

TRAINING IN THE DRAFTING OF TEXTS FOR DOCUMENTATION
IN CONTROLLED LANGUAGE FOR THE PURPOSES OF AUTOMATIC
TRANSLATION

Françoise Blamoutier

Head of information

Institut Economique et Juridique de l'Energie, Grenoble

Abstract

Account of an educational experiment carried out with students of information sciences and with searchers responsible for indexing documents within the context of a sectoral network - to train them to prepare abstracts in controlled language.

The guiding principle is as follows: if an abstract is 'good' in the documentary sense of the term, then the automatic translation of that abstract will present a minimum of problems and the result obtained will also be 'good' (again for documentary purposes) in the target language

On the assumption that the data processing tool exists and that the degree of complexity of the indexing metalanguage has been determined as a function of the field concerned, the quality of the result will depend upon:

- 1) the quality of the linguistic tools: dictionaries and their translations, interplay of relationships;
- 2) the ability of the indexers to manipulate these tools and to extract as much information as possible from a text for a given type of user, taking into account the constraints imposed on him.

The following topics are covered: reasoning as taught to students, conditions in which linguistic tools are fashioned and training methods used.

We give below an account of an educational experiment which has been running for three years with third year students taking a one-year specialized course in information sciences and with searchers who are part of an international cooperative information network.

This experiment is based essentially on the interrelationship between: indexing (in the widest sense of characterization of the text), translation and interrogation. We shall present the experiment as follows:

- 1 - documentation-translation relationship as presented to the students;
- 2 - linguistic tools required for automatic translation of documentary texts ('abstracts');
- 3 - training of students and searchers in the preparation of abstracts in controlled language.

1. DOCUMENTATION AND AUTOMATIC TRANSLATION

The first consideration which led us along these lines stemmed from an observation: the difficulty experienced by students faced with non-integrated training, as is often still the case in France, in which courses in linguistics, data processing documentation, library science, etc. seems to be pieces of a patchwork, in spite of the goodwill of specialist teachers of these subjects who are victims of their own training or of the traditions of their profession and who at best know about the work done by their colleagues, without having had first-hand experience of it.

We have therefore thought it worthwhile not only to coordinate the various types of training, but to give them a single guideline: documentary purpose, whereas up to the present attempts have been geared chiefly - and incidentally with little success - to developing joint methodologies, in particular between linguistics and documentation. We have abandoned this approach which we

feel creates more confusion than clarity.

The case of automatic translation is typical in this respect: when this expression is used imprecisely one thinks immediately of experiments in translation from natural languages, as carried out in Grenoble, for example. What is the aim of this translation? To reproduce in a target language A' the image of a text in a source language A as faithfully as possible, with all its verbiage, even ambiguities sometimes, its ellipses, all aspects of the argument, including personal aspects of the author's style.

In contrast, the documentalist starts with a set of texts in natural language and endeavours to screen them according to predetermined reference axes in order to sift only significant elements for a given type of user. (see Fig.1).

Not only does he eliminate as much of the discursive part as possible insofar as it is irrelevant, but he disregards form itself in favour of real signification and in certain instances goes so far as to compensate for gaps in expression by extracting concepts which were not explicitly stated in the texts analysed but which are useful access keys. Indexing has no ultimate purpose in itself - it is merely the operation necessary to make interrogation possible and the objective in a bibliographical system is to provide the user with keys which will enable him to extract from a system the addresses of documents containing the maximum of elements in answer to his question.

The approach of the documentalist analysing a text is essentially semantic and the metalanguage he uses for communication with the user must be rigorously unambiguous in this respect, whatever its degree of syntactical

complexity which may be reduced to its simplest expression compatible with the aim pursued.

The indexing metalanguage, by virtues of the rules imposed upon it in order to exclude ambiguities and gaps, which are factors of noise or silence in interrogation, may be used as the translation metalanguage for the same purpose, that is a documentary purpose: possibly for production of the same documents (in language X, Y or Z) in the target language in reply to a single question.

We could speak of automatic translation in this case (of abstracts or indices) but subject to very different conditions from the first case and the students must appreciate fully that one may arrive at the operational stage under acceptable operating conditions (cost, time and ... results) in the case of texts prepared for this purpose whereas the translation of complete texts in natural language is a search operation carried out in a totally different environment, even though present methods tend to be similar, due in particular to the interest shown in the semantic approach.

Some work can of course be carried out in joint laboratories and automatic indexing is still an attractive dream which could be realized only by the success of researchers working on the analysis of natural language, but we should like to remain here within the very practical framework of operational systems for which costs (beginning with acquisition) are of paramount importance.

We therefore come back to documentary analyses made by human means - at this juncture still the more economic and productive - and their possibilities for automatic translation. By virtue of a hypothesis put forward by

Gardin ¹ and confirmed in actual practice i.e.

'if we succeed in formulating rules for correspondence between one natural language and a given metalanguage it is probable that we shall be able to formulate such rules for other natural languages vis-a-vis the same metalanguage' - if an analysis metalanguage is good in one language (good: that is, giving the highest pertinency and recall ratio upon interrogation), it has every chance of being good in the other languages provided that:

1. the reference tools exist: descriptor vocabularies, common function words, functors (with a very strict delimitation of their use and their translation within the same limits) and the interplay of possible random relationships between the descriptors (definition of a simplified grammar);
2. the indexer knows exactly how to use them within his field and makes optimum use of them.

The quality of indexing and of translation (which are interrelated) rests on work done at two levels:

- the level of 'once and for all' during elaboration of multilingual vocabularies and the definitions of possible relationships;
- the level of 'step by step' during each indexing.

(This presupposes of course the existence of the data processing tool; we are not concerned here with the technical aspect). To illustrate this we shall take as an example work carried out in the field of energy economics:

1

J. C. Gardin 'Linguistique et documentation', Boll;
d'Informazioni, 1973, No. 2-3, op. 67 - 86.

We shall not here describe the way in which the linguistic tools were developed ¹ but it is of course necessary to describe this in detail to students who must understand all the mechanics of it, and to explain the main guidelines to searchers who will require a complete knowledge of the art and manner of using the system.

2. ELABORATION AND TRANSLATION OF VOCABULARIES

First the list of descriptors: this has been arranged in the form of a faceted thesaurus published in 1974 after several years' work . The system of permanent hierarchical and logical relationships which defines the semantic field of each descriptor in the area concerned is thoroughly scrutinized. It may be advisable at this level, in order to avoid any confusion, to abandon the term 'documentary language' often used to refer to the thesaurus, since the word language must not be allowed to imply the existence of an intrinsic grammar. There is still scarcely any orthodoxy in the vocabulary of information sciences and 'languages' and 'different types of language' are not always distinguished consistently *. It is better to follow Robert Escarpit ² and use the term 'documentary code' to denote a thesaurus by its purpose. There must therefore be absolute insistence upon this rigorous delimitation of the semantic field of descriptors - defined by the relationships, appurtenance to various facets, the addition of explanatory notes where required,

1. See for example F. Blamoutier. 'Les problèmes de vocabulaire dans le thésaurus de l'économie de l'énergie'. La Banque des Mots 1973, No 5, pp.83-96.

*. Translator's Note

An attempt has been made to render the distinction between 'langue' and 'langage' as used in the original French text - unfortunately only the word 'language' exists in English!

arbitrary delimitations (not made at random but selected by the specialists who have drawn up the thesaurus); these may sometimes be artificial but they are the best safeguard against ambiguity of meaning. Through this definition of the content we shall achieve the optimum quality of indexing in the language of the thesaurus. To obtain the same result in another language it 'is sufficient' to have a version of the thesaurus in that language.

We shall not go into the problem raised by translation of the thesauri and the choice of methods: simultaneous elaboration in several languages, elaboration in one source language then translation etc We shall simply say that we have not yet decided, but that we have made all possible efforts to ensure that the historical order imposed by circumstances does not detract too much from the quality by maintaining a permanent team of specialists (in association with outside experts where necessary) for the elaboration, translation (at present into English) and development of the thesaurus.

Naturally the precision given to the definitions when the system is being set up is very useful during translation: if for example we have defined conservation d'énergie as a comprehensive policy at national level including the saving of energy, the adaptation of sources to uses, the exploitation of renewable resources, etc. then energy conservation will have exactly the same content in English; this is possible only if the correspondence exists. Otherwise we should not translate but find an equivalence: an 'abonné' to the gas or electricity board will not be a 'subscriber' but 'contractual user'. If there is no possible equivalence the choice of

1. 'Thesaurus Economie de l'Energie', Paris, Technip, 1974;
2. R. Escarpit - 'L'écrit et la communication'. Paris, PUF, 1973, p.128 (Coll. Que sais-je?, No. 1546):

the term or the initial hierarchy in the source language is amended: thus the existence of a single word tax which covered at the same time the generic impôt and the specific taxe demolished the original French hierarchy (it is worth pointing out however that gallicisms with regard to structure or usage corresponding to a specific fiscal, political, etc. system, had already been eliminated as far as possible during the elaboration of the primitive form in French, for the simple reason that they would have been of no help at all in analysing for the most part foreign texts).

Thus by successive adjustments the two versions of the thesaurus may be regarded as calques which are as identical as possible and which allow equally rigorous indexing in English and French and reciprocal translation of the descriptors. This possibility has so far been used in the Energy Economics network only for French to English (completion of a bilingual statistical directory in 1975, interrogation in English of the data base). Translation of the thesaurus merely makes it possible to translate the indices and descriptors within the body of the sentences of the abstract. But how do we progress if these sentences in the abstract are composed in a more complex way than by simple juxtaposition of keywords?

First by translating the other vocabularies: common function words and functors (work in progress for Energy Economics) with the same attention to clarity; for example ambiguities in the case of plurivalent words must be removed - either by banning their use or by retaining only one use (for example propre could be retained as an adjective of cleanness and dropped as an adjective of possession); for the tool words which are to ensure the essence of syntactical relationships between the descriptors the least onerous and risky solution is to use them

only in a restrictive way: such daunting words as 'de', 'à', 'en', etc. with numerous uses in French may be completely tamed and will present no problem for translation if they are assigned precise and restricted roles in comparison with their roles in natural language.

If the form in which the abstracts are written is completely free, we come back to the problem of automatic translation of natural languages and from the outset we have stated that we do not consider this to be a matter concerning documentation.

So to remain within our framework we would say that, according to the level of elaboration and syntactical complexity regarded as necessary for the 'abstracts' it would be sufficient either to translate the descriptor vocabularies (the syntax of the abstracts would thus be reduced to the juxtaposition of descriptors in subordinate sentences, to their order, to the use of links or of coded rôle indicators), or to translate the vocabularies of common function words and controlled tool words (if more detailed abstracts are required). Thus a certain number of rules (restrictive in comparison with natural grammar) must be determined, the degree of complexity being limited - let us repeat - to the minimum required for the purposes of documentation (clarity, readability, freedom from ambiguity) .

If all appropriate precautions have been taken during the development of these tools then the value of the result in both the source language and in translation will be determined by the quality of the abstract itself; the training of the indexing analysts in controlled vocabulary and syntax is therefore of paramount importance.

3. TRAINING IN INDEXING

Training of students and training of analysts responsible for feeding a precise system are different and we shall treat the two cases separately. First of all as regards students we should begin by explaining that if they are set to work principally on a concrete example this example must be regarded as a study of a particular case and we must show them - in addition to the reasons for selection of this system - the possible variants and their implications. The system selected for this case study is the SPLEEN system of Energy Economics which lends itself particularly well to demonstration, partly because it is supported by a cooperative collection and indexing network - this requires the existence and respect of a certain number of norms, in particular linguistic norms - and partly because it has for two years been developing towards multilingualism in the stages we have already mentioned, and also because the choice of the indexing metalanguage is well suited to automatic translation. It has a highly-simplified syntax: no verbs in the active voice, no morphological variations of the descriptors (selected in the masculine singular except when use requires the plural, for example 'travaux publics' (public works)) etc. Sentence structure is as follows: The first 'sentence' must include an indication of the concept(s), product(s) and country (countries) concerned plus a date or date bracket (except for methodological documents). It is in fact the most significant title possible for the document. Subsequent sentences indicate the facets and particular aspects of the study, forming a kind of table of contents. Supplementary themes may be added to this 'structured commentary' lists of products, countries, concepts, etc. set out in simple form; the descriptors have exactly the same value for the indices or for search purposes whether included in the commentary or set out in the supplementary themes. It is a simple question of readability (see Fig. 2).

So why have we not selected only lists of descriptors linked together by rôle indicators, for example? Because within the field in question the readable 'abstract' is important. In fact the further one goes from scientific fields towards technical or technico-economic fields and still more towards social and human sciences, the less possible it becomes to reduce the argument to significant factual elements alone, and the greater the role of relationships, elements for reflection and interpretation which form the context of these data. It is possible to express these 'ideas' with the aid of a highly simplified, but not non-existent, syntax and the so-called 'common function' words are also significant. The acceptable limits for the simplification of rules are once again those of comprehension. It is a question of finding the metalanguage which, we must admit, may be inelegant but which is in any case non-ambiguous and which will produce the optimum cost/effectiveness ratio.

It must be made clear to the students that we are dealing with the subject from the point of view of the practitioner and not the searcher - the more necessary for them since they intend to become practitioners themselves ¹.

The abstracts (structured commentaries) may vary in length, they may be of an informative rather than an indicative nature, with a more or less restricted syntax, etc. depending upon the aim in view and means available, the decision depends upon a policy to which the techniques would need to be adapted. With this in mind exercises

1. These courses are intended for DESS (Diplôme d'Etudes Supérieures Spécialisées) students, third short cycle (one year) for professional training following a course of four years at the minimum in various disciplines: the exact sciences, biology, economics, legal sciences, etc.

are based on the existing SPLEEN model. The drafting of abstracts constitutes the only stage of human intervention in a process which will subsequently be entirely automatic: they are used both for reading this commentary section, for the compilation of the indices and inverted files by isolating the descriptors in the text between suitable separators ¹, for sequential search by character string to allow interrogation on words or word-groups in the case of waiting descriptors representing new concepts not yet contained in the thesaurus, etc. (see Fig.2).

They must also be adapted to automatic translation. Care must be exercised in this operation - above all, as far as students are concerned, as regards technical training bearing on compliance with constraints since they are not specialists and we cannot therefore expect them to carry out in-depth indexing. On the contrary it is fitting to take advantage of the complementary nature of their original training to lead them to discover all the facets of a given document. All citation studies confirm that there are many standpoints from which a text may be examined and make some contribution to knowledge.

Thus a text on the UN conferences on maritime law will be analysed by a lawyer, an economist, an ecologist, a geographer, etc. - first in free vocabulary but with controlled syntax, then in controlled vocabulary using existing lexicons or, where this is not possible, by 'making it up' with the students for the purpose in hand, lastly by using thesauri; all the possible questions suggested by the students on the texts analysed are put and translated into appropriate interrogation terms (with increasing degrees

1. Special separators are assigned to the 'names cited' keywords consisting of proper names which are not recognizable in any pre-existing corpus since the list is not 'complete'. They will be classified as such in the inverted indices and dictionaries without being checked in advance and will also be reproduced as such in translation.

of complexity: on the indices, on cumulative indices, on the data base possibly in conversational mode) with the same tools all the resources of which are itemized and used.

The training of analysts, searchers or engineers working on the input of a given system follows completely different principles: the limits and constraints of the system, the rules and the tools are imposed upon them in terms of choice and prior work. The aim is to optimize the procedure in three stages: keyword (in the original meaning of the term: word which gives the 'key' to access to the following stages) transformed into a descriptor - abstract or structured commentary (the aim of which is to add the 'concepts' to the data in order to provide more specific background information on the original text) - primary document. The degree of user satisfaction will evidently depend upon accurate screening at these three stages, that is upon the elimination of noise, and also upon any knowledge which the indexer may have added to the explicitly expressed descriptors in the initial text of implicit concepts likely to be good access keys, good 'keywords' to the texts analysed, thus limiting the silence.

It follows that this work can be done only by a specialist: it is never a question of describing the texts - and this for two reasons: the terms which they contain (even if replaced by their equivalents taken from the controlled vocabulary) may be false friends: they may be used incorrectly (for example 'besoin' (need) frequently used in the exact sense of 'demande quantifiable' (quantifiable demand) in economic texts), negatively (for example it is impossible to estimate reserves), or abusively (as frequently happens in technico-commercial texts), etc.; furthermore the appropriate terms for interrogation may be lacking and may have to be derived through reflection.

If in-depth indexing meets these two requirements and produces a structured commentary which takes account of these requirements whilst observing linguistic constraints, then the commentary has every chance of producing the best search results in the source language and a good translation in the target language which in turn produces the best search results in that language; this is exactly the aim set.

How do we persuade indexers to comply with these constraints? First of all by motivating them: any searcher or engineer who makes a contribution to the feed-in system will sooner or later become the 'client' of that system; there are few exceptions to this rule, except professional analysts and with a little imagination they could 'simulate' the necessary interrogations.

How do we subsequently train them? The first task is to place them in the position of potential interrogators: 'to which questions, can this document provide response elements?' and to check that the keywords in their questions are in fact reflected in the descriptors put forward in the abstract.

Vocabulary is more efficiently controlled by direct use of the thesaurus by the indexer than a transformation posterior of keywords into descriptors - first because it does not involve any error of interpretation, and secondly and above all, because it compels the indexer to be precise in his thinking, thus avoiding the totally useless 'artistic haze'. (This pedagogical role of the thesaurus might also prompt him to use it even when it is not essential to work in controlled vocabulary).

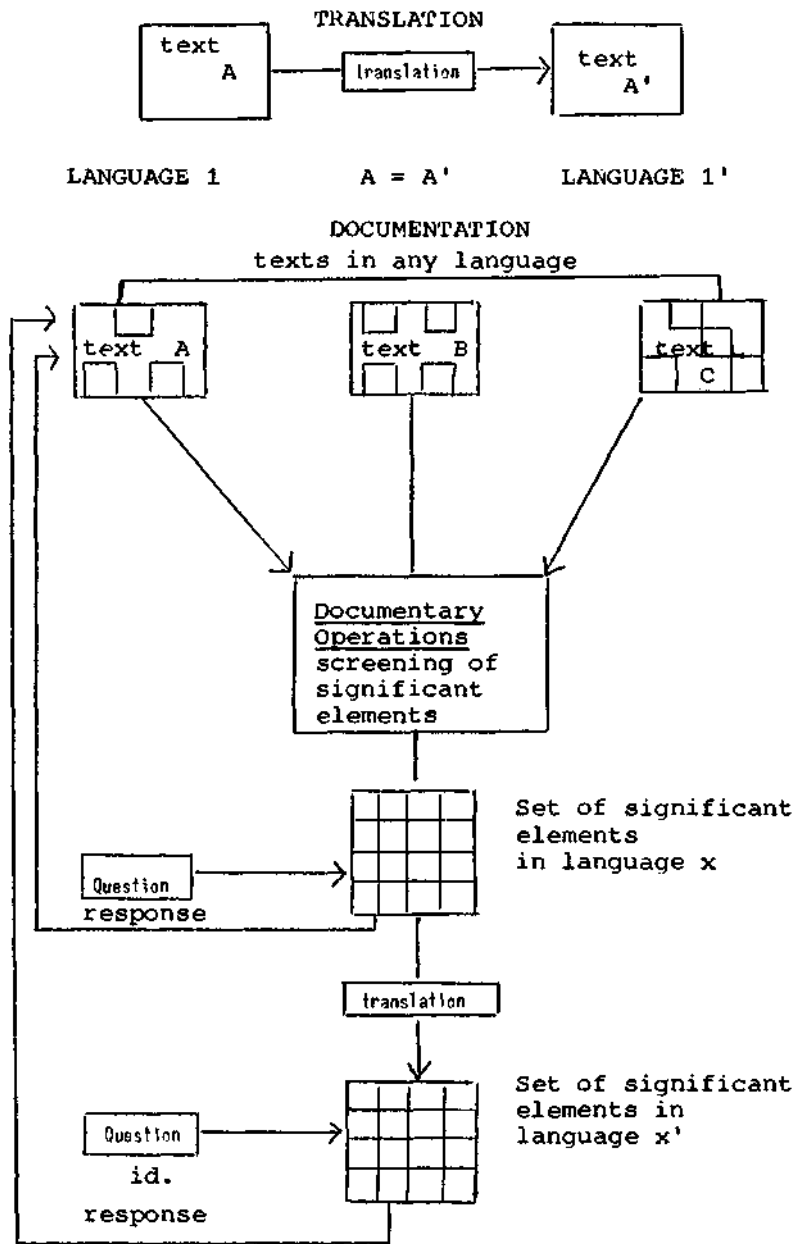
With regard to control of syntax, in view of the simplicity of that of SPLEEN, there has never been any problem; common sense and a knowledge of the existence of

constraints, plus a few rules, are sufficient to carry one through without any great difficulty.

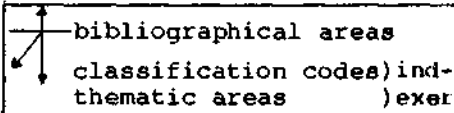
This, at any rate, is the case with searchers at the Institut Economique et Juridique de l'Energie at Grenoble; since searchers and documentalists work permanently side by side it is possible to prepare abstracts after a mutual consultation, each contributing his own knowledge and experience (of the subject; of the system). If permanent dialogue is not possible then training may be carried out in seminars, and refresher courses arranged at regular intervals (since relevant topics, the system and the vocabulary are constantly changing). One very effective means which saves time is use of conversational mode, not only because it is valuable in providing 'motivation by interrogation', as mentioned above, but also because of its pedagogical value in 'successive approximation', but this is an expensive method!

Nevertheless, whatever method is used we believe that the quality of indexing in controlled language is the best guarantee of its 'translatability' under acceptable conditions of cost and quality; and that the training of future indexers - i.e. today's students (of information sciences and also other students within the context of what we call 'user training') - and the further training of analysts already part of a given system, are essential to the achievement of this quality. This is why we feel that it would be false economy to cut back pedagogical investment in this field; if we neglect this task it is to be feared that the increasingly sophisticated system currently being set up will be neither fed nor consequently used to good effect. This consideration applies to both monolingual and - to an even greater degree - multilingual systems and is worthy of some attention on our part.

Fig.1.



Full translation of text A into another language will be undertaken only at a later stage if the interrogator, who has received the indication of this text in response to his question, cannot understand the language in which it is drafted.

area codes			wording	Fig. 2.
P	R	Z	01/1	
P	A	T	UNITED STATES/US Department of Commerce/ / National Bureau of Standards/NBS/ Washington,D.C./	
P	R	Z	02/1	
T	I	T	Evaluating incentives for solar heating	
A	U	T	RUEGG (R.T.)	
O	R	G	Inst.for applied Technology,Building Economics, NBS	
V	I	L	Washington, D.C. 20234	
E	D	I	NBS, U.S. Department of Commerce	
D	A	T	1976	separator codes: // concepts ++ products ?? countries == names cited
T	Y	P	4	
P	A	R	NBSIR 76-1127	
P	P		61 p.	
A	N	N	table,reference,bibliography	
L			E	
D	C	A	/September/1976	
N	D	O	IEJE - 21057E	
R	U	B	2	
T	H	O	(/Economic analysis/and elements of a/policy/ on incentives for the use of+solar energy + for/domestic heating/ in the ?United States?/ 1976-0000/:	
			(/Case study (by region)/on/competition+ between+fossil fuel + and electrical energy+ and suitable measures: (/tax relief/ /subsidy/).)	
T	H	3	= NCSL. National Conference of States Legis- latures (United States) =	
T	H	4	/penetration / /market/ / thermal usage/	