing the use of certain semantic primitives certainly do not apply in all cases. There are certain categories of proper nouns which not only require special treatment at the analysis level, but also raise problems of translation. Typical examples of these are place names, particularly the names of towns and countries, and the months of the year.

In the first case, place names, recognition of addresses is important in analysis. Moreover, in translating into the target language (French) special article and preposition requirements have to be satisfied independently for towns and countries.

It was therefore decided to make use of two levels of semantic primitives for words of this type, on the one hand a fairly general code GEOLOC (geographic location) to be coded with all place names, including towns and countries, liable to occur in addresses, and on the other more specific codes CITY (towns, cities) and COUNTR (countries, provinces) to be coded as applicable.

Lexical routines could then be written around these semantic markers to process articles and prepositions as required. For example:

"In France, Canada and the United States"

would be translated

"En France, au Canada et aux Etats-Unis"

rather than

"Dans la France, le Canada et les Etats-Unis"

which would otherwise have been the translation.

By using the semantic code MONTH with each of the twelve months of the year (and all their various abbreviations), it was also possible to develop special routines for recognizing and translating dates.

"On 1st and 2nd December 1979"

would become

"Les ler et 2 décembre 1979"

and most other date structures could be handled satisfactorily.

4. System limitations

In Section 3, a number of typical examples have been given of how semantics can be used in the system as it now exists to assist analysis and provide the desired translation at the target level.

It is only fair to point out, however, that the semantics component of the software package is not as yet powerful enough to deal with a number of easily recognizable structures which, from the theoretical point of view, could be identified and resolved by the use bf semantics.

4.1 Analysis

At this stage of development, the use of semantics in analysis is restricted mainly to grammatical parsing (at the semanto-syntactic level), to establishing enumerations (frequently at the semantic primitive level) and to developing routines for the resolution of certain types of homograph (at both levels).

What has not yet been attempted, but already seems perfectly feasible, is the use of semantic primitives in creating affinities between different parts of speech where obvious government requirements exist. This approach doubtlessly bears similarities to the templates or paraplates proposed by Prof. Wilks; however, as far as the Commission's development of Systran is concerned, it is interesting to note that the need for a feature of this type emerged independently as a result of practical experience.

A simple example of the need to establish part of speech affinities is to be found in the case of adjectives and nouns of chemistry. In the absence of special lexical entries involving specific words, the phrase

"high sulphuric acid content"

would be analysed as if the first three words were all adjectives qualifying CONTENT (i.e. the content is high, sulphuric and acidic). Using the affinity approach, it would be possible to use a primitive to identify the fact that ACID is a noun of chemistry (the CHCOM code already in use could well be sufficient) and that SULPHURIC is an adjective of chemistry (CHADJ). The analysis passes could then be programmed to detect potential affinities of this type and establish the relationship required. In this case the analysis would then give

"high content of sulphuric acid"

and would produce a correct translation without any special lexical entries.

Similar affinities could be established between subjects, verbs and objects (e.g. verbs of payment would choose nouns or noun phrases of things paid - taxes, duties, funds, etc. - as their objects in cases of ambiguity).

4.2 Text typology

One of the major problems identified in developing a machine translation system for the translation of a wide range of document types, is the inability of the programs to detect the level of language used in the source document and, by extension, the most suitable style and vocabulary required in translation.

An obvious example of this can be seen in the requirements for translating minutes from English into French when past tenses often need to be transformed into present tenses in the target and certain formulations (e.g. "in the chair") take on new meanings. Other examples of text types requiring special treatment are letters, speeches, patents, regulations, calls for tender, sets of instructions, abstracts and forecasts.

Although typology could possibly be specified manually at the encoding stage, a far more satisfactory approach would be to provide for automatic recognition of language level. This would have the added advantage of being able to switch from one level to another within a given text as required.

It is reasonable to suppose that automatic recognition could be triggered by attaching semantic markers to certain words or groups of words which are considered to be typical of a given type of document or level of language. The presence of certain personal pronouns (myself, yours, etc.) or of colloquial forms (don't, isn't) could be used as a general basis on which to distinguish between formal and free style while items in titles (Regulation No., Minutes, Appendix) or in the body of the text (Dear Mr, by virtue of Article, In reply to) could be used to identify the type of document under consideration.

This approach could possibly be extended in turn to enable the subject field(s) to be identified although this could prove far more difficult to handle reliably, if only because most documents cover three or more fields (e.g. agriculture , economics, Community interests, etc.).

If such automatic identification of type or field could be achieved, it would be a fairly simple matter to modify the system at the programming and dictionary levels and so ensure the correct level of translation.

5. Conclusions

5.1 Selection of markers

(a) A semantic component as such is of little use until such time as clear definitions can be drafted and applied in practice.

(b) In order to ensure consistency in the use of semantic markers at all levels, selection criteria must be established in such a

form that they are interpreted in the same way by different coders.

(c) A distinction should be made between generally applicable markers, which should be as few as possible, and specific markers for dealing with particular semantic or semanto-syntactic requirements.

(d) Generalization of marker types should be undertaken wherever possible in the interests of speed, consistency, and downstream reliability and economy of effort.

5.2 Effectiveness of semantic treatment

(a) At the analysis level semanto-syntactic markers are used successfully to resolve the government requirements of various parts of speech while semantic primitives provide a means of establishing enumerations.

(b) At the transfer and target levels, semantic primitives serve as a basis on which to compile lexical rules for inserting special meanings as required by context and, in certain cases, as a means of dealing with more intricate problems of translation such as dates and addresses.

5.3 Future prospects

(a) It would appear feasible to extend semantic analysis to provide for the affinities required between different parts of speech.

(b) Semantic markers could probably also be used as a means of defining document typology and/or level of language in order to meet the stylistic requirements of the target language for different types of text.

Annex I - Semanto-syntactic Markers

Listed below are the various semanto-syntactic markers contained in the Systran system, classified by the parts of speech to which they apply. A further breakdown, where applicable, is made between markers providing information about grammatical government and those relating to basic semantic characteristics. The figures in brackets give an indication of the relative usefulness of each marker: - (l) = extremely useful, (2) = essential for solving specific problems, (3) = of little practical use.

Verbs

Grammatical government:

UTRAN (l) - usually transitive (verb will usually have an object) Most verbs in English come within this category.

UINT (l) - usually intransitive (verb rarely has a direct object - only in special cases will a potential object be analysed as such)

AINT (l) - always intransitive (verb never has a direct object - potential objects will be resolved as having another syntactic function)

N.B. Only one of the above transitivity codes may be selected.

GI (l) - can govern an infinitive (e.g. remember to come)

LINK (l) - can govern a predicate adjective (e.g. it appears useful)

NCO (l) - can open a noun clause (e.g. he said he was coming)

NMR (l) - present participle seldom functions as an adjectival modifier and will normally be analysed as a gerund (e.g. redefining criteria - the redefining of criteria) This code has been used extensively, and to great effect, in combatting the " -ing" problem in English.

GOI (2) - can govern direct object plus infinitive (e.g. I ordered him to come)

GG (2) - can govern present participle (e.g. avoid doing something)

GOG (2) - can govern direct object plus present participle (e.g. I heard him coming)

GOO (2) - can govern two direct objects (e.g. they elected him chairman)

GOA (2) - can govern direct object plus adjective (e.g. he rendered it useless)

GONC (2) - can govern object plus noun clause (e.g. I warned her I was leaving)
Introduced recently for handling this type of structure which is fairly common in English

CLAN (2) - present participle frequently takes an entire clause as antecedent (e.g. the law was passed resulting in additional benefits to farmers)

TENS (2) - present and past tense identical (e.g. put).
By reference to other verbs, enables the correct tense to be selected.

ATRAN (3) - always transitive (verb must have a direct object - otherwise it will usually be resolved as a participle or infinitive)
Was used extensively in the past but has now been largely abandoned as nearly all seemingly transitive verbs often occur in text without direct objects.

Basic semantic characteristics

MOTN (2) - verb of motion.

Useful in resolving adverb/preposition homographs.

INSUB (3) - inanimate subject

ANSUB (3) - animate subject

HUSUB (3) - human subject

ANOB (3) - animate object

ABSOB (3) - abstract object

CONOB (3) - concrete object.

N.B. These subject/object codes are rarely used for reasons explained in the paper.

Nouns

Grammatical government

GG (2) - noun plus preposition frequently governs gerund (e.g. method of writing reports).
Also used to great effect in resolving the "-ing" problem (cf. NMR on verbs).

GI (2) - can govern infinitive (e.g. his decision to come tomorrow)

NAP (2) - can be followed by noun clause in apposition (e.g. the fact that it is difficult ...)

Basic semantic characteristics

CON (l) - concrete (i.e. tangible)

ABS (1) - abstract (i.e. intangible)

CT (1) - countable (plural form has same basic meaning as singular)

MS (1) - mass (partitive characteristics or plural meaning not equivalent to that of singular)

N.B. Combinations of the above codes are used with nearly all nouns as explained in the paper.

HU (2) - human (useful in some cases as a basis on which to introduce special meanings for verbs in context).

QUAN (2) - quantity (e.g. a pound of butter). Useful in identifying prepositional government.

TP (2) - time period (e.g. day, morning) Enables nouns to form the basis of an adverbial phrase where appropriate (e.g. he telephoned this afternoon)

AN (3) — animate

AMB (3) - animate/inanimate ambiguity

GRP (3) - collective noun.

N.B. The last three are seldom used in the absence of dependable selection criteria.

Adjectives

Grammatical government

IMPA (2) - impersonal adjective: "it" when the subject of a LINK verb (see above) followed by this type of adjective is likely to be impersonal (e.g. It is impossible to calculate its effect)

GI (2) - can govern infinitive (e.g. He was happy to come)

GG (2) - can govern gerund (e.g. He is capable of writing reports)

APHI (2) - can initiate an adjectival phrase (e.g. the warning implicit in his remarks)

Basic semantic characteristics

AHAD (3) - only modifies animate nouns
Not used, as no adjectives of this type exist.

COMER (3) - adjective forms comparative with -ER.
Not used, as this characteristic has no influence on analysis or translation.

N.B. More often than not, it is unnecessary to use any of these codes with adjectives as they seldom have any of the attributes covered.

Adverbs

Grammatical government

ADVVB (3) - can modify verb.
Seldom used as nearly all adverbs can modify verbs.

ADVADJ (3) - can modify an adjective or another adverb.
Used only in a few exceptional cases where an adverb is unlikely to modify a verb (e.g. extremely). In many cases, however, this potentiality is better covered by DEG (see below).

Basic semantic characteristics

DEG (2) - degree (e.g. approximately, completely)
Useful in some cases for establishing the affinity with an adjective (completely white light) or figure (almost 10%).

FUT (2) - future time (e.g. tomorrow).
Useful in resolving tenses at the target level.

Also : TI (time),PL (place), MA (manner), FREQ (frequency), DIR (direction), none of which have been found particularly useful in English-French or English-Italian translation but which would no doubt be required for target languages with a rigid adverbial structure such as German.

In the case of adverbs too, more often than not the above markers are of no practical use.

Conjunctions

Grammatical government

The semanto-syntactic markers attached to conjunctions provide information about the potential functions such words have as clause openers. This information is particularly useful is establishing clause boundaries in the first analytical pass as well as in establishing relationships for target agreements. The following types can be specified:

Relative pronoun (which), noun clause opener (whether), interrogative pronoun (how), restrictive conjunction (as far as), time conjunction (before), generalizing conjunction (whatever), comparative conjunction (than), causal conjunction (because), conjunction of purpose (in order that), conditional conjunction (if), concessive conjunction (although), concurrent time conjunction (while) and coordinate conjunction (but).

Annex II - _Semantic primitives

The semantic primitives listed below fall into two groups: generalized markers and markers specifically introduced to deal with particular problems of translation. In nearly all cases, the codes are used with nouns only. All these primitives have two basic functions: to help esta-Rblish enumerations in analysis and to provide a basis on which to intro-Iduce special lexical entries supplying the appropriate translation or meaning in the target language.

I Generalized markers

PRGEN - general process (e.g. translation, use)
Used with all verbal nouns irrespective of subject field to establish enumeration (often with gerund types) and achieve correct article resolution.

PRPPHY - physical property (e.g. colour, possibility)
Now extended to cover properties of all types.

MATER - material (e.g. wood, oil)
Any material or substance whether used as a fuel or as a raw product.

DEV - device (hygrometer, heater)
Any piece of equipment or instrument.
Particularly useful for target preposition requirements.

CONTNR - container (bottle, reservoir)
Any type of receptacle.

SCINO - science nomenclature (chemistry, economics)
Used for the names of all fields.
Useful for article and preposition requirements at target level.

MU - unit of measure (pound, kilometre)

UNABR - abbreviation of unit of measure (1b, km)

DUR - duration (hour, year)
Often used in conjunction with TP (time period) nouns for preposition requirements.

PROF - profession (secretary, accountant)
Mainly for establishing enumerations.

Semi-specific markers

CHCOM - chemical compound (hydrolysate, sulphide)

CHELM - chemical element (zinc, hydrogen)

TRANSP - transport (ship, car)
Any vehicle.

FPROD - food product (sugar, wine)
Anything that may be eaten, with or without further processing.

FINAN - finance (tax, duty)
Anything that requires payment.
Used mainly to ascribe special meanings to verbs.

AGPRO - agricultural processing (harvesting, fertilization)
Now largely replaced by PRGEN but still used in some cases for verbal enumerations.

ENPRIS - enterprise (Commission, United Nations)
Used with the names of institutions and authorities.

GEOLOC - geographic location (Rome, Danube)
Often used in combination with CITY, COUNTR (see below) for identifying addresses.

Specific markers

MONTH - month (January, Feb.)
Used to identify date structures

CITY - city (London, Paris)
Extremely useful for dealing with article and preposition requirements.

COUNTR - country (United Kingdom, France)
Also used extensively for article and preposition requirements.

Typical sample from the alphabetic list of semantic primitives contained in the Systran system (few of which have been used)

PRDIM	PROPERTY (DIMENSION)
PRDTN	PRODUCTION
PREL	PROCESS (ELECTRICAL)
PRELR	PROCESS (ELECTRONIC)
PREP	PREPARATION
PREXP	PROCESS (EXPLANATORY)
PRFI	PROFESSIONAL FIELDS
PRGEN	PROCESS (GENERAL)
PRIOR	PRIVATE INDUSTRY-BANKING
PRLIT	PROCESS (LITERARY)
PRLOG	PROCESS (LOGICAL)
PRMATH	PROCESS (MATHEMATICAL)
PRMECH	PROCESS (MECHANICAL)
PRMENT	PROCESS (MENTAL)
PRMET	PROCESS (METEOROLOGICAL)
PRML	PROCESS (METALLURGICAL)
PROB	PROBLEM
PROF	PROFESSION/TITLE
PROPT	PROCESS (OPTICAL)
PROPUL	PROPULSION
PROX	PROXIMITY
PRPARC	PROPERTY (AERONAUTICS)
PRPAT	PROPERTY (ATMOSPHERIC)
PRPAC	PROPERTY (ACOUSTICS)
PRPBIO	PROPERTY (BIOLOGICAL)
PRPCH	PROPERTY (CHEMICAL)
PRPELC	PROPERTY (ELECTRICAL)
PRPHY	PROCESS (PHYSICAL)
PRPHYC	PROPERTY (HYDRAULICS)
PRPINN	PROPERTY (INNATE)
PRPM	PROCESS (PHYSIOLOGICAL/MECHANICAL)
PRPMAT	PROPERTY (MATHEMATICAL)
PRPMEC	PROPERTY (MECHANICAL)
PRPMET	PROPERTY (METEOROLOGICAL)
PRPMIN	PROPERTY (MINERAL)
PRPML	PROPERTY (METALLURGICAL)
PRPOPT	PROPERTY (OPTICAL)
PRPPHY	PROPERTY (PHYSICAL)
PRPPSI	PROPERTY (PHYSIOLOGICAL)
PRPRE	PROPERTY (RESEARCH)
PRPRR	PROPERTY (RADIATION/RADIOACTIVITY)
PRPSCI	PROPERTY (SCIENTIFIC)
PRPSI	PROCESS (PHYSIOLOGICAL)
PRPTEC	PROPERTY (TECHNICAL)
PRPTM	PREPOSITION OF TIME
PRRE	PROCESS (RESEARCH)
PRTC	PROCESS (TIME CONSUMING)
PRTECH	PROCESS (TECHNICAL)
PSYCH	PSYCHOLOGY
PUB	PUBLICATION
QUAL	QUALITY
QUANT	QUANTITY

Annex III - Dictionary Samples

The three following pages give an idea of how semantic coding is handled at dictionary level.

The first sample is taken from the one-word dictionary. It will be noticed that extensive use has been made of the semanto-syntactic codes CON, CT, MS and ABS on nouns and UTRAN on verbs. Also of interest is the fact that the two adjectives in this sample carry no semantic markers.

Among the semantic primitives used are CHCOM, FPROD, PRPFHY (a clear example of generalization when coded with "endurance") and SCINO.

******

The second (2-page) sample is taken from the expressions dictionary. The numerical information in the left hand column represents the analytical results which must be satisfied if the rule is to be applied.

The presence of semantic primitives here gives a fair idea of how semantic markers can be used at the transfer level to obtain special meanings in the target. IN, for example, when governing words coded with certain semantic primitives (MU, UNABR, SCINO) is rendered EN rather than DANS.

On the following page, there are examples of how noun meanings can be altered by using semantic primitives in connection with analytical information. CONTENT in relation to containers (CONTNR) becomes CONTENU; when governing IN and a food product (FPROD), the translation given is TENEUR.

OC	STEM/ID/EXPRESSION	POS BPQ		T G	M I O	MEANING 095 06-27-78 PAGE 617	D P Q	SYN C G	P B R	V E F	R E P	I N I	d D	DATE LAST UPDATED	C D R
0C	ENDOSULFAN	101C	GN=(N,S)	0	0	ENDOSULFANE	1	003A2	0	0	0	0	0	11-22-75	P
4D	ENDOSULFANS		SYN-CON,MS												
			SEM-CHCOM												
4C	ENDOSULFANS	101C	GN=(N,P)	0	0	ENDOSULFANE	1	003A2	0	0	0	0	0	11-22-75	X
0D	ENDOSULFAN		SYN-CON,MS												
			SEM-CHCOM												
0C	ENDOSULPHAN	101C	GN=(N,S)	0	0	ENDOSULFANE	1	003A2	0	0	0	0	0	07-15-77	L
			SYN-CON,CT												
			SEM-CHCOM												
0C	ENDOTHIA	101C	GN=(N,S)	0	0	ENDOTHIA	1	000A140	0	0	0	0	0	07-15-77	L
			SYN-CON,MS												
			SEM-FPROD												
0C	ENDOTOXIN	101C	GN=(N,S)	0	0	ENDOTOXINE	1	003A1	0	0	0	0	0	05-08-77	L
4D	ENDOTOXINS		SYN-CON,CT												
			SEM-CHCOM												
4C	ENDOTOXINS	101C	GN=(N,P)	0	0	ENDOTOXINE	1	003A1	0	0	0	0	0	05-08-77	X
0D	ENDOTOXIN		SYN-CON,CT												
			SEM-CHCOM												
0C	ENDPOINT	101C	GN=(N,S)	0	0	POINT	1	003 2	0	0	0	0	0	10-02-75	P
4D	ENDPOINTS		SYN-ABS,CT			CDS-WD2									
					1	FINAL	2	019	0	0	0	0	0		
4C	ENDPOINTS	101C	GN=(N,P)	0	0	POINT	1	003 2	0	0	0	0	0	10-02-75	X
0D	ENDPOINT		SYN-ABS,CT			CDS-WD2									
					1	FINAL	2	019	0	0	0	0	0		
0C	ENDRIN	101C	GN=(N,S)	0	0	ENDRINE	1	003A1	0	0	0	0	0	05-14-76	D
			SYN-CON,MS												
			SEM-CHCOM												
4I	ENDS	0404	HMRTN=25,GN=(S),3S,PR,PW,	0	0	TERMIN	3	001	0	0	0	0	0	07-27-76	X
0I	END		SU			CDS-NAREFL									
			SYN-UTRAN												
4C	ENDS	101C	HMRTN=25,GN=(N,P),PW,SW	0	0	EXTREMITE	1	003A1	0	0	0	0	0	07-27-76	X
0C	END		SYN-ABS,CON,CT												
0C	ENDURANCE	101C	GN=(N,S)	0	0	DURABILITE	1	003 110	0	0	0	0	0	05-17-77	P
			SYN-ABS,MS												
			SEM-PRPHY												
0C	ENDURE	0404	HMRTN=1,GN=(S,P),1S,2S,1P,	0	0	SUPPORT	3	001	0	0	0	0	0	10-03-77	L
4D	ENDURES		2P,3P,PR												
4C	ENDURING		SYN-GG,UTRAN												
4D	ENDURED														
4C	ENDURED	0404	HMRTN=20,GN=(S,P),1S,2S,3S,	0	0	SUPPORT	3	001	0	0	0	0	0	10-03-77	X
0D	ENDURE		1P,2P,3P,PA												
			SYN-GG,UTRAN												
4C	ENDURES	04C4	GN=(S),3S,PR	0	0	SUPPORT	3	001	0	0	0	0	0	10-03-77	X
0C	ENDURE		SYN-GG,UTRAN												
4C	ENDURING	4C4C		0	0	SUPPORT	3	001	0	0	0	0	0	10-03-77	X
0C	ENDURE		SYN-GG,UTRAN												
4C	ENEMIES	101C	GN=(N,P)	0	0	ENNEMI	1	003A2	0	0	0	0	0	06-09-75	X
0D	ENEMY		SYN-ABS,AMB,CON,CT,HU												
01	ENEMY	2035	HMRTN=38,GN=(S,P)	0	0	ENNEMI	2	00140	0	0	0	0	0	10-24-74	Y
0C	ENEMY	101C	HMRTN=38,GN=(N,S)	0	0	ENNEMI	1	003A2	0	0	0	0	0	06-09-75	C
4D	ENEMIES		SYN-ABS,AMB,CON,CT,HU												
0D	ENERGETIC	202C	GN=(S,P)	0	0	ENERGETIQUE	2	002A0	0	0	0	0	0	01-06-77	V
0D	ENERGETICS	101C	GN=(N,S)	0	0	ENERGETIQUE	1	00C 110	0	0	0	0	0	06-09-77	L
			SYN-ABS,MS			CDS-AS									
			SEM-SCIND												

DC	STEM/ID/EXPRESSION	POS UPQ	W N	T G	M I D	MEANING 080 01-15-79 PAGE 195	D P Q	SYN	P	V C G	R B R	E F	N P	I D I	DATE LAST UPDATED	C D R
91	IN $C-80 $C-88,E,20 $C-80 FORM $C-81,E,10 $C-84,0,01 $C-830,E,00		1	0	0	EN CDS-DELART	0	000A0	0	0	0	0	0	0	12-15-77	A
			5	0	0	FORME	1	000	1	0	0	0	0			
91	IN $C-80 $C-88,E,27 VIEW		1	0	0	2A	0	000A0	0	0	0	0	0	0	06-14-78	P
			4	0	0	AVIS	1	000A240		0	0	0				
91	IN $C-80 $C-MU		1	0	0	EN CDS-DELART	0	000A0	0	0	0	0	0	0	03-21-78	P
91	IN $C-80 $C-UNABR		1	0	0	EN CDS-DELART	0	000A0	0	0	0	0	0	0	03-21-78	P
91	IN $C-80 FRENCH $C-816,E,00		1	0	0	EN FRANCAIS	0	000A0	0	0	0	0	0	0	05-15-77	P
91	IN $C-80 PARTICULAR $C-816 $C-84,Z,02		1	0	0	EN PARTICULIER	0	0004Q	0	0	0	0	0	0	03-21-78	P
91	IN $C-80 THAT $C-80 $C-81,E,3C		1	0	0	PARCE	0	000	0	0	0	0	0	0	12-13-76	P
			3	0	0	QUE	0	000	0	0	0	1	0			
91	IN $C-80 THAT $C-80,M IT		1	0	0	PARCE	0	000	0	0	0	0	0	0	11-23-76	P
			3	0	0	QUE	0	000	0	0	0	1	0			
91	IN $C-80 TRACE $C-84,0,02		1	0	0	EN TRACES	0	0004Q	0	0	0	0	0	0	06-29-77	P
91	IN $C-80,M OVER		1	0	0	DANS	0	000	0	0	0	0	0	0	11-14-77	P
91	IN $C-80,M VARIOUS $C-816 $C-84,0,02		1	0	0	DANS CDS-DELART	0	000	0	0	0	0	0	0	03-21-78	P
91	IN $C-8105,0,02 THE CHAIR		1	0	0	CSPRESIDENCE	1	000	140	0	0	0	0	0	11-09-78	P
91	IN $C-8105,0,02 THE FIRST PLACE		1	0	0	CSEN PREMIER LIEU	0	000A0		0	0	0	0	0	12-12-76	P
91	IN $C-8105,0,02 THE MAIN $C-80 $C-81,E,60		1	0	0	ESSENTIELLEMENT	0	000A0		0	0	0	0	0	11-09-78	P
91	IN $C-8105,0,02 THE SECOND PLACE		1	0	0	DEUXIEMEMENT	0	000AC		0	0	0	0	0	11-09-78	P
91	IN $C-818 $C-80 EMERGENCY		1	J	0	EN CAS D'URGENCE	0	000A0		0	0	0	0	0	11-14-75	F
91	IN $C-818 $C-81,E,A0		1	0	0	EN	0	000	0	0	0	0	0	0	08-31-76	B
91	IN $C-818 $C-811,0,04 $C-80,-,1 $C-81,E,A0		1	0	0	DANS	0	000	0	0	0	0	0	0	08-31-76	B
91	IN $C-818 $C-8132,0,80		1	0	0	PENDANT	0	000	0	0	0	0	0	0	12-13-76	P
91	IN $C-818 $C-864,E,04 $C-81,E,A0		1	0	0	EN	0	000A0		0	0	0	0	0	10-02-76	B
91	IN $C-818 $C-83,E,1F		1	J	0	EN	0	000A0		0	0	0	0	0	12-03-75	L
91	IN $C-818 $C-INFO		1	0	0	$$$ CDS-END2A	0	000	0	0	0	0	0	0	08-31-76	B
91	IN $C-818 $C-MATER		1	0	0	EN CDS-DELART	0	000A0	0	0	0	0	0	0	03-21-78	P
91	IN $C-818 $C-MATER $C-826,E,00 $C-84,Z,02		1	0	0	EN CDS-DELART	0	000A0	0	0	0	0	0	0	08-11-78	P
91	IN $C-818 $C-MU		1	0	0	EN CDS-DELART	0	000A0	0	0	0	0	0	0	08-11-78	P
91	IN $C-818 $C-PRPCH		1	0	0	EN	0	000A0	0	0	0	0	0	0	08-26-76	B
91	IN $C-818 $C-SCIND		1	0	0	EN CDS-DELART	0	000A0	0	0	0	0	0	0	04-03-78	P
91	IN $C-819 CLOCKWISE DIRECTION		1	B	0	DANS LE SENS DES AIGUILLES D'UNE MONTRE	0	000	0	0	0	0	0	0	11-16-74	B
91	IN $C-818 FORCED-AIR		1	0	0	2A	0	000A0	0	0	0	0	0	0	03-21-78	P
			3	0	0	AIR FORCE	1	000A240		0	0	0				
91	IN $C-818 MEETING $C-81,E,1Q		1	0	0	$$$ CDS-END2A	0	000	0	0	0	0	0	0	11-17-77	P
			3	0	0	REUNION	1	003	1	0	0	0	0			
91	IN $C-818 ORDER $C-824 OF		1	K	0	PAR	0	000	0	0	0	0	0	0	06-14-76	B
			3	K	0	ORDRE	1	003	240	0	0	0				
			5	K	0	$$$ CDS-ENDDE	0	000	0	0	0	0	0			
91	IN $C-816 POSITION $C-80,-,1 $C-19		1	J	0	EN	0	000A0	0	0	0	0	0	0	11-22-75	L
			3	J	0	POSITION	1	003	140	0	0	0				
91	IN $C-818 RELATION $C-84,0,01 $C-824,M WITH		1	0	0	EN	0	000A0	0	0	0	0	0	0	10-03-77	P
			3	0	0	RELATION	1	003	140	0	0	0				
91	IN $C-818 RESPECT		1	0	0	2A	0	000	0	0	0	0	0	0	04-01-77	P
			3	0	0	REGARD	1	003A2		0	0	0				

DC	STEM/ID/EXPRESSION	POS BPQ	N	G	T I D	MEANING 080 01-15-79 PAGE 84	D P Q	SYN	P	G	V B R	R E F	I M P	W D I	DATE LAST UPDATED	C D R
92	CONSUMER TAX		1	0	0	CONSOMMATION	1	003	1	0	0	0	0	0	11-05-78	P
			2	0	0	TAXE	1	003	1	0	0	0	0			
92	CONSUMING SECTOR		1	0	0	CONSOMMATION	1	003	1	0	0	0	0	0	11-14-77	P
			2	0	0	SECTEUR	1	003	2	0	0	0	0			
92	CONSUMPTION CREDIT		2	0	0	CR3EDIT 2A LA CONSOMMATION	1	003	2	0	0	0	0	0	01-04-79	P
92	CONSUMPTION LOAN		2	0	0	PR4ET 2A LA CONSOMMATION	1	003	2	0	0	0	0	0	01-04-79	P
91	CONSUMPTION OF HIGH-PROTEIN PRODUCT		1	K	0	CONSOMMATION	1	003	1	0	0	0	0	0	04-30-76	S
			2	K	0	DE	0	000	0	0	0	0	0			
			4	K	0	PROT3EAGINEUX	1	000	2	0	0	0	0			
92	CONTACT PERIOD		1	0	0	EX3ECUTION	1	003A1	0	0	0	0	0	0	02-27-78	P
			2	0	0	D3ELAI	1	003	2	0	0	0	0			
91	CONTAIN $C-B1,E,38 $C-B16 $C-B11,0,02 $C-824,M IN		1	J	0	PAPAISSANT	3	000	0	0	0	0	0	0	11-14-75	L
91	CONTAIN $C-B1,E,4C $C-B0,-,1 $C-FPROD		1	0	0	CONTENANT	0	000	0	0	0	0	0	0	11-23-76	P
93	CONTAIN $C-823,M MORE		3	0	0	PLUS	0	000	0	0	0	0	0	0	01-07-77	P
						CDS-DEWORD,ENODE										
92	CONTAINER BARGE		2	0	0	BARGE-	1	003	1	0	0	0	0	0	07-10-78	D
						CDS-WD2										
					1	CONTENEUR	2	002	0	C	0	0	0			
			2	SEM-TRANSP												
92	CONTAINER FLOW		1	K	0	DES R3CIPIENTS	1	000	0	0	0	0	0	0	06-14-76	B
						CDS-NLPREP										
			2	K	0	ACHEMINEMENT	1	00342	0	0	0	0				
92	CONTAINER JOINT BOLT		1	B	0	R3CIPIENT	0	000	2	0	0	0	0	0	11-13-74	A
			3	B	0	BOULON DE JOINT	1	003	2	0	0	0	0			
92	CONTAINER SEAL		1	B	0	R3CIPIENT	1	003	2	0	0	0	0	0	11-13-74	A
			2	B	0	JOINT D'3TANCH3EIT3	1	003	2	0	0	0	0			
92	CONTAMINATING PARTICLE		2	J	0	PARTICULE	1	003	1	0	0	0	0	0	11-20-75	B
						CDS-WD2										
					1	3ETRANGER	2	014AC	0	0	0	0	0			
91	CONTENT $C-B0,M OF $C-B18,M TEXT		1	0	0	CONTENU	1	000	210	0	0	0	0	0	12-12-78	P
						CDS-AS										
91	CONTENT $C-820,NE,00 $C-824 OF		1	K	0	TENEUR	1	003	1	0	0	0	0	0	12-13-76	P
						CDS-GOVEN										
			4	K	0	$$$	0	000	0	0	0	0	0	0		
						CDS-ENODE										
91	CONTENT $C-824 BETWEEN $C-B18 $C-B11,0,04		1	0	0	TENEUR	1	003	1	0	0	0	0	0	09-03-76	B
						CDS-WD2										
					1	COMPRIS	2	001	0	0	0	0	0	0		
			3	0	0	ENTRE	0	000A0	0	0	0	0	0			
91	CONTENT $C-824 IN $C-B0 $C-FPROD		1	0	0	TENEUR	1	003	1	0	0	0	0	0	08-11-78	P
			3	0	0	EN	0	000	0	0	0	0	0	0		
						CDS-DELART										
91	CONTENT $C-824,M OF $C-B18 $C-CONTNR		1	0	0	CONTENU	1	003	230	0	0	0	0	0	03-16-77	P
						CDS-AS										
91	CONTENT $C-826,NE,00 $C-824 OF		1	K	0	TENEUR	1	003	1	0	0	0	0	0	04-01-77	P
			4	K	0	$$$	0	000	0	0	0	0	0	0		
						CDS-ENODE										
91	CONTENT $C-84,0,02 $C-8105,0,02 $C-B0 $C-B1,E,CC		1	0	0	TABLE DES MATI2ERES	1	000	1	0	0	0	0	0	11-09-78	P
						CDS-AS										
91	CONTENT OF $C-B18 $C-CONTNR		1	0	0	CONTENU	1	003	2	0	0	0	0	0	09-03-76	B
						CDS-AS										
			2	0	0	$$$	0	000	0	0	0	0	0	0		
						CDS-ENODE										

Annex IV - Sample translations

Three typical samples of Systran English-French output are appended to illustrate the quality of raw output which can now be obtained without any pre- or post-editing.

The two sentence-by-sentence samples are from a working document and the Commission's annual report respectively, while the third upper/lower case printout is a translation of a journal article.

All three do of course contain a number of errors, some of which could be eliminated by improvements to the system at either the dictionary or program level while others would normally be left to the post-editor.

The fact that the same dictionaries were used for three different subject fields is evidence of the system's ability to cope with different types of text.

COMMUNICATION MULTILINGUE .

637 IF THIS TOOL IS TO REMAIN USABLE , THE FORMAT OF THE DATA BANK MUST BE CONSTANTLY ADAPTED TO USER REQUIREMENTS .
SI CET OUTIL DOIT RESTER UTILISABLE , ON DOIT ADAPTER CONSTAMMENT LE FORMAT DE LA BANQUE DE DONNEES AUX BESOINS D' UTILISATEUR .

638 CAREFUL MAINTENANCE AND CONSTANT UPDATING OF THE DATA BASE ARE ABSOLUTELY INDISPENSABLE .
L' ENTRETIEN SOIGNEUX ET LA MISE ZA JOUR CONSTANTE DE LA BASE DE DONNEES SONT ABSOLUMENT INDISPENSABLES .

639 3 .
3 .

640 EURODICAUTOM IS A MULTILINGUAL AND THEREFORE AN INTERNATIONAL TERMINOLOGY BANK .
EURODICAUTOM EST MULTILINGUE ET DONC UNE BANQUE INTERNATIONALE DE TERMINOLOGIE .

641 NATIONAL ORGANS AND INSTITUTIONS CONCERNED WITH THE PRESERVATION OF CORRECT USAGE , TERMINOLOGY WORK AND STANDARDIZATION MUST THEREFORE BE CONSULTED WHEN THE DATA BASES ARE COMPILED .
DES ORGANES ET DES INSTITUTIONS NATIONAUX CONCERNES PAR LA CONSERVATION DE L' UTILISATION , DU TRAVAIL DE TERMINOLOGIE ET DE LA NORMALISATION CORRECTS DOIVENT DONC ÊTRE CONSULTES QUAND LES BASES DE DONNEES SONT COMPILEES .

642 4 .
4 .

643 EURODICAUTOM IS ONLY ONE OF A NUMBER OF INSTITUTIONS CONCERNED WITH THE COLLECTION AND STORAGE OF LINGUISTIC INFORMATION .
EURODICAUTOM EST SEULEMENT L' UNE D' UN CERTAIN NOMBRE D' INSTITUTIONS CONCERNEES PAR LE RAMASSAGE ET L' ENTREPOSAGE DES INFORMATIONS LINGUISTIQUES .

644 CLOSE COLLABORATION WITH SIMILAR INSTITUTIONS IS DESIRABLE FOR OTHER REASONS THAN THE SIMPLE COST FACTOR .
LA COLLABORATION ETROITE AVEC DES INSTITUTIONS SEMBLABLES EST SOUHAITABLE POUR D' AUTRES RAISONS QUE LE FACTEUR DE COUT SIMPLE .

645 IT IS A CRUCIAL PRECONDITION FOR ATTAINING THE COMMON GOAL , WHICH IS TO FACILITATE AND IMPROVE MULTILINGUAL COMMUNICATION .
C' EST UNE PRECONDITION CRUCIALE POUR ATTEINDRE LE BUT COMMUN , QUI DOIT FACILITER ET AMELIORER LA COMMUNICATION MULTILINGUE .

646 5 .
5 .

647 AS A FIRMLY ESTABLISHED SERVICE , EURODICAUTOM REQUIRES A CONSIDERABLE INITIAL INVESTMENT IN MANPOWER .
COMME UN SERVICE FERMEMENT ETABLI , EURODICAUTOM EXIGE UN INVESTISSEMENT INITIAL CONSIDERABLE DANS LA MAIN-D'OEUVRE .

648 A CONTINUING AND CONSTANT EFFORT WILL SUBSEQUENTLY BE NEEDED IF THE INITIAL INVESTMENT IS NOT TO HAVE BEEN SPENT IN VAIN .
UN EFFORT CONTINUANT ET CONSTANT SERA EXIGE PLUS TARD SI L' INVESTISSEMENT INITIAL DOIT NE PAS ÊTRE DEPENSE DANS VAIN .

ARTICLE 92 (3) , EN PRENANT L' 3ETUDE D4UE DES BESCINS DE LOI ET CE POLITIQUES CCMMUNAUTAIRES , NOTAMMENT LA POLITIQUE AGRICOLE COMMUNE .

359 FARM ACCOUNTANCY DATA NETWORK .

R3ESSAU L' INFORMATIONS CE COMPTABILIT3E DE FERME .

360 348 .

348 .

361 THE RESULTS OF THE FARM ACCOUNTANCY DATA NETWORK (FADN) , ON WHICH THE COMMISSION PRESENTS AN ANNUAL REPORT TO THE COUNCIL AND PARLIAMENT , ARE BASED ON DATA COLLECTED IN THE COMMUNITY BY SOME 250 ACCOUNTANCY OFFICES FROM SOME 15,000 SAMPLE HOLDINGS SELECTED BY ABOUT FIFTY REGIONAL COMMITTEES .

LES R3ESULTATS DU R3ESEAU D' INFORMATIONS DE COMPTABILIT3E DE FERME (CC$FADN) , SUR LEQUEL LA C$OMMISSION PR3ESENTE UN RAPPORT ANNUEL AU C$CONSEIL ET AU C$PARLEMENT , SONT FONC3ES SUR DES INFORMATIONS RASSEMBL3EES DANS LA C$OMMUNAUT3E PAR QUELQUES 250 BUREAUX DE COMPTABILIT3E DE QUELQUES 15,000 EXPLOITATIONS D' 3ECHANTILLON CHOISIES PAR CINQUANTE COMIT3ES ENVIRON R3EGIONAUX .

362 THIS REPORT PROVIDES THE COMMUNITY AUTHORITIES RESPONSIBLE FOR THE COMMON AGRICULTURAL POLICY WITH A SET OF UNIFORM AND DETAILED DATA , BY TYPE OF FARMING AND BY REGION , ON THE SITUATION AND INCOMES OF AGRICULTURAL HOLDINGS ;

CE RAPPORT FOURNIT AUX AUTORIT3ES COMMUNAUTAIRES RESPONSABLES DE LA POLITIQUE AGRICOLE COMMUNE UN ENSEMBLE D' INFORMATIONS UNIFORMES ET D3ETAILL3EES , PAR LE TYPE D' AGRICULTURE ET PAR R3EGION , SUR LA SITUATION ET LES REVENUS DES EXPLOITATIONS AGRICOLES ;

363 THESE DATA ARE OF PARTICULAR VALUE IN THE FORMULATION OF PROPOSALS ON PRICES (OBJECTIVE METHOD) AND ON AGRICULTURAL STRUCTURES .

CES INFORMATIONS SONT DE VALEUR PARTICULI2ERE DANS LA FORMULATION DES PROPOSITIONS SUR DES PRIX (M3ETHODE OBJECTIVE) ET SUR DES STRUCTURES AGRICOLES .

364 THE NUMBER OF HOLDINGS IN THE FADN SAMPLE INCREASED FROM 19.000 IN ' 1976 ' TO 24.500 IN ' 1977 ' .

LE NOMBRE D' EXPLOITATIONS DANS L' 3ECHANTILLON CC$FADN A AUGMENT3E DE 19,000 EN ' 1976 ' JUSQU' 2A 24,500 EN ' 1977 ' .

365 THIS INCREASE OF ALMOST 30% IN THE ACCOUNTING SAMPLE IN ONE YEAR SUGGEST THAT THE FINAL GOAL OF SOME 28.000 RETURNING HOLDINGS SHOULD BE ACHIEVED IN ' 1978 ' .

CETTE AUGMENTATION DE PRESQUE 30% DE L' 3ECHANTILLON DE COMPTABILIT3E DANS UN AN SUGG2ERE QUE LE BUT FINAL DE QUELQUES 28,000 EXPLOITATIONS COMPTABLES DEVRAIT 4ETRE R3EALIS3E EN ' 1978 ' .

366 THERE ARE , HOWEVER , CERTAIN DIFFICULTIES IN TWO MEMBER STATES WHERE A CONSIDERABLE EFFORT REMAINS TO BE MADE - THE TWO WHERE THERE HAS BEEN DELAY IN IMPLEMENTING THE ACCOUNTANCY AID PROVIDED FOR IN ARTICLE 11 OF DIRECTIVE 72/159 ON THE MODERNIZATION OF FARMS .

IL Y A , CEPENDANT , CERTAINES DIFFICULT3ES DANS DEUX C$ETATS MEMBRES O2U UN EFFORT CONSID3ERABLE RESTE 2A 4ETRE FAIT - LES DEUX O2U IL A Y EU RETARD EN METTANT EN APPLICATION L' AIDE DE COMPTABILIT3E PR3EVUE 2A L' ARTICLE 11 DE LA DIRECTIVE 72/159 SUR LA MODERNISATION DES EXPLOITATIONS .

367 TO OBTAIN A BETTER UNDERSTANDING OF THE IMPACT ON THE ECONOMY AND INCOMES OF AGRICULTURAL HOLDINGS OF CERTAIN INCREASINGLY IMPORTANT FACTORS , SUCH AS INVESTMENT , INFLATION AND DIRECT AID , THE COMMISSION ADOPTED ON 23 SEPTEMBER A NEW FARM RETURN WHICH TAKES ACCOUNT OF EXPERIENCE GAINED IN THE LAST TEN YEARS AND RECENT PROGRESS IN DATA PROCESSING .

POUR OBTENIR UNE COMPR3EHENSION MEILLEURE DE L' IMPACT SUR L' 3ECONOMIE ET DES REVENUS DES EXPLOITATIONS AGRICOLES DE

les Français parle anglais. Adroitement, mais intrinsèquement peu probable. Champion Schmidt. Si les allemands semblent généralement maladroits sur la question de langue, ce n'est pas le défaut de monsieur Helmut Schmidt, leur chancelier. Il a soutenu la cause de langue anglaise et a fait son meilleur pour réduire le nombre de langues employées aux sommets de la CEE. Bien que les entretiens formels de sommet encore comportent la panoplie entière de langues communautaires, il est parvenu à introduire anglais pour les réunions officieuses. Au sommet Copenhague en avril les chefs de gouvernement pour la première fois se sont seulement réunis après dîner pour examiner monsieur le plan monétaire de Schmidt's (puis ultra-secret) - entièrement en anglais. Seulement le premier ministre italien, monsieur Giulio Andreotti, a exigé un interprète. Le même expédient a été adopté au sommet Brême en juillet.

Monsieur Schmidt a également a essayé de couper sur employer un excès de langues pour d'autres buts de la CEE. Par exemple, il a suggéré dans le coffret allemand que le passeport de la CEE propose devrait être en français et anglais seulement, pour la garder plus.

Toma, P. - Systran - A machine translation system of the 3rd generation. Luxembourg, CEC, ref : CETIL/94/78, 12 p.

Van Slype, G. and Pigott, I. - Description du système de traduction automatique Systran de la Commission des Communautés Européennes. Luxembourg, CEC, 1979, 28 p.

Chaumier, J., Mallen, M.C. and Van Slype, G. - Evaluation du système de traduction automatique Systran. Luxembourg, CEC, 1977, 17 p.

Van Slype, G. - Deuxième évaluation du système de traduction automatique Systran anglais-français de la Commission des Communautés Européennes. Luxembourg, CEC, 1978, 179 p.

Masterman, M. - Interim Report of Study Contract TH-17 on the development potential of the Systran system. CEC internal document, 1979, 15 p.

Systran dictionary coding manual and systems documentation. WTC, La Jolla, California. 1978, approx. 400 p.