

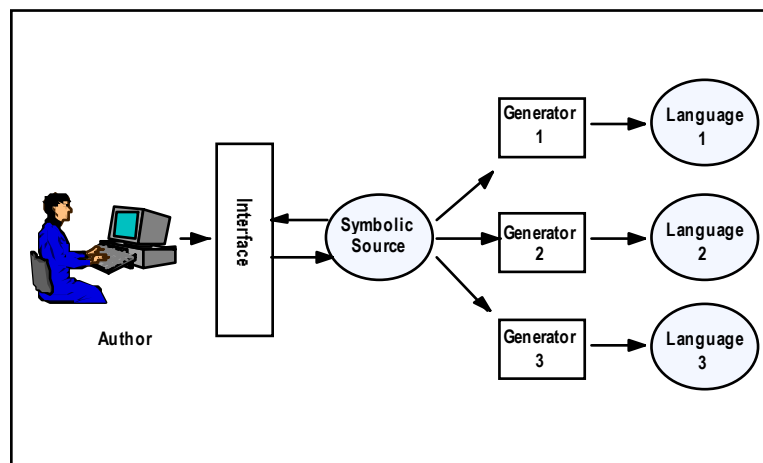
Localisation, Knowledge-base Editing and Multilingual Natural Language Generation

*Donia Scott, Information Technology Research Institute, University of Brighton, Brighton, U.K.
Email: donia.scott@itri.bton.ac.uk*

One of the core activities underpinning the information society is document management — the ability to create, maintain and update documents, and especially sets of related documents, in a co-ordinated way. Multilingual document management, where documents need to be maintained in several languages, will similarly be the foundation of an effective multilingual information society. Current approaches to multilingual document management employ manual or automatic translation (or a combination of both) from master monolingual sources. Recent developments in Natural Language Generation (NLG) technology suggest an alternative approach in which the master source is language-neutral, and documents in all languages are generated independently and automatically. This eliminates ‘source language bias’, makes subsequent updates to a document easier, faster and probably cheaper, and facilitates multilingual maintenance of the document base.

Technological support for monolingual document management is now quite well-established. Template and stylesheet facilities are common on many word processors, access and version control supports co-ordinated document development, macros and conditional constructions can be used to support different variants of the same basic document. But the multilingual situation brings with it additional problems. The most fundamental is how to maintain over time versions of the same document in different languages. The techniques for management of variants of a document in the same language are in general not powerful enough to support the relationship between the same document in different languages, even when they are quite direct translations of each other. To make matters worse, in general direct translations are not what is required: different languages and cultures have their own ways of expressing the same ideas and the most effective document is one which conforms in style as well as language to the reader’s expectations. Supporting this requires techniques far beyond the abilities of most current document management systems.

At present, the principal way of producing versions of a single document in several languages is through translation: the document is initially written in one language and then translated into other desired languages. Manual translation is big business, but it is costly (good translators are relatively rare and therefore expensive) and always



Symbolic authoring architecture for multilingual document production

under time pressure. Automatic translation is potentially quicker and cheaper, but current systems still lack the quality, coverage and adaptability required to deliver final copy of important public documents.

In addition, translation-based multilingual document management tends to favour the source language. The appropriate style, register and distance from the reader for a particular document type varies from language to language, as does the linguistic realisation of these features. For example, in instructional texts, French is more likely to use indirect constructions than English, and also more likely to express them using impersonal pronouns rather than passive constructions (Paris and Scott 1994). Expert translators (with no time constraints!) can accommodate these differences, but more often echoes of the source language detract from the quality of the translated document.

The alternative approach that we and our colleagues have been exploring uses a technique called **Symbolic Authoring** to generate language-neutral symbolic representations of the content of a document, from which documents in each target language are generated automatically, using NLG technology. NLG has been developing steadily in recent years, and a number of commercial or near commercial systems now exist.¹ Many of these systems take their input from some external data

¹ For example, **AlethGen** (Coch 1996), **CORECT** (Levine et al 1996), **DRAFTER** (Paris et al 1995), **EXCLASS** (Caldwell & Korelsky 1994), **FOG** (Goldberg et al 1994), **GhostWriter** (Marchant et al 1996), **GIST** (Power et al 1995), **IDAS** (Reiter et al 1995), **ILEX** (Knott et al 1996), **LFS** (Iordanskaja et al 1992), **ModelExplainer** (Lavoie et al 1996), **PlanDoc** (McKeown et al 1994) and **PostGraphe** (Fasciano & Lapalme 1996).

source. The idea of Symbolic Authoring is simply to allow the user to specify the generator input directly.

In essence, a Symbolic Authoring system comprises a natural language generator coupled to an interface that supports the manual creation of the generator's input (that is, the authoring of the symbolic (conceptual) content of the document). Such a system becomes interesting if we add additional generators for other languages — see figure 1. Now a single (symbolic) authoring process supports multilingual variants of a document directly: one update to the document is reflected in all languages simultaneously. Furthermore, each generator can be tuned to its own language and cultural settings, choosing its own most appropriate realisation strategy independently of the others.

As well as the NLG technology, it is clear that the other key requirement of a Symbolic Authoring system is an effective user interface. The 'symbolic content' required by an NLG system is typically a LOOM-like knowledge base (MacGregor, 1988), and the user interface must enable the author to construct such a knowledge base. This is a significant problem, which different systems have addressed in different ways. Our own most recent work uses a technique called **WYSIWYM** ('what you see is what you meant' (Power et al 1997)) to present the knowledge base to the author as text (using the same NLG technology as the authoring component itself). Early experiments suggest this could be a very effective and general solution to the input interface problem.

Symbolic Authoring allows the simultaneous management of a document in several languages, through the use of a language-neutral content representation. These 'symbolic sources' can themselves be managed as documents (sharing structure, using macros and templates etc.). The symbolic nature of the information also allows for more powerful authoring support such as cross-referencing, consistency checking and stylistic control. Additionally, because the source documents are language-neutral, they can be maintained equally well by authors of any nationality (using appropriately localised interface tools — and with WYSIWYM, this localisation comes for free). The authoring language is purely a feature of the interface, not the underlying document.

How much of what we have described is feasible right now? Current NLG works best with fairly short documents in well-understood genres (such as instructional texts). In addition, existing input representations tend to be quite application-specific. Nevertheless, systems such as DRAFTER, GIST and Ghostwriter show that useful applications can be created within those constraints. Effective symbolic authoring user interfaces exist, and there are exciting developments in this area, such as WYSIWYM. Full integration into a real document management system also remains an outstanding task, but a primarily technical one. In summary, most of the key pieces of this potential cornerstone of MLIS are there, just waiting to be put together.

References

- Caldwell, D. and Korelsky, T. (1994), *Bilingual generation of job descriptions from quasi-conceptual forms*, in Proceedings of the Fourth Conference on Applied Natural Language Processing.
- Coch, J. (1996), *Evaluating and comparing three text production techniques*, in Proceedings of the Sixteenth International Conference on Computational Linguistics (COLING-96).
- Fasciano, M. and Lapalme, G. (1996), *PostGraphe: a system for the generation of statistical graphics and text*, in Proceedings of the Eighth International Workshop on Natural Language Generation Herstmonceux, Sussex, UK.
- Goldberg, E., Driedger, N. and Kittredge, R. (1994), *Using Natural Language Processing to Produce Weather Forecasts*, IEEE Expert, 9(2).
- Iordanskaja, L., Kim, M., Kittredge, R., Lavoie, B. and Polguere, A. (1992), *Generation of Extended Bilingual Statistical Reports*, in Proceedings of the fourteenth International Conference on Computational Linguistics (COLING-92).
- Knott, A., Mellish, C., Oberlander, J. and O'Donnell, M. (1996), *Sources of Flexibility in Dynamic Hypertext Generation*, in Proceedings of the Eighth International Workshop on Natural Language Generation, Herstmonceux, Sussex, UK.
- Lavoie, B., Rambow, O. and Reiter, E. (1996) The ModelExplainer, Demonstration presented at the Eighth International Workshop on Natural Language Generation, Herstmonceux, Sussex, UK.
- Levine, J., Rogers, I., Bennington, T. and Pattison, C. (1996), *Class hierarchies as a Multi-Purpose Knowledge Representation in a Requirements Capture and Design Tool*, in Proceedings of Expert Systems'96, Cambridge, UK.
- MacGregor, R. (1988), *A Deductive Pattern Matcher*, in Proceedings of the 1988 Conference on Artificial Intelligence. St. Paul, Mn, USA.
- Marchant, B.P., Cerbah, F. and Mellish, C. (1996), *The GhostWriter Project: a demonstration of the use of AI techniques in the production of technical publications*, in Proceedings of Expert Systems'96, Cambridge, UK.
- McKeown, K., Kukich, K., and Shaw J. (1994), *Practical Issues in Automatic Documentation Generation*, in Proceedings of the Applied Natural Language Processing Conference, Stuttgart, Germany.
- Paris, C. and Scott, D. (1994), *Stylistic Variation Multilingual Instructions*, in Proceedings of the Seventh International Workshop on Natural Language Generation, Kennebunkport, Maine, USA.
- 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 International Joint Conference on Artificial Intelligence (IJCAI-95), Montreal, Canada.
- Power, R., Cavallotto, N., and Pemberton, L. (1995), *The GIST Specification Tool*, LRE Project 062-09, Deliverable PR-3b.
- Power, R., Scott, D., and Evans, R. (1997), *What You See is What You Meant: direct knowledge editing with natural language feedback*, Technical report ITRI-97-03, Information Technology Research Institute, University of Brighton, UK.
- Reiter, E., Mellish, C. and Levine, J. (1995), *Automatic Generation of Technical Documentation*, Applied Artificial Intelligence, vol. 9, no 3.