Multi-language Machine Translation through Interactive Document Normalization

Aurélien Max
Groupe d'Etude pour la Traduction Automatique (GETA)
Xerox Research Centre Europe (XRCE)
Grenoble, France
aurelien.max@imag.fr

Abstract

Document normalization is an interactive process that transforms raw legacy documents into semantically well-formed and linguistically controlled documents with the same communicative intention content. A paradigm for content analysis has been implemented to select candidate semantic representations of the communicative content of an input document. This implementation reuses the formal content specification of a multilingual controlled authoring system. As a consequence, a candidate semantic representation can not only be associated with a text in the language of the input document, but also in all the languages supported by the system. This paper presents how multilingual versions of an input legacy document can be obtained interactively with a proposed implementation, and discusses the advantages and limitations of this kind of *normalizing translation.*

1 Introduction

Translating unrestricted text by machine is a problem that has been involving a lot of research for the past decades, but is still far from solved (Cole et al., 1996). This task arises so many problems in computational linguistics, most of them only partially solved, that a lot of research is still to be carried out before one can ask a personal computer to translate accurately an arbitrary piece of text from one language to another. The performance bottleneck due to the lack of linguistic and knowledge resources has led the builders of practical translation systems to constrain the input to controlled languages, and/or to have recourse to human expertise on the source language and on the discourse domain (e.g. (Boitet and Blanchon, 1996; Baker et al., 1994)). Unsurprisingly, the most successful systems to date operate on text in very limited domains, exemplified by the weather forecast translation from English to French of the METEO system.

There exist many situations where documents belonging to a constrained domain have to be translated in several languages, as is the case of official documents in multilingual communities or product descriptions for international companies. In these situations one have at least the following expectations:

- **high-quality translation,** which implies that it be accurate and not necessarily literal
- outputs in possibly **many target languages**
- **consistency** across documents of the same class (e.g. drug leaflets, experiment reports), so that concepts are always expressed in the same unambiguous manner and the texts produced can be regarded as **gold standards** for the meaning they convey

Different methods that do not impose constraints on the input text have been proposed to achieve high-quality translation. Interaction with a user can be used to disambiguate the input text, and could be preferred to to post-editing as this has to be done only once for all languages, thus reducing the time and efforts needed. Interlingual representations (Hutchins and Somers, 1992) are well-adapted to support the production of the target text in several languages, and they can also be effectively used to check the semantic coherence and well-formedness of a document. Reusing previous translations, as proposed in the different flavours of Example-based Machine Translation (Somers, 1999), is an interesting alternative to purely rule-based approaches and allows the selection of non-literal high-quality translation candidates.

This paper starts with a short presentation of an authoring system that allows the creation of

multilingual documents with all the above properties. *Document normalization,* which is described next, stemmed from the question of whether such an authoring system could be vised in a reversed mode to analyze existing documents from the class of documents supported by it. After providing a motivating example, we will briefly introduce *fuzzy inverted generation,* a paradigm we proposed to normalize documents reusing the formalism of the above mentioned authoring system, and describe a document normalization system. We will then attempt to define how *normalizing translation* can be achieved through document normalization, and we will discuss the advantages and limitations of such an approach.

2 Controlled Document Authoring

Controlled Document Authoring is an active field of research comprising approaches such as the *What You See Is What You Meant* (WYSIWYM) paradigm (Power and Scott, 1998) and *Multilingual Document Authoring* (MDA) (Dymetman et al., 2000). The systems allow authors to specify document content representations interactively in their own language, and then produce versions in several languages using parallel resources.

In MDA, a system developed at XRCE, the author of a document has to select valid semantic choices in active fields interspersed with the evolving text of the document in her language until the document is complete (see figure 1). The system can at any time produce current versions of the documents from the content representation in all the languages it supports. The documents thus obtained are of high-quality, and are not necessarily literal translations but rather adaptations to a given language.[1] In fact, the linguistic structures of two documents can be completely different in two different languages, and communicative intentions can be conveyed in quite different ways. Moreover, since the generator of an MDA system is deterministic, the texts produced will be consistent across documents.

The specification of well-formed document content representations in MDA is recursively described in a grammar formalism that is a variant of Definite Clause Grammars (Pereira and Warren, 1980). Text strings can appear in right-hand sides of rules, which allows text realizations to be associated to content representations, and thus provides a close coupling between semantic modelling and generation. Figure 2 shows an abstract typed tree in the MDA formalism and its realizations as English and French sentences.[2] Non-terminals are typed semantic elements whose type appears after the two colons. Dependencies can be enforced through the use of shared variables between semantic elements. The granularity of text fragments in rules is not necessarily a fine-grained predicate-argument structure of sentences commonly used in NLG, so this is an intermediate level between full NLG and templates (Reiter, 1995). This approach proved to be adequate for classes of documents where the productivity of certain choices could be rendered as entire text spans, as is the case for example of warning sections in drug leaflets (Bruii et al., 2000).

3 Document normalization

3.1 A Motivating example

The pharmaceutical domain produces yearly publications which are compendiums of documents initially produced by pharmaceutical companies which are presented in a consistent way (e.g. (ABPI, 1996; OVP Editions du VIDAL, 1998)). Several kinds of variations were observed in a corpus study we conducted on a corpus of 50 patient pharmaceutical leaflets for pain relievers from different drug vendors (Max, 2002). First, the structures of the leaflets could vary considerably, as well as the locations where certain communicative goals were expressed. (Paiva, 2000) showed the presence of significant stylistic variation in a corpus of 342 patient leaflets. Our study also revealed that similar communicative intentions could be expressed in a variety of ways conveying more or less subtle semantic distinctions. Seeing the content of such documents as goal-driven communication, a given utterance can be seen as an attempt to satisfy some communicative goal on the part of the writer of the document. We argue that for documents of the importance of pharmaceutical leaflets consistency of expression and of information presentation can be beneficial to the reader by allowing a clear and unambiguous understanding of the communicative goals contained in different documents. It can indeed sometimes be confusing for a reader to find various ways to express the same communicative intentions, as in the following examples:

- *This product should not be taken for more than*

[1] Different parts of the document can thus be easily localized: for example, disclaimers and contact information can be adapted to the targeted community.

[2] This example and its specification are inspired from the Nespole! project, a speech-to-speech translation project.

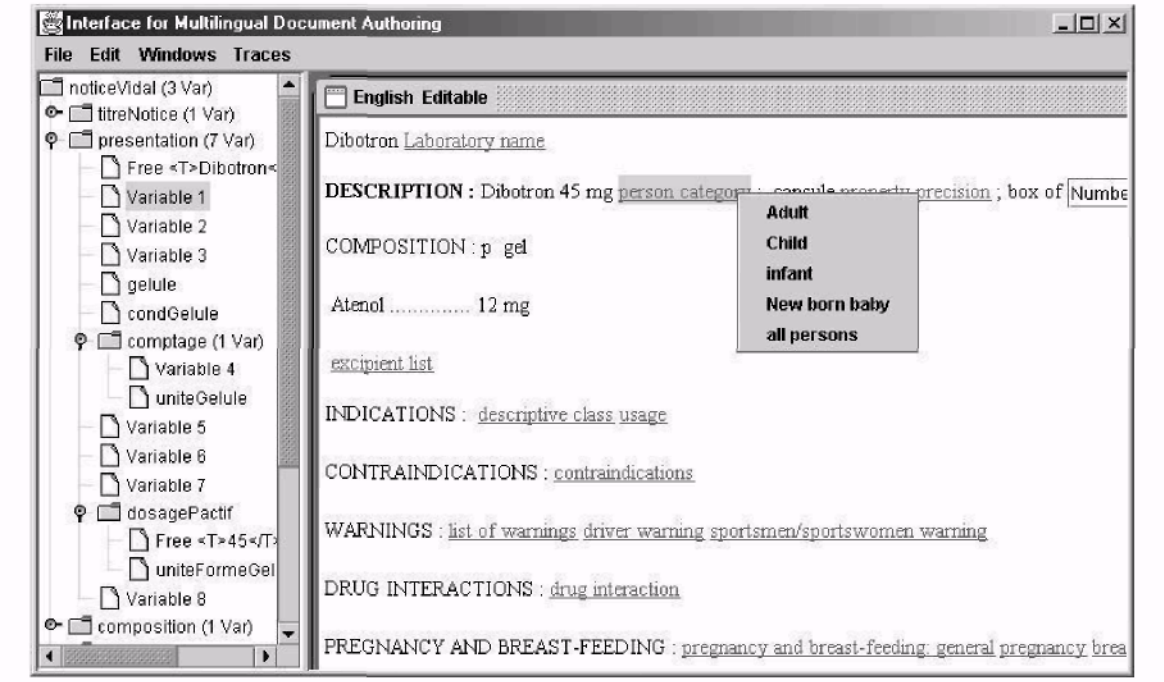


Figure 1: View of the MDA system during the authoring of a patient information leaflet

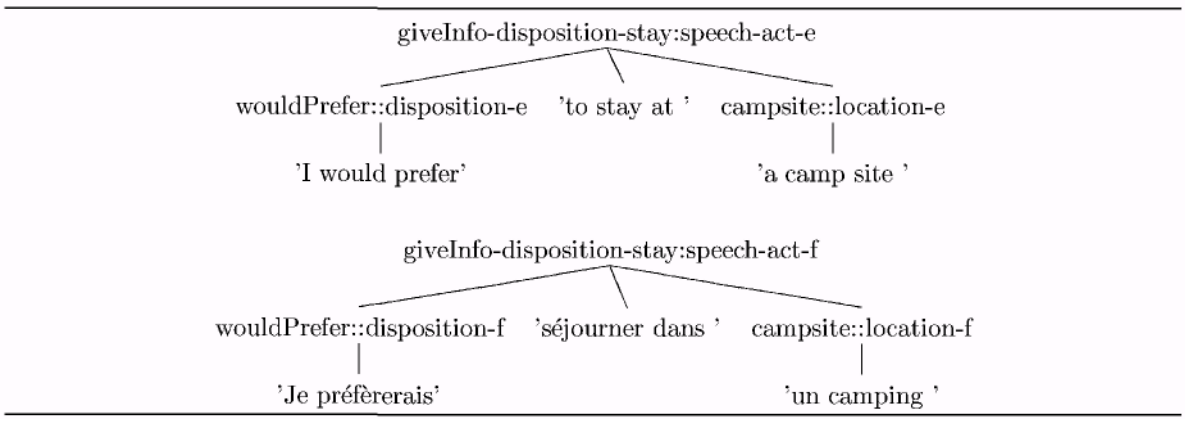


Figure 2: Abstract typed trees in English and French for the sentence / *would prefer to stay at a camp*

14 days without first consulting a health professional.

- *If pain persists after 14 days, consult your doctor before taking any more of this product.*
- *If symptoms persist for 2 weeks, stop using this product and see a physician.*

Document normalization can be achieved by analyzing a legacy document into a semantically possible content representation, and producing a normalized version from that content representation. This normalized version expresses *predefined content,* which is conveyed in the input document, in a structurally and linguistically controlled way. Predefined content reveals *communicative goals,* which should typically be described by an expert of the discourse domain. Control on the production of text from some content representation allows to produce messages that can be seen as some sort of 'gold standard' for the communicative goal that are conveyed and that can be augmented to be made *self-explaining* (Boitet, 1996), and to obtain consistent document structures as well as to impose terminological and stylistic guidelines.

3.2 Fuzzy inverted generation

For the purpose of document normalization we would like to match texts that do not carry significant communicative differences in a given class of documents but may be of quite different surface forms. Therefore, we proposed to concentrate more on what counts as a well-formed document semantic representation rather than on surface properties of text, as the space of possible content representations is vastly more restricted than the space of possible texts.

Bridging the gap between deep content and surface text can be done by using the textual predictions made by the generator of an MDA system from well-formed content representations to match an input document. Indeed, an MDA system can be used as a formal device for enumerating well-formed document representations in a constrained domain and associating textual representations with them. If we can compute a relevant measure of semantic similarity between the text produced for any document content representation and the text of a legacy document, we could possibly consider the representations with the best similarity score as those best corresponding to the legacy document under analysis. Since this kind of analysis uses predictions made by a natural language generator, we named it *inverted generation* (Max and Dymetman, 2002) (see fig. 3). We also qualified it *fuzzy,* because as a generator will seriously undergenerate with respect to all the texts that could be normalized to the same communicative intention, the matching procedure has to be performed at a more abstract level than on raw text to evaluate commonality of communicative content. Considering the types of documents that could be analyzed using this paradigm, it seemed relevant to expand the generative power of the system, so that different texts could be associated with the same content representations to increase the robustness of the analysis. Although this non-determinism proves beneficial for inverted generation, we implemented it in such a way that the generation process would still be done deterministically.

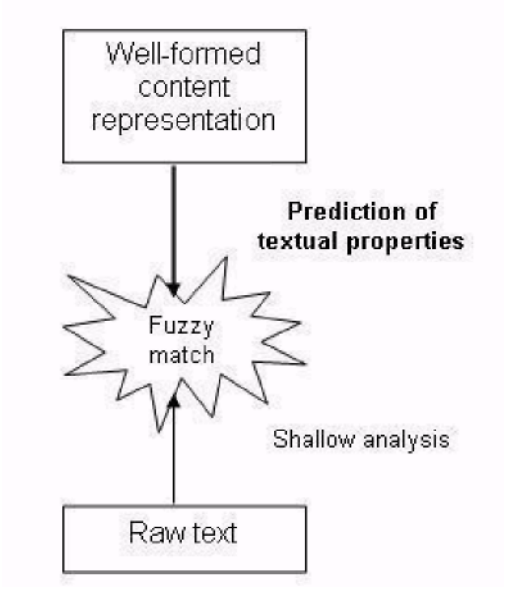


Figure 3: Deep content analysis through fuzzy inverted generation

To normalize an input document, we would like to find the *virtual document*[3] that is most similar to the input document in terms of the communicative content it conveys. The space of virtual documents for a given class of documents being potentially huge, we proposed an admissible heuristic search procedure (Nilsson, 1998), so that the candidate structures are returned in an order of decreasing similarity with the input text. The evaluation function it uses is an optimistic measure of similarity that corresponds to a weighted intersection between the *lexical profile* of the input

[3] We call *virtual document* a document that can be predicted by the authoring system but does not exist *a priori.*

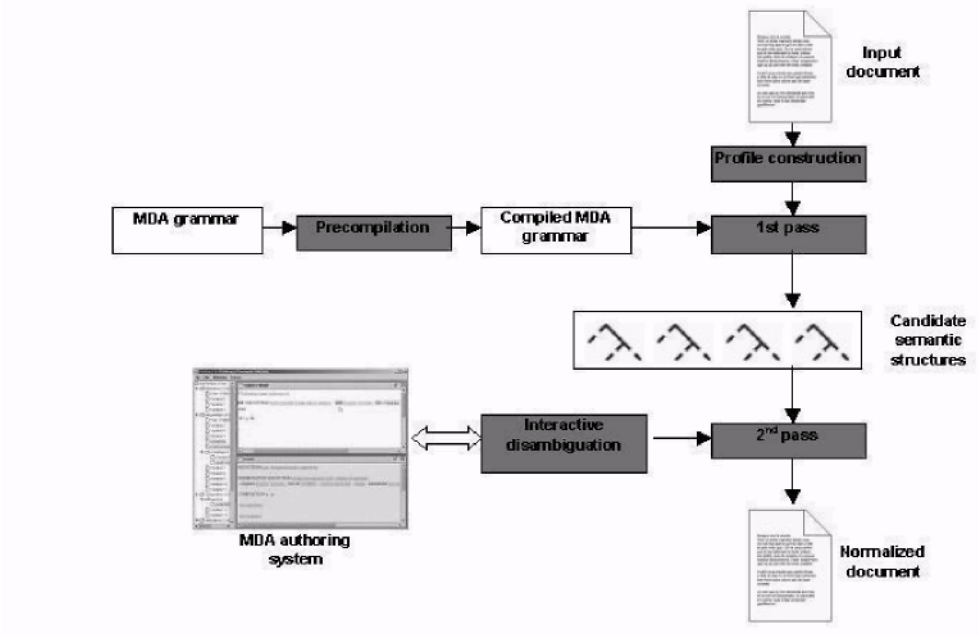


Figure 4: Architecture of the document normalization system

document and that computed for a partial content representation.[4] The lexical profile for a text fragment is defined as a vector of informative synsets5 associated with their number of occurrences, and the lexical profile for an MDA semantic type gives the maximum number of occurrences of any given synset that could be attained by performing any derivation from that type.

3.3 Document normalization system

Figure 4 shows an overview of the document normalization system that we have started to develop. An MDA grammar is first compiled off-line to associate profiles with all its semantic types by percolating profiles in the grammar from the terminals up to the root type. This compiled version of the grammar is used in conjunction with the profile computed for the input document in a first pass analysis. The aim of this first pass analysis implementing fuzzy inverted generation is to isolate a limited set of candidate content representations. A second pass analysis is then applied on those candidates, which are now actual texts associated with their content representation, using more fine-grained linguistic analysis, in conjunction with interactive disambiguation when needed.[6]

4 Normalizing translation

Using the resources of a multilingual authoring system to analyze a legacy document offers a natural possibility: once the semantic content representation is obtained through document normalization, the generative capability of the authoring system can be reused to produce the documents corresponding to that representation in all the supported languages (see figure 5). Normalizing translation uses the same resources for both analysis and generation, and shares some properties with a pivot approach. A significant difference with previous approaches to translation using *reversible grammars* is that fuzzy matching is used. As this approach to machine translation relics on the matching with existing texts (those that can be produced by the generator of the authoring system), it shares some properties with Example-based Machine Translation (Somers, 1999), with the specificity that matched text fragments correspond first to semantic types in the MDA formalism and then eventually to their appropriate trans-

[4] More details on how fuzzy inverted generation can be implemented in MDA can be found in (Max, 2002).
[5] Synsets from WordNet, or ideally from a specialized thesaurus, have been preferred to lemma in order to account for lexico-semantic variation (Gonzalo et al., 1998).

[6] Typically, interactive disambiguation will allow an expert to prefer one of several ambiguous candidates on the basis of the legacy document.

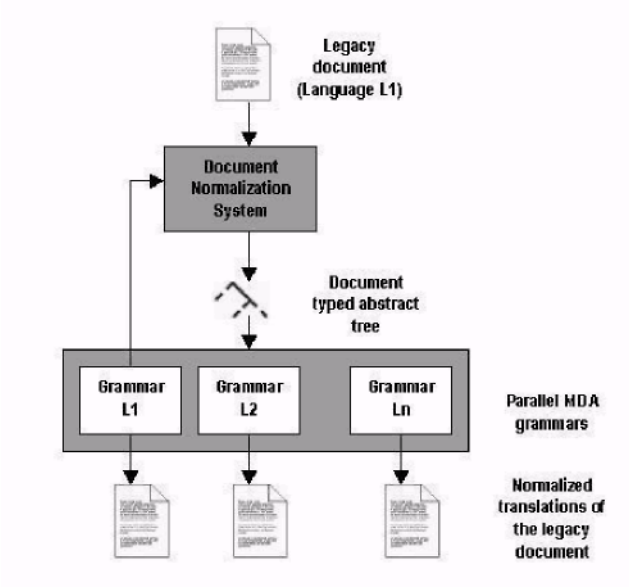


Figure 5: Normalizing translation using MDA grammars

lations in other languages. Under the assumption that the authoring system produces high-quality documents in all the languages it supports, then evaluating normalizing translation can be limited to evaluating the performance of document normalization. Forthcoming publications will attempt to address this issue.

A simple example of normalizing translation is given on figure 6. The discourse domain is assumed to be that of travel information, where some semantic distinctions are considered uninformativc. In this case the English speaker wishes to mention his *family's first choice,* but this is lost in the translation. The utterance is best matched with **wouldPrefer-3**, a possible English realization for the semantic type **disposition**, in the context of **giveInfo-disposition-stay**. It is then normalized to the deterministic choice of the English generator, **wouldPrefer**. The structure to which it belongs, which is of type **speech-act**, can then be rendered in English as *I would prefer to stay at a camp site.* Using parallel MDA grammars for French and Spanish allows to obtain the corresponding abstract trees from which the French and Spanish versions of the normalized sentence can be obtained.

5 Discussion

The proposed approach to translation has important limitations: first, only documents in constrained domains which can be modeled with the MDA formalism can be dealt with. This excludes arbitrary pieces of text, and requires an initial development of the grammatical resources and its transposition to as many languages as the system should support. However, this would allow to reuse the document modeling for authoring new documents from scratch, which could modify in a beneficial way the documentation practices of technical writers. As opposed to 'traditional' machine translation, normalizing translation can only translate those elements of a text that fit in well-formed document content representations. Consequently, elements that are not modeled in the grammar used for analysis will not appear in the normalized version of the document and its translations, which makes normalizing translation performing a kind of content selection. Another delicate aspect of this approach is that if the normalization goes wrong, even though an expert could control and validate the whole process[7], then the multilingual versions of the resulting document will not be accurate translations of the input document.

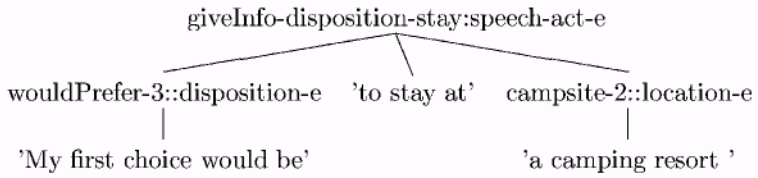
Despite the limitations given above, we think that this approach proposes enough advantages to be a viable solution for some well-defined contexts. First and foremost, if normalization goes well, so will translation, provided the parallel grammars of the authoring system are correct. The fuzzy inverted generation we have proposed has the inherent property of only producing candidates that are semantically well-formed and coherent, thus providing a means to the expert to correct or to reject an ill-formed legacy document. Validated documents are richer than usual textual documents since they are associated with their semantic description, which can for example be used to index the documents in a knowledge base for subsequent retrieval.

On the architectural side, the same resource is used for both analysis and generation, thus reducing considerably development time. Moreover, the fuzzy approach and the non-determinism of the inverted generation makes it possible to match a large range of inputs that could be more difficult to recognize using more traditional approaches to content analysis, such as syntactic parsing followed by semantic composition (Allen, 1995). Provided the necessary resources are available, notably a lemmatizer, a lexico-semantic database such as WordNet, and a human expert fluent in the appropriate language, any grammar of the authoring system could be used for analysis, therefore pos-

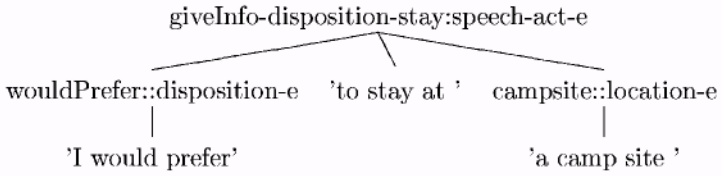
[7] The normalized document in the original language can be used by the expert to validate the normalization process, similarly to *feedback texts* in WYSIWYM.

Input text: *Staying at a camping resort is always my family's first choice.*

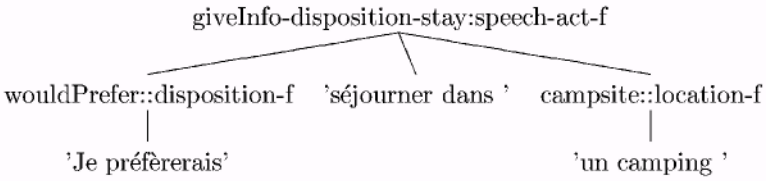
Best matching English abstract tree:



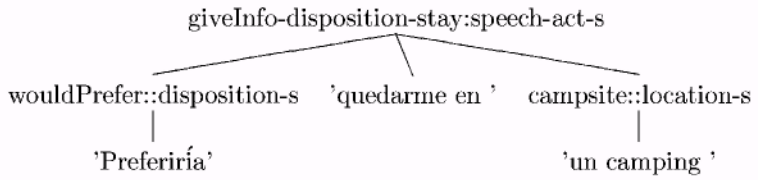
Corresponding normalized English abstract tree:



Corresponding French abstract tree:



Corresponding Spanish abstract tree:



Input text normalized translations:

- English: I *would prefer to stay at a camp site.*
- French: Je *préfèrerais séjourner dans un camping.*
- Spanish: *Preferirla quedarme en un camping.*

Figure 6: Example of normalizing translation for the English sentence *Staying at a camping resort is always my family's first choice* in the context of travel information

sibly allowing an N-to-N normalizing translation architecture. Finally, if the system has some supervised learning ability, for example by augmenting its generative power with examples validated by the expert, then it could be expected to perform better as more normalizations are done, as is the case with translation memories.

Acknowledgements The author wishes to thank Marc Dymetman and Christian Boitet for their supervision of his PhD work. This work is funded by a grant from ANRT.

References

ABPI. 1996. *ABPI Compendium of Patient Information Leaflets.* Datapharm Publications.

James Allen. 1995. *Natural Language Understanding.* Benjamin/Cummings Publishing. Redwood City, 2nd edition.

Kathryn L. Baker, Alexander M. Franz, Pamela W. Jordan, Teruko Mitamura, and Eric H. Nyberg. 1994. Coping with Ambiguity in a Large-Scale Machine Translation System. In *Proceedings of COLING-94, Kyoto, Japan.*

Christian Boitet and Hervé Blanchon. 1996. Multilingual Dialogue-Based MT for monolingual authors: the LIDIA project and a first mockup. *Machine Translation,* 9:99-132.

Christian Boitet. 1996. Dialogue-based machine translation for monolinguals and future self-explaining documents. In *Proceedings of MIDDIM-96, Le Col de Porte. France.*

Caroline Bruii, Marc Dymetman, and Veronika Lux. 2000. Document Structure and Multilingual Authoring. In *Proceedings of INLG 2000, Mitzpe Ramon. Israel.*

Ronald A. Cole, Joseph Mariani, Hans Uszkoreit, Annie Zaenen, and Victor Zue, editors. 1996. *Survey of the State of the Art in Human Language Technology.* Cambridge University Press.

Marc Dymetman, Veronika Lux, and Aarne Ranta. 2000. XML and Multilingual Document Authoring: Convergent Trends. In *Proceedings of COLING 2000, Saarbrucken, Germany.*

Julio Gonzalo, Felisa Verdejo, Irina Chugur, and Juan Cigarràn. 1998. Indexing with WordNet Synsets Can Improve Text Retrieval. In *Proceedings of the COLING/A CL Workshop on the Usage of WordNet in Natural Language Processing Systems.*

W.J. Hutchins and Harold Somers. 1992. *An Introduction to Machine Translation.* Academic Press. London.

Aurélien Max and Marc Dymetman. 2002. Document Content Analysis through Inverted Generation. In *Proceedings of the workshop on Using (and Acquiring) Linguistic (and World) Knowledge for Information Access of the AAAI Spring Symposium Series, Stanford University, USA.*

Aurélien Max. 2002. Normalisation de Documents par Analyse du Contenu à l'Aide d'un Modèle Sémantique et d'un Générateur. In *Proceedings of TALN-RECITAL 2002, Nancy, France.*

Nils J. Nilsson. 1998. *Artificial Intelligence: a New Synthesis.* Morgan Kaufmann, San Francisco.

OVP Editions du VIDAL, editor. 1998. *Le VIDAL de la famille.* Hachette, Paris.

Daniel S. Paiva. 2000. Investing Style in a Corpus of Pharmaceutical Leaflets: Result of a Factor Analysis. In *Proceedings of the ACL Student Research Workshop. Hong Kong.*

Fernando Pereira and David Warren. 1980. Definite Clauses for Language Analysis. *Artificial Intelligence,* 13.

Richard Power and Donia Scott. 1998. Multilingual Authoring using Feedback Texts. In *Proceedings of COLING/ACL-98, Montreal, Canada.*

Ehud Reiter. 1995. NLG Vs Templates. In *Proceedings of ENLGW-95, Leiden, The Netherlands.*

Harold Somers. 1999. Review Article: Example-based Machine Translation. *Machine Translation,* 14:113-157.