TYPOLOGY STUDY OF FRENCH TECHNICAL TEXTS, WITH A VIEW TO DEVELOPING A MACHINE TRANSLATION SYSTEM

B.ROUDAUD
B'VITAL (SITE group)
35 rue J. Chanrion - 38000 Grenoble - FRANCE
tel : +33 76 51 04 96
e-mail : B_Roudaud@site-maisons-alfort.fr

ABSTRACT

Within the industrial context of the information society, technical translation represents a considerable commercial stake. In the light of this, machine translation is considered as being an application of paramount importance. It is for this reason that the activities of B'VITAL have always centered around the processing of technical texts.

The following article gives an account of the various tasks carried out over the last few years on corpus analysis. We have drawn conclusions as to the validity of the notion of *text typologies*, applied in particular to technical matter, with a view to developing a machine translation system. The study was conducted using a fair amount of French documents and has led us to observe in particular, that a same typology may be identified in texts originating from varying fields.

INTRODUCTION

The technical literature of a nuclear plant represents about 150 000 pages (figure quoted by EDF, France, in 1990). About 20 000 pages make up the maintenance literature of a an aircraft, of which 5 000 are subject to revision every three months on average.

In France, the cost per page for a translation varies from 250 to 400 francs for a translation from French to one of the other European languages (according to Bossard Consultants), which amounts to 6 million francs for the maintenance literature of an aircraft.

If a machine translation system is able to reduce considerably the time spent per page by translators, even if the result is not perfect, it will signify a gain with respect to delivery times and costs.

To avoid problems concerning non standardised terminology, we have endeavoured to stick to the terms familiar to traditional grammar, even though they may not always be appropriate.

I wish to thank my colleagues, D. Bachut, O. Gamrat and M.C. Puerta, for their help.

1 - DEFINITION OF THE CORPUS STUDY

According to our method, based on GETA's works (Grenoble), if the MT system comes across a non handled phenomenon, the sentence is not rejected and translation is carried through using those parts of the sentence which have been successfully analysed (Cf. [1] & [2]). Within the framework of an industrial application in which the ratio of development costs to gain in quality is preponderant, rare phenomena may thus be handled in a very simplified way, or even not handled at all.

Midway between controlled syntax systems and systems 'which translate everything', our approach to MT favours the development of systems adapted, a posteriori, to the texts to be translated. This theory goes by the assumption that it is possible to define and then recognise the typology of texts, specifying in particular their form, the linguistic phenomena present or absent and the general vocabulary used (excluding terminology).

The study, which was strongly influenced by the MT application, began in 1984, during the French *Projet National de Traduction Automatique* (PNTAO), with an intensive research phase. It is still currently in progress although only sporadically, with the study of new texts. The first study aimed at proving that it was possible to define a typology for the given corpus. Its result was the definition of an initial typology (less restrictive than the METEO typology for instance, see [6]), which was further refined (though not radically changed) during the years which followed. The definition of the typology retained consisted of a list of the phenomena handled, in other words, a list of the *static grammar charts* (Cf. [2] & [3]) which are part of the linguistic specifications of the system. We will not give a full formal explanation of the defined typology, however, we will view it from a more informal point of view.

The whole corpus was made up of documents in French provided by different firms (Sonovision, SITE, Aérospatiale, EDF, Rhône Poulenc, Syseca...). It consists mainly of aeronautical documents (maintenance documents, job cards...), data processing texts (extracts from reference documents or from user guides, or

software error messages), minutes of meetings, extracts from work schedules or from technico-commercial documents.

The aeronautical documents which make up the greater part of the corpus (about 50%), maintenance manuals and job cards, were initially chosen as the basic corpus for the *PNTAO* for several reasons :

- they were available in machine readable form, at Sonovision (which later became the SITE group),
- they corresponded to a considerable commercial need,
- they concerned a sector of strong export activity,
- they were representative of a large number of technologies (mechanics, data processing, fluid mechanics, strength of materials, etc.).

It is for these reasons that they constitute even today, an important source of enrichment for the corpus.

The initial corpus was made up of about 400 pages taken from maintenance and service guides of Marcel Dassault aircraft. Part of the texts, provided with the English translation, were made up of job cards. Each job card indicates how to go about a specific servicing operation. They contain general recommendations and precise instructions :

NOTA :
L'avion doit être amarré lorsque la vitesse du vent risque de dépasser 50 noeuds.
Pour les vents très violents, il est préférable de mettre l'avion sous abri.

A - Préparation du travail
a. Procéder au campement de l'avion (voir 02.11).

B - Exécution de l'opération
01 Déposer les portes 1905, 1907, 4106
02 Visser une attache d'amarrage sur chaque ferrure

The other texts were service notes describing an apparatus or a mechanism, how it functions and the servicing procedures to be followed. The following is an illustration of this second type of text :

section 10 - UTILISATION
11 - Description - Fonctionnement
11.1 - Généralités
Le raccord auto-obturable d'un équipement mécanique a pour but d'assurer le raccordement rapide d'une génération hydraulique à un banc de test et permet la mise en fonction des différents organes de cette génération hydraulique sans perte de liquide et sans entrée d'air.
Le raccord auto-obturable sert également au remplissage et au dégazage des circuits.

The study enabled us to pinpoint the characteristic of these two types of texts. It turned out that their typology was identical. The main difference is at form level : the first type of text contains enumerated instructions, generally described using short phrases without full stops and which do not exist in the second type of text.

The purpose for extending the study was to investigate the possible existing differences amongst various types of documents in various fields. It turned out that most of the literature intended for technicians corresponds to the original typology.

Amongst the other corpora studied (about 800 pages form various fields, e.g. electrical installation manuals, computer manuals and advertising, software manuals and advertising, aeronautical texts, description of composite materials, maintenance documentation for mechanical installations, etc.), we have classified the texts into two groups :

- texts *pertinent* to the initial corpus, in other words those texts whose typology was identical or closely ressembled that of the first corpus,
- *non-pertinent* texts containing new phenomena. This second type of text will probably require a sub-categorisation into other typologies (a task not as yet carried out).

Typology is field independent. We have received extracts from aeronautical documents, provided by Aérospatiale. Some of these texts fall into the non-pertinent category, while others fall into the pertinent category. Furthermore, it was found that texts originating from different fields (data processing or electrical engineering, for example) fell into the typology.

Amongst the non-pertinent texts we found mainly texts which require rewriting rather than translation, for instance user manuals (at least those intended for non specialists), scientific articles, advertising and legal texts. Such texts are obviously less adapted to MT. We also found minutes of meetings and client service reports, whose syntax is often very fanciful.

As far as quantity goes, the volume of technical matter is much greater than that of advertising and legal matter (Cf. Van Dijk consultancy figures). Maintenance and service guides, intended for specialists, and reference guides represent a much greater volume than that of user guides, intended for non-specialists (which are less homogenous typologically speaking).

2 - RESULT OF THE CORPUS STUDY

The study laid the emphasis on the problems directly related to translation, in particular with respect to accuracy.

The sub-sections which follow illustrate the defined typology and instance the linguistic phenomena not encountered in the so-called pertinent texts studied. Several points must be underlined :

1- any phenomenon classed as being absent may nevertheless be marginally present in a text ;
2- the corpus study showed that few of the phenomena absent from the initial study were later added.

2.1. FREQUENT PHENOMENA

It is not possible to list fully all the phenomena encountered. We will however endeavour to give the main features of the typology of the texts studied, giving the most unusual or significant examples (from the pertinent corpus). The following hence plays more of an illustrative role than a descriptive role.

Punctuation

1 - Commas

Generally speaking, puntuation is used somewhat irregularly. There are few commas, and they do not always appear where they would normally be expected. Consequently, it appears pointless to use them for linguistic purposes, for example, in the determining of the limit of a coordination.

2 - Full stops

As previously mentioned, the job cards come in the form of enumerated operations. Each operation is described by one or more small paragraphs containing sentences. As a rule, there are no full stops indicating the end of a paragraph :

01 Ouvrir les verrières
NOTA :
L'inspection extérieure de l'avion s'effectue par le pilote accompagné du mécanicien. Elle est exécutée dans le sens des aiguilles d'une montre, en partant de la gauche du poste avant
02 Contrôler que...

Enumerations

Every possible type of enumeration can be found. In each case, processing is delicate :

S'assurer que :
- l'assiette de l'avion reste voisine de la ligne de vol
- la manoeuvre ne présente pas de danger

NOTA :
...le mécanicien doit :
- s'assurer de l'absence de fuites d'huile...
- inspecter le revêtement : sur les éléments en stratifié listés ci-après, s'assurer de l'état de la peinture et de l'absence d'entaille ou blessure
  - karmans voilure-fuselage intrados et arrière
  - saumons voilure

Infinitive clauses

They are very frequent in the texts, especially as operations to be carried out are described using the infinitive form (and not using the imperative form : *faites ceci*, or personal form : *vous faites ceci*). They appear as the main clause of a sentence, as an object or as an adverbial :

Après avoir soigneusement asséché les partie externes du raccord, soumettre celui-ci à une pression interne de 0,6 bar pendant 5 minutes...

...il est préférable de mettre l'avion sous abris.

Ne pas tendre exagérément les cordes.

Agir sur la valve pour la faire reculer.

Infinitive clauses introduced by the preposition *à*, with a passive adjectival value are frequently encountered. They bear a certain modality :

fixer un obturateur sur la buse d'entrée du circuit d'air à refroidir *(to be cooled)*

... l'effet à obtenir *(to be obtained)*

Les coupures de frettes ne sont pas à prendre en considération

Finally, the various infinitive clauses can be inter-coordinated or coordinated with different types of clauses (object or adverbial clauses). Coordinations (apart from enumerations) are generally limited to two or three clauses.

Conjugated verbal clauses

We class a conjugated verbal clause as being any clause governed by a conjugated verb. The clauses encountered may appear either as a main (or independent) clause or a subordinate clause (object or adverbial). All the common types of clauses can be found (personal, impersonal, active, passive...). Coordinations (enumerations aside) are generally limited to two or three clauses. Examples of characteristic sentences :

Pour les vents très violents, il est préférable de mettre l'avion sous abri.

Si le liquide polluant s'infiltre dans la zone située entre la jante de la roue et le talon, il faut impérativement démonter le pneu pour le nettoyer.

Amongst the remarkable clauses, we encountered several examples of inverted subjects (which could pass as a stylistic turn unlikely to be found in a technical text). For example :

Au cadre 19 se trouve un boîtier étanche...

Relative clauses

Very few are encountered in the texts (on average 1 relative clause every 2 pages), most of them are introduced by the relative pronoun *qui* (for more than 95% of the texts studied). Here are a few examples :

Il est basé sur la conductibilité thermique... et sur le mode de circulation adopté qui permettent un échange... entre les deux circuits d'air qui la traverse.

... elles possèdent chacune un anneau à travers lequel passe la sangle de rappel...

Noun phrases

We class noun phrases as being all groups with a nominal value whether or not they are introduced by a preposition. It is likely that all the possible types of noun phrases exist in the corpus. We observed with interest that in some cases, the groups were quite often barely grammatically acceptable. The principle of juxtaposing nouns in French is a rare phenomena. All the same, juxtaposition is quite a generalised pratice in the texts studied (this is probably the case for a lot of technical texts). The following juxtapositions are normally accepted :

l'antenne TACAN
l'ensemble raccord-bouchon
un système de type baïonnette

Whereas the following are less acceptable grammatically speaking :

l'ensemble de hissage avion
longueur partie lisse
l'axe avion

The following complex examples illustrate the degree of complexity noun phrases may attain (the part of the sentence which is not the noun phrase is bracketed) :

(Il procède) de la technique des refroidisseurs à surfaces secondaires, appelés refroidisseurs à lames et à intercalaires ou refroidisseurs compacts qui convient particulièrement bien au gaz ayant un mauvais coefficient d'échange thermique.

Verb nominalisation (verb-action noun derivation) is also frequently used :

... effectuer une vérification du tarage...
(rather than *vérifier le tarage*)
... procéder à la dépose des panneaux...
Faire une vérification du réglage...

These structures are built using a small number of French verbs (*faire, effectuer,...*), and they are used with all nominalisations of verbs of action.

Idiomatic expressions

We class idiomatic expressions as being variable or non-variable expressions, which require identification within a text in order to ensure correct translation. Idiomatic expressions exist in all the syntactic categories. Nominal idiomatic expressions are probably the most frequent, particularly in specialised terminology (more than 80 % of the terms are nominal idiomatic expressions, in aeronautical terminology).

Without going into further detail, it can be said that they are all present in the corpus. However, only a relatively small number of verbal expressions can be found (in comparison with the potential wealth of the French language) and there has been no example of a transformation of these expressions (for example *les mesures qui doivent être prises*) in the corpus. They are generally made up using a limited number of verbs (*prendre, tenir, mettre...*) :

Maintenir en place l'embase...
Cette pression de remplissage tient compte d'une perte de charge...
... mettre les commandes du régulateur oxygène sur ON...

We must bear in mind that expressions of the *effectuer*-plus-a-nominalised-verb type do not belong to the specific idiomatic expression class, and are considered as being a commonly accepted transformation.

2.2. RARE OR NONEXISTENT PHENOMENA

The following is a list of significant examples of linguistic phenomena rarely or never encountered. It is not an exhaustive list.

Interrogative clauses

No direct nor indirect interrogatives were encountered in any of the texts.

Imperative clauses

No imperatives (imperative mood) were encountered in the pertinent corpus, they are replaced by infinitive clauses (such is the case for most service or maintenance notes in French) :

procéder au campement de l'avion
ne pas tendre exagérément les cordes

Amongst the non-pertinent texts, messages for users of a data processing system, user guides or training guides may contain imperatives or (sometimes) interrogatives.

Direct speech

No examples of direct speech were encountered.

Comparative phrases with a comparative reference

Comparative phrases containing a comparative reference were extremely rare in the corpus studied. Only one example was encountered :

La bâche à eau est plus volumineuse que la cuve d'évaporation, composée d'élément en alliage léger, soudés.

Concessive clauses

No concessive clauses were encountered.

Clausal subjects

No examples of personal clauses with a syntactic role of subject were encountered. It is hence possible to find :

...Il est recommandé de ne pas inverser la palette...

but not :

Inverser la palette n'est pas recommandé...

Human personal pronouns

Although not totally absent, personal pronouns are rare in technical texts (one per page on average). None of the personal pronouns encountered referred to a human (in fact, humans are rarely referred to in the texts) in the pertinent corpus.

This information meant that for the French-English system, personal pronouns would always be "it" (or "they") as it is not necessary to search for the reference.

Rhetorical figures

Rhetorical figures are rare and correspond to established usages.

Only one metaphor was encountered in the texts :

... les commandes cheminent sur le coté droit sous le plancher passager...

The problem is solved at dictionary level, given that a *cable* can *cheminer* (*make its way*).

The metonyms encountered appear to be accepted metonyms which are transposable into the target language. Thus, in the above example, the term *commandes* which represents the cable or cables which propagate the controls, is literally translated into English, without shocking the technicians of the field.

Anaphora are rare, and as with personal pronouns, if encountered could be handled in a simplified way.

Finally, with regard to the aspectual phenomena, in French, syntactically speaking, there is very little aspectual indication. English, which was our target language very often uses the progressive form. We therefore paid particular attention to the reconstitution of the progressive form. After having carried out a comparative study of the texts, we were able to draw the conclusion that the progressive form was virtually nonexisent in the translated texts.

5 - CONCLUSION

The corpus study has enabled us to draw several conclusions. Firstly, it has enabled the definition of a text typology which is pertinent to a considerable volume of texts of different origins and different technical fields (although this excludes in particular computer user guides and documents for the general public).

The definition of this typology has enabled us to draw up accurate linguistic specifications, simplifying and sometimes ignoring the handling of certain linguistic phenomena. The specifications were then used to develop an MT system (based on GETA's system, ARIANE), whose grammars are less complex, and hence easier to maintain. The quality of translation, tested on a important number of aeronautical texts (maintenance guides) and on several specimens of pertinent texts, have proved the validity of this approach : revision times varied from 20 to 30 minutes per page (average translation times in SITE is about 1 page per hour).

The study of texts of different typologies has enabled us to pinpoint the limits of a minimal language, used in all types of texts, and hence the indispensable kernel of any new typology.

From an economic point of view, the texts targeted by the typology described here represent a considerable amount of the documents to be translated.

BIBLIOGRAPHY

[1] BACHUT, D., & al., "Industrialisation d'un système de TAO français-anglais pour la documentation technique", Génie Linguistique 91, Versailles, 1991.

[2] BOITET, Ch., & al., "ARIANE-78 : an integrated environment for automated translation and human revision", COLING-82, Prague, 1982.

[3] BOITET, Ch., "The French National MT-Project: Technical organization and translation results of CALLIOPE-AERO", IBM Conf. on Translqtion Mechanization, Copenhague, 1986.

[4] BOITET, Ch., "Current Machine Translation systems developed with GETA's methodology and software tools", ASLIB, London, 1986.

[5] BOITET, Ch., "Current state and future outlook of the research at GETA", MT Summit, Hakone, 1987.

[6] CHANDIOUX, J., "METEO : un système opérationnel pour la traduction des bulletins météorologiques destinés au grand public", Meta 21, 1976.

[7] CHAPPUY, S., "Formalisation de la description des niveaux d'interprétation des langues naturelles", Thèse 3e cycle informatique, Grenoble, 1983.

[8] HUTCHINS, W.J., "MACHINE TRANSLATION : past, present, future", Chichester, Ellis Horwood series in Computer and their applications, 1986.

[9] KITTREDGE, R., & LEHRBERGER, J., "Sublanguages : study of language in restricted semantic domains", Berlin, de Gruyter, 1982.

[10] VAUQUOIS, B., & CHAPPUY, S., "Static grammars : a formalism for the description of linguistic models", International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Language, Colgate University, 1985.