Raising the Interlingual Ceiling with Multilingual Text Generation*

Keith Vander Linden and Donia Scott
Information Technology Research Institute
University of Brighton, Lewes Road
Brighton BN2 4AT, UK
email: knvl,drs2@itri.bton.ac.uk

1 May 1995

Abstract

In a typical interlingual machine translation (MT) system, the tasks of text planning and content selection are not explicitly performed. Rather, they are assumed to be implicit in the interlingual representation derived in the source language analysis phase. This simplifies the task of target language generation, but greatly restricts the flexibility of the resulting text. Recent MT systems have pushed the level of the interlingua to higher, more semantic/pragmatic levels of representation, and have correspondingly included some text planning capabilities in their generation phases, but the process remains closely bound by the structure and content of the source text. This places a "ceiling" on the extent of the capability of interlingual systems. In this paper we argue that in the general case, the structure and content of "parallel" texts need not be the same. We then propose that multilingual text generation (MLG) has the potential to address this phenomenon by allowing a more flexible variation of text structure and of content selection in a manner unrestricted by a "source" text.

*This work is partially supported by the Engineering and Physical Sciences Research Council (EPSRC) Grant J19221, the Commission of the European Union Grant LRE-62009, and by BC/DAAD ARC Project 293.

1 Introduction

It is well known that languages vary at the level of sentence structure, not just in the range of grammatical constructions provided by the language, but also in the functional impact of those constructions. Interlingual representations in Machine Translation (MT) address this issue; they are intended to rise above the details of grammatical form in the source text, representing the language independent "content" of the source text from which the functionally appropriate target text can be constructed (Hutchins and Somers, 1992). Although most if not all interlingual MT systems build these representations on a sentence-by-sentence basis, it is possible to imagine such an interlingual representation embodying the complete textual, semantic, and pragmatic content of the source text. This content invariance constitutes a theoretical limit of the current MT paradigm, an interlingual "ceiling" of sorts. The assumption is that this as high as one need go to perform adequate translation.

Variation of content selection, which operates above the interlingual ceiling, presents a problem for interlingual approaches. It is common within a single language. For example, one of the typical differences between tutorial guides and reference manuals is the content selected for inclusion in the document. The tutorial tends to focus on a particular course of action while the reference guide attempts to cover all courses of action.

A similar phenomenon has been observed between languages, where it is often the case that two "parallel" texts cannot be seen as "source" and "target", but rather as "two different presentations of the same subject matter" (Hervey and Higgins, 1992, page 178). Hervey and Higgins identify two sources of variation, external to the grammatical construction of the languages, that might lead to content variation in multilingual texts:

- **Cultural expectations** — Different cultures respond to different things. For example, a culture may prefer a certain argumentative structure, or may respond to different imagery.
- **Assumed background knowledge** — Different cultures may have different levels of background knowledge on certain issues and may have access to different facilities. A commonly touted example of this is the terseness of French recipe books compared to their British counterparts (Hervey and Higgins, 1992).

Hervey and Higgins do not pursue either of these points in any detail, but we have observed examples of them in our work on multilingual instructions. Consider the following example of instructions for choosing the appropriate shoe size, taken from a multilingual promotional brochure (Birkenstock, 1995):

(1) (a) (English)

A little extra room when you stand is needed when you walk. If toes have become curled from wearing tight shoes, make sure your sandals are long enough to accommodate them once they have resumed their natural shape

(b) (French)

Beaucoup de personnes choisissent mal la pointure de leurs chaussures, ce qui plus tard aboutira souvent à des douleurs.

'Many people choose the wrong size of their shoes, which often leads to pain later on.'

(c) (German)

Bei der richtigen Größe und Länge machen viele Schuhkäufer Fehler, die sich durch Schmerzen später bemerkbar machen.

'Many shoe sellers/salesmen make mistakes concerning the correct size and length, which makes itself felt later through pain.'

(d) (Italian)

La misura in lunghezza deve essere quella giusta, altrimenti son dolori! Acquistate i vostri sandali Birkenstock tenendo presente che, in posizione eretta, il vostro piede ha bisogno di un paio di millimetri in più, sia dietro al tallone che cavanti alle dita.

'The length measurement must be the right one, otherwise there will be pains! Buy your Birkenstock sandals bearing in mind that, in the wrong position, your feet need an extra pair of millimetres, both behind at the heel and in front at the toes.'

The radical differences in content and style of these texts is presumably due to the variation of expectations between the language cultures. It is not clear, for example, that the English and the French versions have any propositional content in common; they must have been generated with relative independence or at the very least have been translated by someone who had intimate knowledge of the product and considerable freedom in the nature and content of the resulting text. While we may argue the relative merits of the texts in their own cultural context, one must at least concede the possibility of this sort of variation in content between languages.

Multilingual texts may also vary according to the facilities assumed to be relevant in that particular culture. As an example, we find in a multilingual telephone guide (Hyatt Hotels & Resorts, 1988) that the English instruction section contains a list of helpful telephone numbers for a variety of countries (i.e., those countries without a separate section in the manual) along with a full description of how to use each of them, while the French section leaves all of these out and inserts the number for "FRENCH DIRECT, opérateur de FRANCE TÉLÉCOM" and a discussion of how to use it.

These examples exhibit variation of content selection, a variation above the level of the interlingual ceiling. In light of them, we argue that in the general case, the structure and content of "parallel" texts need not be the same. We propose multilingual text generation (MLG) as a paradigm which addresses this phenomenon by allowing a more flexible variation of text structure and of content selection in a manner unrestricted by a source text. In this paper, we will review the notion of an interlingua as it is used in MT. We will then present an architecture for MLG and discuss the knowledge sources it employs, noting how they depart from the standard interlingual mold.

2 The Interlingua in Machine Translation

MT systems vary along two dimensions:

1. The units of analysis and synthesis;
2. The level of the analysis and synthesis of these units.

The potential units of analysis and synthesis cover the strata of language, ranging from words (common in early MT systems) through sentences (common in current MT systems) to full texts. A limitation of most current MT systems is that the units of analysis and synthesis are restricted to the sentence level and below. Any pertinent information that could not in principle be derived from the text (e.g., textual relations between sentences) is not accessible. As Sager points out:

> Since MT is limited to sentence-level transfer, rhetorical and logical structures are not specifically considered by the analysis modules. Any continuity between sentences in a target language MT output is therefore purely coincidental. (1994, page 255)

This is a practical limitation of current MT systems, not of the general MT paradigm. There is no theoretical limitation preventing the MT analysis and synthesis modules from considering textual units and the semantic and pragmatic information they would bring to bear; this is simply an artifact of the present state of the art.

The potential levels of analysis and synthesis range from relatively low-level direct transfer systems, through transfer and interlingual systems to what might be called "content-varying" systems. These levels of analysis can be carried out on any of the above mentioned units of analysis. The first three levels (i.e., direct, transfer, and interlingual) are common in MT systems and are shown in figure 1. As mentioned in the previous section however, the level of analysis may not rise to the level of content selection. This limitation dictates that MT systems can only perform content selection in a limited way (i.e., content may only be removed, but not added). This is a theoretical limitation of the MT paradigm, a limitation as we will see below is addressed in part by the MLG paradigm.

We should note here some more recent MT systems have attempted to put limited reasoning capabilities in their interlingual knowledge representation systems (e.g., Nirenburg et al., 1992). These capabilities could potentially add content to the target texts that was not in the source text, but it is not clear how far this can be pushed. At present the process remains closely bound by the structure and content of the source text.

3 An Architecture for MLG

We are currently engaged in the construction of DRAFTER, an instructions drafting tool intended to support technical authors in the generation of multilingual instructions. We will discuss it here as an exemplar of the MLG paradigm. The idea is that DRAFTER should allow technical authors to specify the procedures that underlie instructions, and then to generate multilingual drafts of the instructions for those procedures. Authors do not input a natural language text, as they would in the MT paradigm, but rather a formal specification of the

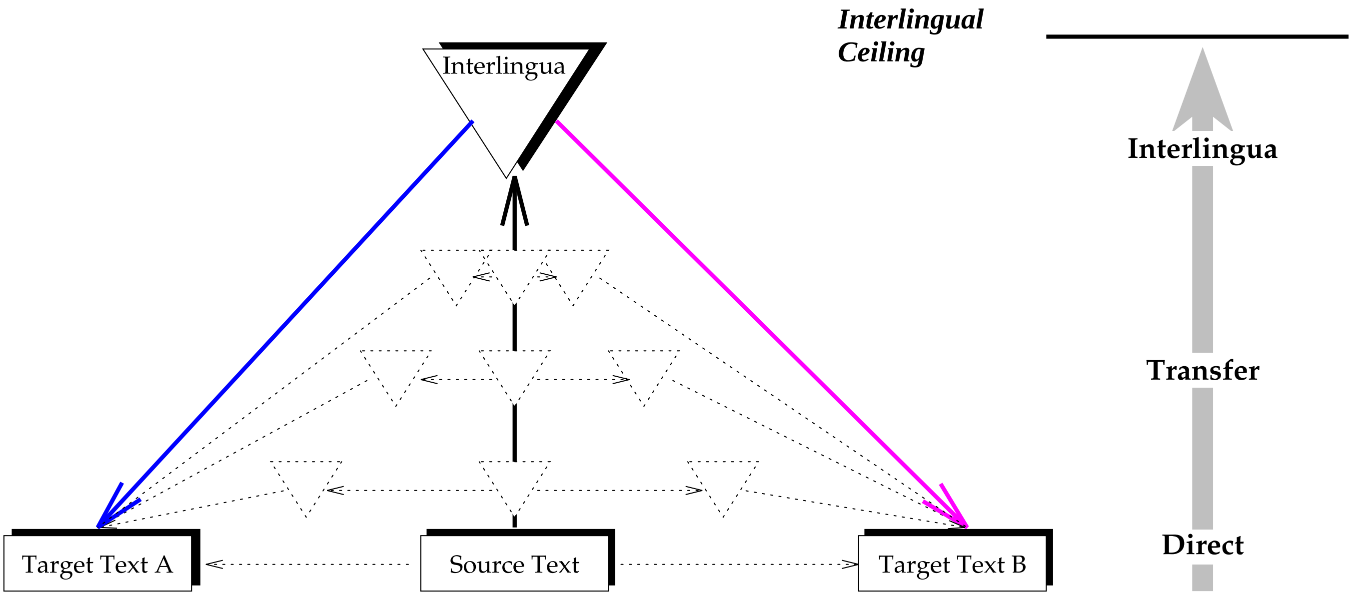

Figure 1: The Interlingua in MT

procedure being documented.[1]

DRAFTER contains three basic components, a Knowledge Base and two processing modules: an Author Interface and a Drafting Tool (see figure 2). The Knowledge Base lies between the two processing components and is the main repository of the semantic knowledge in the system. It represents knowledge at three levels:

- **The Upper Model** — This structure is the root of the Knowledge Base and provides an ontology used by the Tactical Generator to determine how to express concepts in natural language (Bateman, 1990; Henschel, 1993). It provides, among other things, the basic distinction between objects, processes, and qualities, and the definition of grammatical case-roles. It is a static resource provided with the DRAFTER system which may not be edited by the author.

- **The Domain Model** — This structure codes the concepts and relations fundamental to instructional texts. It is a multi-layered structure, maintaining at its highest levels, concepts and relations common to all instructions (e.g., actions, states, objects, and procedural relations) and at its lowest levels, concepts and relations specific to window-based computer interfaces (e.g., windows, clicking, and opening). The inventory of concepts and relations contained here have been derived from a study of a multilingual corpus of software instruction manuals. Currently, this is also a static resource which is provided with the DRAFTER system.

- **The Proc./Desc. Instances** — This structure represents the procedural and descriptive information that underlies the instructions for using a particular device or tool. The author builds this structure interactively by instantiating relevant concepts provided in the Upper and Domain Models, and relating them as appropriate for the procedure being documented.

The two processing modules which interact with this Knowledge Base are as follows:

- **The Author Interface** — This interface allows technical authors to specify the procedures involved in using the device in question (i.e., the Procedural/Descriptive Instances). It also allows them to control the Drafting Tool.

[1] DRAFTER's current domain of application is instructions for window-based computer interfaces in French and English.

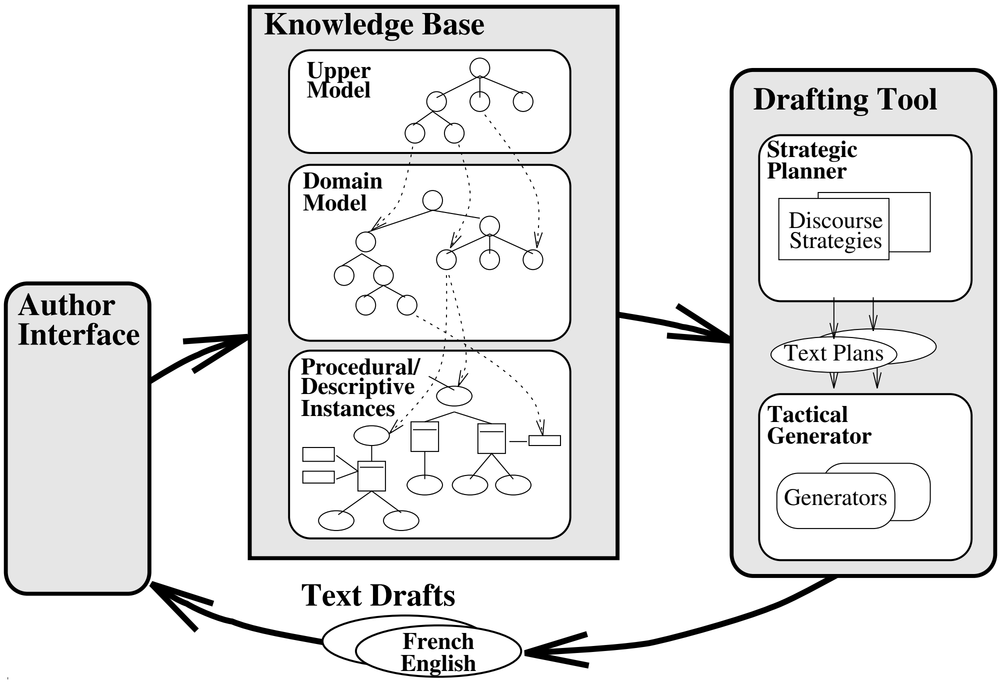

Figure 2: DRAFTER Data Flow Diagram

- **The Drafting Tool** — This tool contains the two standard language generation tools, namely a *strategic* component for planning the content and the text structure, and a *tactical* component for planning sentence structure. Given the information specified in the Knowledge Base, this tool produces draft instructions in French and English.

The basic scenario for using the DRAFTER system is that the author should identify and specify the procedures being documented (see Paris et al., 1995 for more details on this). The resulting representation, as with an MT interlingual representation, is treated as language independent. The author then initiates the Drafting Tool. This tool performs content selection, text planning, and sentence realization, passing the resulting multilingual drafts back to the author who then has the opportunity to edit the raw text or to modify the procedural/descriptive instances and re-draft. All aspects of the drafting process are driven by potentially language-specific resources (i.e., the discourse strategies and grammatical resources). We further intend to allow the author to tailor the drafting process by setting a collection of pragmatic parameters which will control the resulting form of the instructions.

4 Where is the Interlingua?

In the MLG paradigm just described, the most obvious correlate to the MT interlingua is the procedural/descriptive instances that the author builds. There are really no other legitimate candidates; all the data structures within the Drafting Tool are potentially language specific. The procedural and descriptive instances are treated as a language independent source from which to generate the target texts. They are not, however, an interlingua in the strictest sense of the term because they do not dictate content selection. The author has the potential to enter a variety of information that is perhaps relevant only to certain languages. Correspondingly, the Drafting Tool has the potential to select a sub-set of the content of the full Knowledge Base depending upon language specific discourse strategies. Here, the MLG paradigm rises above the interlingual ceiling of textual and sentential structure, into the realm of content selection (see figure 3).

As an example, consider the multilingual tele-

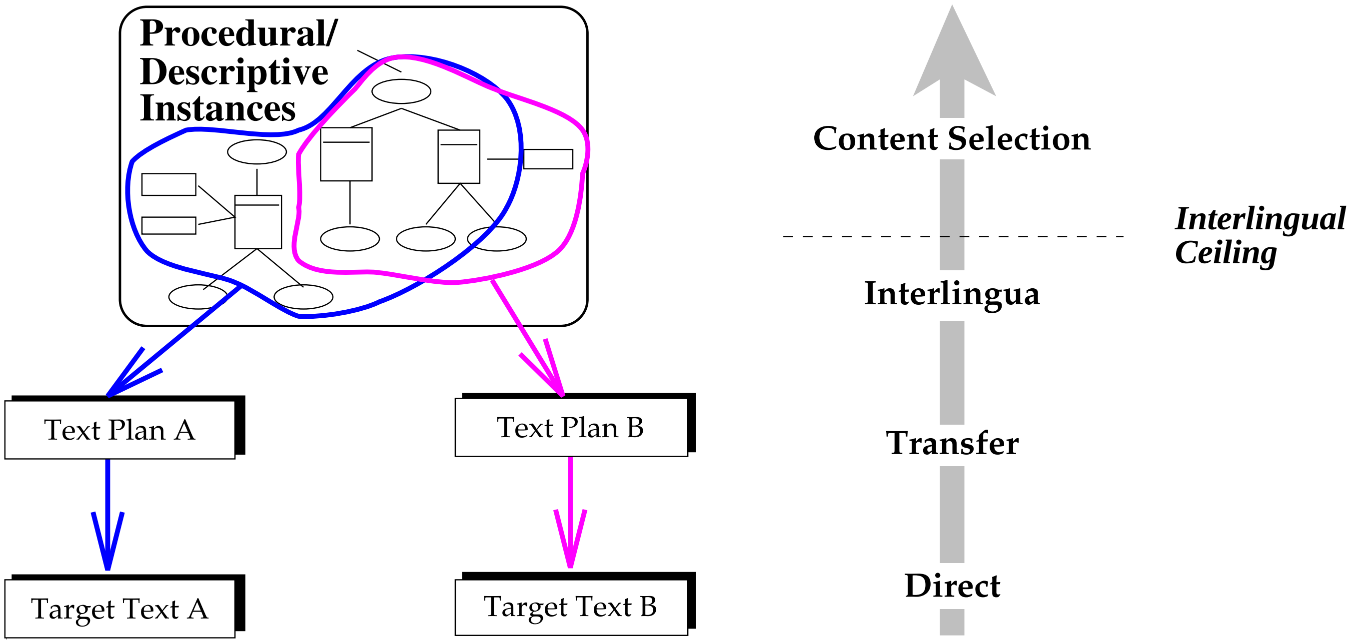

Figure 3: The Interlingua in MLG

phone guide discussed above where the authors included information in the full guide for 11 different country direct services, but did not feel it necessary to include a translation of all of them in each language portion of the guide. Rather, they selected the service or services most relevant to that language (e.g., FRANCE Direct for the French language section). In a MLG scenario, the author would simply include information on all of them in the Knowledge Base and then either ear-mark portions for inclusion in certain language versions, or perhaps, if the discourse strategies would support it, allow the Strategic Planner to distribute them as required.

Similarly with the Birkenstock guide, the author could specify the range of information relevant to the process in question and allow the discourse strategies to independently determine the content and structure of the instructions.

5 Discussion

Texts intended to elicit the same or similar effects in different language/cultures can vary not only in sentence structure, but also in text structure and content selection. The former has been the subject of the majority of the work in the MT paradigm, while the latter has received relatively little attention. We believe that the automated production high-quality multilingual texts requires attention to all these issues and have thus been studying multilingual instructions, particularly with respect to their content selection and text structure.

The majority of the translated texts in our corpus have relatively parallel text structure and content. It is not clear, however, whether these parallel structures were actually appropriate, or whether they arose because the texts were translated in a parallel fashion. We suspect the latter because of the growing number of examples of variation we are finding, particularly in independently generated multilingual texts. Certainly in practical terms it would be simpler to assume that parallel structures were appropriate and rely on post-editing to deal with the anomalous non-parallel cases. Unfortunately, post-editing gives rise to its own problems related to effort and reuse.

There are at least two approaches for dealing with this variation in structure and content. The first is to augment the MT paradigm by giving the translator the ability to modify, and in particular to augment, the knowledge in the interlingual representation. To our knowledge, no one has attempted to do this, presumably because of the difficulties in making MT interlingual representations understandable and modifiable by translators. We have pursued a second approach, that of MLG. In this alternate paradigm, the author enters a range of content into a knowledge base from which the multilingual generation system picks and chooses as appropriate for each particular language. Although MLG technology is in the early stages of its development, we argue that in certain domains it provides a useful alternative to MT, one which allows a more flexible variation of text structure and of content selection in a manner unrestricted by a "source" text.

References

Bateman, J. A. (1990). Upper modeling: Organizing knowledge for natural language processing. In McKeown, K. R., Moore, J. D., and Nirenburg, S., editors, *Proceedings of the Fifth International Workshop on Natural Language Generation,* June 3–6, Dawson, PA.

Birkenstock (1995). Besser gehen. Birkenstock - Orthopädie GmbH, Rhein, Germany.

Henschel, R. (1993). Merging the English and the German upper model. Technical report, GMD/Institut für Integrierte Publikations- und Informationssysteme, Darmstadt, Germany.

Hervey, S. and Higgins, I. (1992). *Thinking Translation*. Routledge, London.

Hutchins, W. J. and Somers, H. L. (1992). *An Introduction to Machine Translation*. Academic Press, London.

Hyatt Hotels & Resorts (1988). International telephone guide. AT&T.

Nirenburg, S., Carbonell, J., Tomita, M., and Goodman, K. (1992). *Machine Translation: A Knowledge-Based Approach*. Morgan Kaufmann, San Mateo, CA.

Paris, C., Vander Linden, K., Fischer, M., Hartley, A., Pemberton, L., Power, R., and Scott, D. (1995). A support tool for writing multilingual instructions. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence,* August 20–25, Montréal, Canada.

Sager, J. C. (1994). *Language Engineering and Translation: Consequences of automation*, volume 1, Benjamins Translation Library. John Benjamins, Amsterdam.