# V. TEXT TYPE, TERMINOLOGY AND TRANSLATION QUALITY.

After the initial period of MT (1950's and early 1960's), in which it were mainly the "computer people" who attempted to put together a translation machine, a gradually increasing use was made of new developments and insights in linguistics, a field which made rapid progress during the last decades [see fig. 1-1 for a very cursory chronology of developments in linguistics, parallel to those in the computer industry].

MT today is certainly connected with a lot of linguistics, but it remains something different! Though some universities or institutions may experiment with MT for the more academic purpose of model testing or theory verification, MT as highlighted in this report is the applied use of linguistic and other know-how to perform a concrete task in the most economic way.

Present-day linguistics is by no means devoted to the particular requirements of the automation era. In fact, linguistics seems to be predominantly concerned with: spoken language, language differences (sociolinguistics), dialects, language acquisition and instruction. One could perhaps characterize this roughly as 'anti-normative'. If one comes across formalisms in modern linguistics, it is often their cognitive or psychological basis that matters and from which they derive their importance.

A practical MT system as DLT, on the other hand, appears to be more compatible with structured, 'normative' language [the use of a normative nineteenth-century element (Esperanto) internally within DLT, is only a coincidence].
In this chapter of the report, we will look at the text type, the processing of specialistic terminology and the translation quality as regards DLT.

## 1. Text type to be handled by DLT.

As a first characterization, the following principal limitations must be emphasized:

- written language
  i.e. keyboard-entered text only; speech input is outside the scope of this report and the DLT project in general;

- informative language
  the primary purpose of the text must be to inform, not to incite or invoke, nor to convey subjective feelings, personal emotions, esthetic appeal etc.;

- normative language
  the text must adhere to certain norms: it must be correct, conventional language; often, it will be directed at an audience (Information Distribution), which may range from small and narrowly defined to wide and heterogeneous; but even if the text is directed at a single receiver (Electronic Mail), a certain formality is required: slang, familiarities, spoken-language expressions etc. must be banned.

LSP (Language for Special Purposes, Fachsprache, Langue de Spécialité) certainly overlaps with the DLT text type. We quote two fragments found among the many attempts to define LSP:

> "the supranational activities of industrial and commercial interests carried out with an eye to the language and the reasoning capacity of the computer" [Opitz, 1980];
>
> "to communicate information .... in the most economic, precise, unambiguous terms available" [Draskau, 1983].

Not only scientific journals (including the more popular ones), also high-quality daily papers as Le Monde and The Times can be reckoned to the informative text type, which covers quite a range of texts of different degrees of syntactic complexity and readability.
In making texts for international journals as well as for manuals, instructions for export products etc., simplicity seems to become more and more a demand:

> "when writing for an international audience, language must be simple, with no ambiguities, Americanisms or slang"
> [Tiefenthal, 1980].

An important characteristic of the informative text type as a whole is that it exists of denotative language. There is no intentional use of connotations (subtle second meanings or associations of a word) or non-literal interpretations related with mental states and attitudes [cf. Kay, 1982; Hendrix, 1981].

## 2. Linguistic properties of this text type.

Due to the wide range of different materials within the informative text type, the following treatment can only be illustrative. The most striking features of this text type, as known from general linguistic and LSP literature, have played a role in the IL design described in Chapter IV.

## 2.1. Dominant sentence patterns.

For French, Bourquin [1983] investigated sentence patterns in Le Monde, a study out of which we can only mention a few items. The dominant clause pattern in Le Monde is:

x S x V x O x
1 2 3 4

i.e. primarily an SVO pattern with four 'insertion spots' for various secondary constituents. It appears that free adjuncts (PPs, NPs denoting time, etc.) most frequently occupy position 1 or 4, which has reinforced the IL design decision to place FADJs either at the clause beginning (in GENeral clauses) or at the clause end (for INFinitival clauses).
Another interesting issue found by Bourquin for Le Monde, and also mentioned by Ulijn [1978] in his thesis about technical French is NP expansion due to verb nominalization. This has stimulated extra attention to complex NP structures in the IL, and has lead to a solution in which the underlying clausal structure around the nominalized verb is preserved.
In her analysis of 'English for Economists', Gallais-Hamonno [1982] gives interesting statistics of NP composition (including separate counts for different numbers of premodifiers etc.). Studies as this one, with many systematically ordered fragments of relevant text material, will be of great help in further DLT development.

## 2.2. Absent grammatical elements.

"Many of the question forms, stylistic inversions, and exclamatives of conversational English are totally absent from technical literature" [Kittredge, 1982].
The same applies to the use of certain past tenses. In an analysis of 25 articles on chemistry in a number of prominent French periodicals, Wojnicki [1981] found an extremely low frequency of any past tense other than the 'prèsent indicatif' and the 'passè composè'.
Pigott [1982a] reported, that the '1st person' had long been neglected in SYSTRAN, because of its rareness in the MT text type. However, if one thinks of information distribution in offices, documents with the (plural) 1st person will be no exception.
Also the '2nd person' forms an element that will be absent in many informative texts, apart from a certain subset of the more user-friendly written manuals, operational instructions, man-computer dialogues etc.

2.3. Specific translation requirements.

If the primary purpose of informative language is to convey information, then the first requirement for this type of text is reliability. As a professional translator once stated at an MT conference, it makes quite a difference whether the TL output reads 'ten meter square' or 'ten square meter'.
Good style and smooth readability of the target text is the next important requirement. Instructional texts should not loose their clearity and conciseness during translation.
In general, the translation quality required must fit in fast electronic information-distribution situations where the audience is large and where its 'distance' to the information provider prevents clarification of any difficulties.

A requirement which is of crucial importance in technical texts is the ability to handle specialistic terminology [see below]. Another requirement, of a less linguistic nature but equally important, is the correct handling of all kinds of literal strings, numbers, acronyms, quotations etc. that appear in the SL text but are NOT to be translated.

## 3. DLT's aptitude for the translation of terminology.

3.1. Problems and requirements.

Esperanto, the basis of DLT's IL, has been the focal point of attention of a small number of linguists and terminologists, the most prominent of them being Wüster [1975, 1978, 1979]. A recent overview of the status and problems of terminology in Esperanto is [Fritsch, 1979]. In 1981, a conference on the application of Esperanto in science and technology [Žilina, 1981] was partially devoted to the same subject.
As with the provision of Esperanto dictionaries [see IV.4.5], the development and standardization of Esperanto terminology generally suffers from insufficient backing by commercial or governmental interests.

As to the linguistic aspects of Esperanto terminology, there is one issue which is worth mentioning here: the naturalists vs. purists controversy [Wells, 1978].
The naturalists hold the view that in Esperanto, technical terms should as much as possible resemble the international forms, often derived from Latin or Greek, and being currently used in English as well as French and other languages, e.g.:

imperialismo ekspresionismo komputero

[notice the international '-ism' ending].
On the other side, the purists adopt international forms only as

Esperanto roots, and apply endings or typical Esperanto affixes directly to these:

imperiismo esprimismo komputilo

In DLT, a pragmatic approach will be taken when a new term has to be added to the IL lexicon. If at least two of the languages English, French and German use the same international term, then the naturalistic method will be followed. Otherwise the puristic creation will be applied.

### 3.2. Candidate fields of terminology for DLT.

For a DLT pilot system, the terminology of the discipline or application area aimed at, should already exist for a large part in Esperanto. This is the case with: mathematics, zoology, international business and - to some extent - international law.
The first two of these are rejected because of the relatively small market and therefore low return of investment they would eventually mean for DLT. The attractively limited size of mathematics terminology (approx. 5000 terms) cannot compensate this.
From a market point of view, international business and economics is attractive. Besides, it is excellently covered by the standard work of [Munniksma, 1975].

An interesting and almost unavoidable extension of international business terminology is public international law. Here, we rely on the advice of Lapenna [1983], who asserts that completion and effective use of the already existing Esperanto terminology for this field is quite feasible [see also section VII-6]. Lapenna is a well-known expert on international law, and has written numerous articles on language problems, translation and Esperanto [Lapenna 1972, 1982].

## 4. Translation quality attainable with DLT.

### 4.1. Expected stylistic faults.

The quality level at which DLT aims lies in the range of categories 7 and 8 of the scale of Carroll [Van Slype, 1982], which are defined as follows:

> "Category 8: Perfectly or almost clear and intelligible, but contains minor grammatical or stylistic infelicities, and/or mildly unusual word usage that could, nevertheless, be easily 'corrected'. "

"Category 7: Generally clear and intelligible, but style and word choice and/or syntactical arrangement are somewhat poorer than in category 8. "

The exact level that can be attained will much depend on the comprehensiveness of the DLT lexicons, in particular with regard to the quantity of phraseological expressions, idioms and collocations covered. Of the stylistic faults that can be expected, we mention:

- monotony in word order (loss of theme/rheme);
- monotony in word choice;
- word repetition instead of anaphoric references;
- literal translation of idioms.

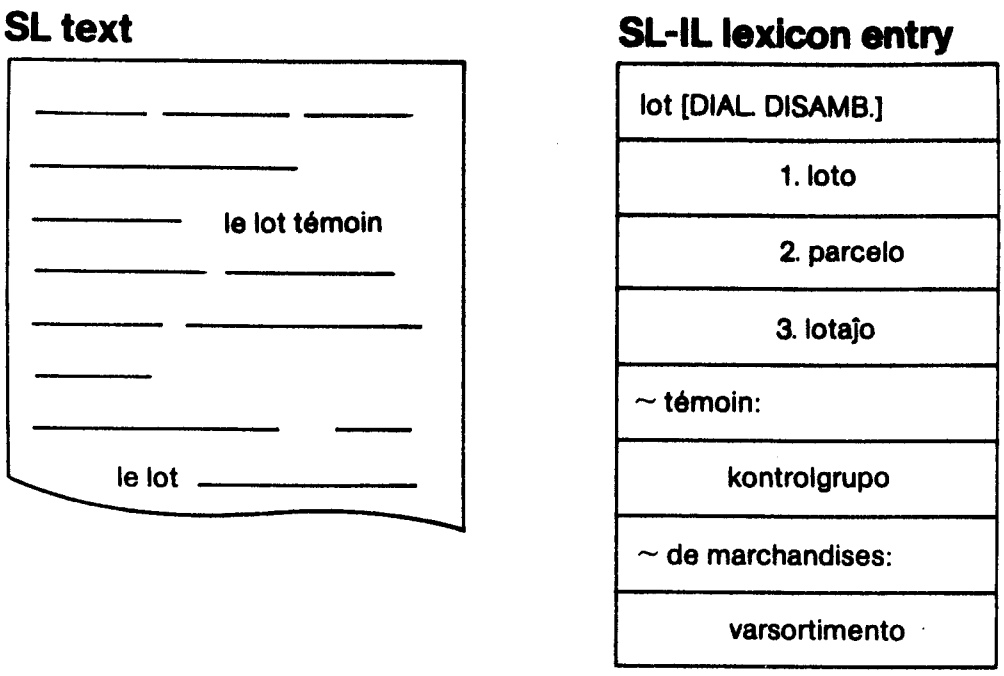
In addition, the first versions of DLT will show a number of deficiencies of a more grammatical nature: incorrect use of articles, wrong past tenses, false verbal aspects etc.

4.2. Translation precision with DLT.

An opinion often heard in discussions about MT or translation in general, is that the use of a pivot language such as the Esperanto-based IL (which, as we admit in section III-1.4, implies a double translation) would cause an increased information loss.
We will briefly explain here, that this is not the case and that, on the contrary, the DLT translation process even can produce an enrichment of the text. This is easy to understand if one thinks of the addition of human intelligence (via the interactive disambiguation dialogue) to the translation process.

The IL version of the text must always be unambiguous, also with respect to anaphoric references. DLT will therefore force the human attendant to explicitize any ambiguous reference, and the human response to such a question will be documented in the IL by monotonous word repetition (if enough sophisticated, the TL-module can of course replace full forms by anaphorics again). Such a case is illustrated in fig. V-1.

Fig. V-2 gives an impression of the precision that can be attained with DLT in the near future. Note the constant appearance of the 'passè simple' and the not quite correct '60 kg lourds' [line 24] in the output text.




## Disambiguation dialogue

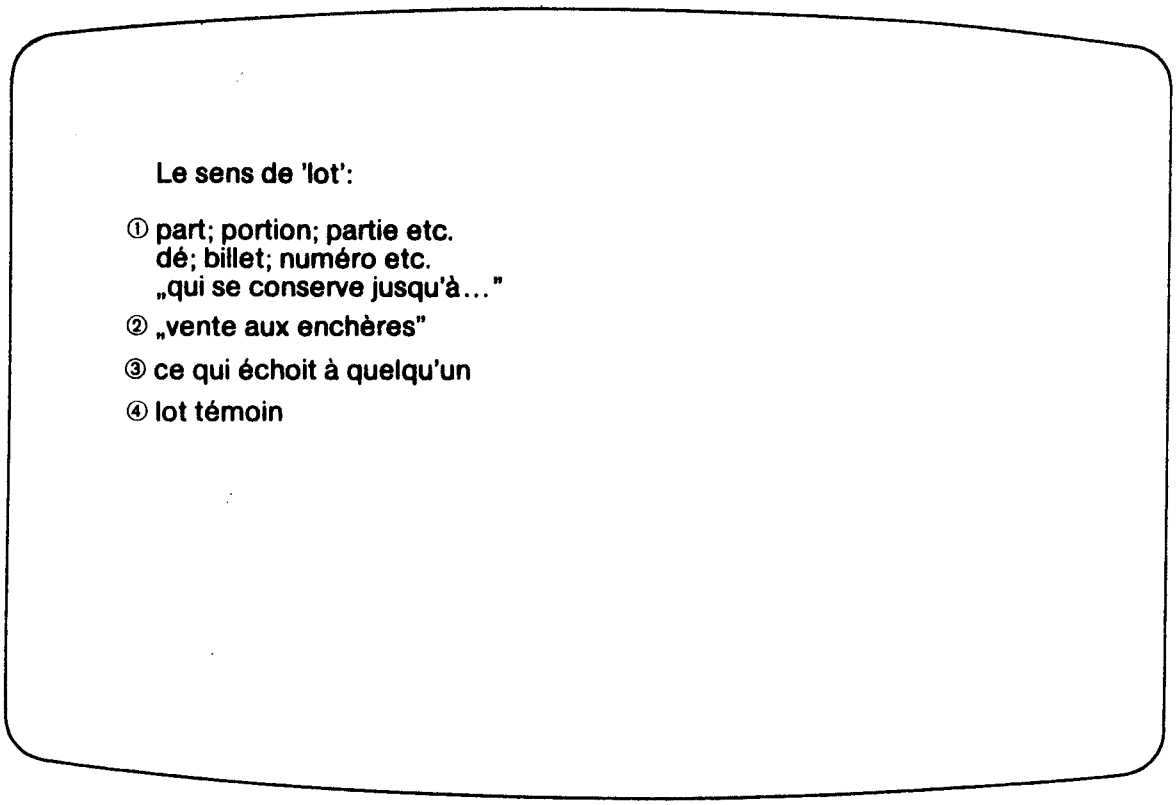



Fig. V-1. Example of 'referential enrichment' in DLT. At the first occurrence of 'le lot' (with 'témoin') in the SL text, it matches one of the collocations in the lexicon entry ('kontrolgrupo'). At the next occurrence of 'le lot' (without 'témoin'), the computer-generated menu therefore presents the same collocation as a possible interpretation. If the human operator agrees with this interpretation, the IL version will contain 'kontrolgrupo' (and a translation back to French would read 'le lot témoin') at both occurrences.

**SL**

Pour déterminer ce facteur, le Pr. Paquay a pris douze brebis qu'il a nourries pendant aquelques dizaines de jours avec une alimentation d'entretien juste suffisante pour assurer la survie des animaux.
Puis, il a divisé les animaux en deux groupes: l'un, le lot témoin, étant constitué de trois brebis, et l'autre de neuf brebis.
Le lot témoin reçut pendant toute la durée de l'expérience la même ration d'entretien, tandis que les neuf brebis de l'autre lot furent nourries à volonté.
Puis, régulièrement, les douze brebis furent pesées et soumises à des examens sanguins.
Au début de l'expérience, tous les animaux pesaient 60 kg.

Or, dès qu'elles furent nourries à volonté, les neuf brebis du second lot augmentèrent leurs ingestions et prirent du poids.
Puis, lorsque celui-ci se stabilisa à 90 kg (leur poids normal) les ingestions diminuèrent.
A ce moment-là, on observa dans le sang des animaux une augmentation du taux d'acide linoléique et inversement une diminution du taux d'acide oléique.

[Andeol, 1979]

**IL**

Por determini tiun faktoron la profesoro 'Paquay' prenis dek du ŝafinojn kiujn dum kelkaj dekoj da tagoj li nutris per vivtena nutrado, ĝuste sufiĉa por certigi la postvivon de la bestoj.
Poste, li dividis en du grupojn la bestojn: la unun, la kontrolgrupon, konsistantan el tri ŝafinoj, kaj la aliun el nau ŝafinoj.
Dum la tuta dauro de la eksperimento la kontrolgrupo ricevis la saman vivtenan porcion, dum-kiel la nau ŝafinoj de la aliu loto nutrajtis lauvole.
Poste, je regulaj tempoj, la dek du ŝafinoj pesajtis kaj submetajtis al sanganalizoj.
Komence de la eksperimento, ĉiuj bestoj estis je '60'-b kilogramojn pezaj.
Nu, tuj kiam iŝi nutrajtis lauvole, la nau ŝafinoj de la dua loto pliigis siajn enstomakigojn kaj grasiĝis.
Poste, kiam la pezo stabiliĝis en '90'-b kilogramoj (ilia normala pezo) la enstomakigoj malpliiĝis.
En tiu momento, ie-en la sango de la bestoj oni observis pliiĝon de la titro de linolata acido kaj inverse malpliiĝon de la titro de oleata acido.

**TL**

Pour déterminer ce facteur, le Pr. Paquay prit douze brebis qu'il nourrit pendant quelques dizaines de jours avec une alimentation d'entretien juste suffisante pour assurer la survie des animaux.
Puis, il divisa en deux groupes les animaux: l'un, le lot témoin, se composant de trois brebis, et l'autre de neuf brebis.
Le lot témoin reçut pendant toute la durée de l'expérience la même ration d'entretien, tandis que les neuf brebis de l'autre lot furent nourries à volonté.
Puis, régulièrement, les douze brebis furent pesées et furent soumises à des examens sanguins.
Au début de l'expérience, tous les animaux furent 60 kg lourds.

Or, dès qu'elles furent nourries à volonté, les neuf brebis du second lot augmentèrent leurs ingestions et prirent du poids.
Puis, lorsque le poids se stabilisa à 90 kg (leur poids normal) les ingestions diminuèrent.
A ce moment-là on observa dans le sang des animaux une augmentation du taux d'acide linoléique et inversement une diminution du taux d'acide oléique.

Fig. V-2. DLT dry-run from French to IL and again back to French.