Corpus-based annotated test set for Machine Translation evaluation by an Industrial User

Eva DAUPHIN and Véronika LUX
AEROSPATIALE - CCR
12, rue Pasteur, BP 76
92152 Suresnes Cedex - France
eva.dauphin@siege.aerospatiale.fr
veronika.lux@siege.aerospatiale.fr

Abstract

This article is concerned with the building of a test data set for assisting the industrial user in machine translation evaluation. The emphasis is laid on the interest of an approach based on the study of bilingual corpus pragmatic characteristics. The study of one chapter of the maintenance manual of the Super Puma helicopter made it possible to identify the pragmatic characteristics relevant in the choice of the morpho-syntactic structures and translation processes actually used. The textual test set consists in a SGML file including the source text sequences aligned with the reference translation sequences and also including the pragmatic, formal and translational characteristics in the form of annotations (labels and formal descriptions).

Introduction

Corpus studies appear to be one of the most appropriate techniques to identify the linguistic constraints and needs which will be used as evaluation measurements and criteria to judge the adequacy of a machine translation system to an industrial user's environment. In this article, the linguistic constraints correspond to the linguistic characteristics of the corpora to be treated by the machine translation system. These constraints are illustrated in the source language corpora to be submitted to machine translation. The linguistic needs correspond to the minimal level of quality required for the produced translation. These needs are illustrated in a human attested translation of the chosen source corpora. The identification and formalisation of the linguistic constraints and needs illustrated in corpora represent a major step during the evaluation process of machine translation applications by an industrial user. Corpus study is not a new concept in the NLP domain, but the methods used can be quite different depending on the expected results and applications. In this article, we will describe how we built a reusable annotated test set through the study of a bilingual corpus.

1 A highly structured documentation

The corpus we chose to study is the maintenance manual for the Super Puma helicopter written in French and its attested human translation in English[1]. An important characteristic of this corpus is that it is written and used in compliance with the so-called ATA[2] 100 specification.

1.1 ATA 100: a short presentation

The ATA 100 specification role is to provide a set of rules for the writing and exploitation of aircraft after-sale documents. Both document writers and users are supposed to be familiar with this specification. As such, the ATA 100 specification defines the document production and use environment. In particular, the specification imposes a very strict way of structuring the text in terms of topical and discursive organisation, and also in terms of practical document production by providing a dedicated SGML DTD for the maintenance manuals (ATA 100 DTD). The relevance of ATA 100 in the writing and exploiting process of the document encouraged us to consider the corpus from a quite new point of view in the NLP domain: the pragmatic approach. The ATA 100 specification can indeed be considered as a sociolectal system or standard (i.e. a system that rules the communicative usage inside a restrained community of persons).

From a communicative point of view the ATA specification defines the types of discursive genres (illocutionary force of the utterances) the writer has to adopt according to a pre-defined document structure. For example, all the maintenance documents must be divided into tasks and subtasks. Each maintenance task description should be preceded by a definition of this task. Any task to be performed should be described as a succession of subtasks which are explained in the form of a succession of orders. Each task and subtask has a precise denomination that takes the form of titles in the document. Because the maintenance manuals are submitted to annual updates, the document contains also a large number of factual information such as dates, version numbers, aircraft type reference, page numbers etc. The production and

---

[1] Because of the huge volume of the document (7 Mo of plain text), we chose to select a representative sample of it : chapter 10 - Parking and Mooring.

[2] Air Transport Association of America [ATA:92]

exploitation environment having a strong impact on the way the documents are written, it was quite natural to first characterise the corpus we were intending to study from a pragmatics point of view.

1.2 Pragmatic labelling of sequences

For us, the pragmatic labelling was a first step in the classification of utterances based on their communicative value. Practically, we decided to assign to each utterance of the text a label indicating its textual and discursive status according to the ATA 100 indications. The pragmatic study of the corpus resulted in the definition of 4 types of labels: the meta-textual indicators, the topical meta-utterances, the discursive meta-utterances and the illocutionary typed utterances (orders, definitions, etc.). For example, the METNORM and METXNORM labels correspond to the Task and Subtask titles in the text:

Stockage des instruments
Déstockage des atterrisseurs
Remise en service de l'appareil

The illocutionary aim of the METNORM or -XNORM is to help the user understanding the topical organisation of the document. They actually illustrate the operational organisation of the maintenance work to be done by the user.

2 Underlying syntactic behaviours

The second step of the corpus study consisted in the observation of the formal structures of the utterances previously labelled from a pragmatic point of view. As we can notice in the above examples, the pragmatic value of the utterances has a very clear incidence on their morpho-syntactic structure.

2.1 Some observations

In our corpus, we observed that the Meta-Textual Indicators usually present a phrasal structure (they are not complete sentences) and include a large number of brachygraphical signs (acronyms, codes, alpha-numerical references, etc.). The meta utterances are all nominal phrases resulting from the nominalisation of verbal groups:

Stockage des instruments
Déstockage des atterrisseurs

or from the topicalisation of an object:

Éléments stockés en containers pressurisés

Nominalisation and topicalisation can thus be considered as processes used by the writer to "textualise" knowledge for its reader.

As far as the Directive - Operation Utterances (EDOPER) are concerned, we also could observe very strong regularities which are all based on the same morpho-syntactic basic scheme: VERB + OBJECT, the verb is always an infinitive one, the real subject (the reader) is never mentioned and adverbial complements may be inserted at specific places in the sentence depending on their semantic value (time, manner, place, mean, etc.).

2.2 Typical morpho-syntactic schemes

This type of morpho-syntactic observations has been carried out for all the utterances of the text and resulted in the definition of twelve morpho-syntactic basic schemes presenting the characteristics of the linguistic structures used by the writer. Our morpho-syntactic schemes are based on the concept of syntagmatic components which are further specified using a set of features. We used two kinds of features: morphological features (tense, mode, voice, derivation, etc.), and functional features (manner adverbial complement, direct object, subject, agent, etc.). For example, most of the Topical Meta-Utterances (MET) correspond to the morpho-syntactic scheme SNDEV which is the following:

$$N + (AJ) + (SN_1) + (SP_1|SN_2)^3 + (SAV|SP_2)^4$$

with the following features:

- N: *deverb* = *+*, indicating that the noun (N) is the result of the nominalisation of a verb (deverb)
- AJ: *fonction* = *épithète*, indicating that the adjective (AJ) has the function of modifier,
- SN_1: *fonction* = *COMPADV* and *type* = *temps*, indicating that the nominal phrase has a function of adverbial complement (*COMPADV*) with a "time" semantic (*type* = *temps*)
- SP_1: *fonction* = *dev-OBJ* and *prep* = *de*, indicating that the prepositional phrase has a function of object of a nominalisation (*dev-OBJ*) introduced by the preposition de (*prep* = *de*).
- SN_2: *fonction* = *dev-OBJ*, indicating that the nominal phrase has a function of direct (there is no prepositional introducer) object of a nominalisation (*dev-OBJ*)
- SAV and SP_2: *fonction* = *COMPADV* and *type* = *manière*, indicating that the SAV and/or the SP_2 have the function of adverbial complement (*COMPADV*) with a "manner" semantics (*type* = *manière*).

The scheme presentation reflects the results of a textual study and in order to formalise some particular phenomena, we had to introduce some specific features such as "*deverb*" for the nouns resulting from the nominalisation of verbs. These schemes are actually generic representations that allow us to characterise all the textual sequences of our corpus using only 12 scheme labels.

2.3 Co-description

The study of the possible co-description of an utterance by a pragmatic label on one hand and by a morpho-syntactic scheme on the other hand made it possible to assess compatibilities and incompatibilites between the pragmatic value of an utterance and its linguistic structure. The following table shows that for each pragmatic value, we can find a typical underlying morpho-syntactic structure.

---

[3] The sign "|" means an "or".

[4] The brackets indicate that the component is optional.

	IMTENT	IMTPDP	IMTLAP	IMTTDM	IMTRDT	EADEF	EADESCR	EDIND	EDOPER	METNORM	METXNORM	MED
SNDEV										X	X	
SNDEV2				X	X			X				
SNTHEM										X	X	
SNTHEM2				X	X							
SNADET	X		X	X	X			X			X	X
SNADET2		X		X	X			X				
SPADV											X	
PACT							X					
PPAS							X					
PCOP			X			X	X					
PINF									X		X	
SBRA	X	X	X									

At this stage of the corpus study, we described each textual sequence of our text with two labels: a pragmatic one (indicating the textual and illocutionary status of the sequence) and a morpho-syntactic one (describing its formal behaviour).

3 Underlying translational behaviours

The third step of the corpus study intended to show that it was possible to add translational information on the already obtained pragmatic and morpho-syntactic information. To get these translational information, we carried out a contrastive study of our French text with its attested human translation.

3.1 Some observations

The pragmatic characteristics of the text apparently implied some translation choices. Indeed, for us, only the directive illocutionary value of utterances could explain the choice of translating the French infinitive verbs in English imperative verbs. Infinitive, in French has no intrinsic value of imperative ; all the infinitive verbs in French are not necessarily translated by an imperative verb in English. Also, the pragmatic phenomena of "terminologisation" which are not the same in French and in English explain the possible structural non correspondence between some nominal phrases:

Appareil entreposé non stocké

--> *Aircraft Stored-No Preservation Measures*

or between some sentences. The keeping of the pragmatic value from French to English can also lead to the restitution in English of some missing elements in French:

rotation des roues

--> *rotate the wheels.*

3.2 Translational annotations

The contrastive study led us to identify some recurrent (in our corpus) translational consequences due to the pragmatic value preservation from French to English. This part of the study resulted also in the identification of some phenomena that have to be strictly formalised if we want them to be correctly handled by a machine translation system. This is the case for the terminological elements that may have quite unpredictable translations (or equivalents):

GTM --> Engines
Circuits anémo-barométriques --> Air Data Systems

Concerning the problem of term translation, the best solution we found for annotating the test set consists in tagging them in the French sequence and in the English corresponding sequence using SGML tags (<T> and </T>):

Vérifier la <T>BTP</T>

--> *Check <T>MGB</T>*

Concerning the morpho-syntactic translational observations, we chose to express them in the form of an "oriented rule" attached to the French-English pair of corresponding sequences. If we take the example of determination in the Topical Meta-Utterances (topical titles), we observe that the nominal phrases which are the object of a nominalisation are nearly always determined:

*Déstockage de **la** structure*
*Nettoyage **des** parties métalliques*

whereas they are systematically undetermined in English:

Depreservation of airframe
Cleaning of metal parts

The annotation concerning the omission of determiner in English is the following:

V + SP --> V + SP(DT-)[5]

and is attached to the concerned pairs of English-French sequences.

4 Building an annotated test set

The corpus study allowed us to get a large number of information concerning the French text on one hand, and the English text on the other hand. We also have information on the corpus-specific translational processes used from French to English. In order to build the test set, it appeared necessary to structure this information so that it could be exploited by the industrial evaluator.

4.1 Defining an annotation scheme

To structure our annotated test set, we defined a so-called "annotation scheme" and we adopted the descriptive language SGML. The test set is based on the notion of equivalent textual sequence pairs directly extracted from the aligned French-English original studied corpus. Each pair of aligned sequences compose what we called a test unit. The information concerning the French sequence is directly attached to it (morpho-syntactic scheme,

[5] V for Verb, SP for Prepositional Phrase, DT for Determiner.

complete morpho-syntactic description and tagged terms) ; the information concerning the English sequence is directly attached to it (complete morpho-syntactic description and tagged terms) and the information concerning the sequence pair (pragmatic label, factual data) is attached to the created test unit.

4.2 The use of SGML

To really get a structured file of annotated test data, we chose to build it in compliance with the SGML ISO standard. We thus wrote an SGML DTD in order to formalise the conceptual annotation scheme. The result obtained for the following sequence pair:

Stockage appareil complet

--> *Storage of complete aircraft*

is the following:

```
<UTEST NUM="0001" CHAPNUM="10"
PRAGTYPE="METNORM" PARTRAD="1-1">
<SEQS STRGEN="SNDEV">
<LIB>Stockage appareil complet
<LIBD>N_deverbatif + dev-OBJ[SN(N + AJ)]
<SEQC> LANG="EN">
<LIB>Storage of complete aircraft
<LIBD>N_deverbatif_lex + dev-OBJ[SP(PP + SN(AJ +
N))]>
<ETRAD CDLANG="FR-EN">
<TRADN NUM="01">N_deverbatif --> N_deverbatif_lex
<TRADN NUM="02">dev-OBJ[SN] --> dev-OBJ[SP]
```

This format, though a bit complex for an human eye, has the advantage to clearly separate annotations from the original textual data. Moreover, this format allows an easy exploitation of the contained data provided the evaluator uses SGML tools with which selection and extraction of subset of data become really easy (each tagged data is a potential selection criteria).

Conclusion

The interest of building this kind of annotated test set from corpora is multiple. First, it allows the evaluator to have at his disposal a whole set of potential test data which are clearly representative of his real industrial needs. Being enriched by pragmatic and morpho-syntactic annotations, it considerably helps the evaluator to clearly identify the phenomena well or badly handled by a machine translation system. Indeed, using the annotations, the evaluator can easily link the mistakes of a machine translation system with the concerned linguistic units.

The presence of a reference annotated translation is also of great help for keeping the evaluator as impartial as possible (even if a réference human translation may not always be the best one). This is particularly true when dealing with terminology. Finally, using SGML for the building of the test set file allows one to perform targeted evaluations by giving extraction criteria such as the morpho-syntactic schemes or even the pragmatic labels: an evaluator can decide to select all the test units including the pragmatic METNORM label in order to carry out a specific evaluation of the MT system performances when translating the task and subtask titles of the Super Puma helicopter maintenance manual.

As stated above, our initial aim in this corpora study was the building of a corpus-based annotated test set, to be used for the evaluation of MT systems. The results nevertheless seem interesting for some other purposes. Firstly, these results are potential contributions to the specification of dedicated NLP systems. In particular they allow one to suggest heuristics for the processing of linguistic phenomena which, in general, are known to be complex problems for NLP. One example, in the case of MT, is the translation of French determiners into English. Another example, in the case of automatic analysis, is the resolution of anaphora: more than 80% of the pronouns in our corpus refer to the object complement of the last sentence.

On the long run, future MT systems could take advantage of the pragmatic information contained in the SGML tags of the source text, to drive both the analysis and the transfer phases. For example, a verbal form in the infinitive, when occurring in a procedural part of French text (identified as such via SGML tags) would be analysed as a sequence with injunctive value, and translated into English by a verbal form in the imperative. When occurring in a tittle, a similar infinitive verbal form could be translated by an "ing" verbal form.

References

ADAM J.-M., *Éléments de linguistique textuelle, théorie et pratique de l'analyse textuelle*, Liège, Mardaga, 1990.

Traitement de l'information, systèmes bureautiques, Langage normalisé de balisage généralisé (SGML), NF EN 28879, ISO 8879, Paris, AFNOR 1990.

ARNOLD D., Text typology and machine translation : an overview, in *Translating and the computer*, n°10, London, ASLIB, 1990.

Bureau de Normalisation de l'Aéronautique et de l'Espace, *Spécification ATA n°100*, Traduction française de la mise à jour n°29. Issy les Moulineaux: BNAE, 1992.

AUSTIN J.-L., *Quand dire, c'est faire*, Paris, Seuil, 1970.

BEAUGRANDE R. de, *Text, discourse and process*, Norwood, Ablex Publishing Corporation, 1980.

BÜHLER K., *Sprachtheorie*, Stuttgart, Fischer, 1965.

DAUPHIN E., Étude de corpus : un préalable pour l'adaptation des systèmes de traduction automatique aux besoins des utilisateurs, in *Actes des 5èmes journées scientifiques TA-TAO*, Novembre 1993, Université de Montréal, Montréal, UREF, 1993.

DAUPHIN E.: *Élaboration d'un support de données textuelles bilingues annotées pour l'aide à l'évaluation de la traduction automatique par un utilisateur industriel: l'apport de la pragmatique*, Thèse de Doctorat en linguistique, Université de Paris 7, 1995.

FLICKINGER D. et al., Toward evaluation of NLP systems, *Special session at the 25th annual meeting of the Association for Computational Linguistics*, 1987.

HUMPHREYS L., User-oriented MT evaluation and texte typology, in Falkedal K. ed., *Proceedings of the evaluator's forum*, Les Rasses, Vaud, 21-24 avril 1991, Genève, ISSCO, p.55-64, 1991.

ISABELLE P. et WARWICK-ARMSTRONG S., Les corpus bilingues : une nouvelle ressource pour le traducteur, in BOUILLON P. et CLAS A. ed., *La traductique*, Montréal, Presses de l'Université de Montréal, p.288-306, 1993.

KAY M., RÖSCHEISEN M., Text translation alignment, *COMPUTATIONAL LINGUISTICS*, 19/1, p.121-142, 1993.

KING M. et FALKEDAL K., Using test suites in evaluation of machine translation systems, in H. Karlgren ed., *Proceedings of COLING*, University of Helsinky, p.211-216, 1990.

KITTREDGE R. et LEHRBERGER J., *Sublanguages, studies of language in restricted semantic domains*, Berlin, De Gruyter, 1982.

KOCOUREK R., *La langue française de la technique et de la science*, Wiesbaden, Brandstetter verlag, 1991.

MARCUS P.M. *et al.*, Building a large annotated corpus of English: The Penn Treebank, *COMPUTATIONAL LINGUISTICS*, 19/3, p.313-330, 1993.

MELBY A., La typologie des textes : son importance pour la traduction automatique, in BOUILLON P. et CLAS A. ed., *La traductique*, Montréal, Presses de l'Université de Montréal, p.35-40, 1993.

SAGER J.C., *A practical course in terminology processing*, Amsterdam, John Benjamins, 1990.

SPERBERG-McQUEEN C.M. et BURNARD L. ed., *Guidelines for Electronic Text Encoding and Interchange: TEI P3*, Oxford, ACH, ACL ALLC, 1994.

VANDERVEKEN D., *Les actes du discours*, Liège, Mardaga, 1988.