Practical Considerations in Building a Multi-Lingual Authoring System for Business Letters

John Tait, Huw Sanderson
School of Computing + Information Systems
University of Sunderland
Sunderland SR6 0DD, U.K.
{John.Tait, Huw.Sanderson}
@sunderland.ac.uk

Peter Hellwig
Universität Heidelberg
Lehrstuhl fur Computerlinguistik
D69117 Heidelberg, Germany
hellwig@novell1.gs.
uni-heidelberg.de

Jeremy Ellman
MARI Computer Systems Ltd.
Wansbeck Business Park
Northumberland NE63 8QZ, U.K.
Jeremy.Ellman@mari.co.uk

Periklis Tsahageas
SENA S.A.
Byzantiou 2,
GR-142 32 N. Ionia, Greece
P.Tsahageas@senanet.com

Ana Maria Martinez San José
Sistemas y Tratamiento de Informacíon S.A.
Avenida del Tomillar, 13
28250 Torrelodones Madrid, Spain
sysnet@bitmailer.net

Abstract

The paper describes the experiences of a multi-national consortium in an on-going project to construct a multilingual authoring tool for business letters. The consortium consists of two universities (both with significant experience in language engineering), three software companies, and various potential commercial users with the organisations being located in a total of four countries. The paper covers the history of the development of the project from an academic idea but focuses on the implications of the user-requirements orientated outlook of the commercial developers and the implications of this view for the system architecture, user requirements, delivery platforms and so on. Particularly interesting consequences of the user requirements are the database centred architecture, and the constraints and opportunities this presents for development of grammatical components at both the text and sentence level.

1. Introduction

This paper describes our experience in working on the European Union (EU) Framework IV Language Engineering project, MABLe (Multi-lingual Authoring of Business Letters). One of the aims of the Language Engineering sector of the programme is to develop applications which help improve information access and interchange across languages (Telematics, 1996), within the broader objective of assisting Small and Medium sized Enterprises (SMEs).

Framework IV projects are expected to form the basis of systems put into practical use in the medium term, and the focus of the paper is the practical and technical problems faced by the consortium in moving the fruits of academic research towards a widely used practical product.

The MABLe project is being undertaken by a consortium consisting of :

MARI Computer Systems Ltd. (UK) - Industrial Developer

SENA S.A. (Greece)
- Industrial Developer

STI S.A. (Spain)
- Industrial Developer

Pan Hellenic Exporters Assoc. (Greece)
- Industrial User Association

Centro de Sondi E Imagen S.L. (Spain)
- Lead Industrial User

University of Sunderland (UK)
- Academic Research

University of Heidelberg (Germany)
- Academic Research

The project targets SMEs who conduct international trade as the main potential users of the system. Its goal is to build a system which will help people who wish to write a good quality business letter in a language with which they are not fluent; via an interaction with the computer conducted in their native language. The current project will build a prototype allowing Greek and Spanish users to produce letters in English. However, the system has been designed to allow relatively straightforward extension to other languages, both by making the user interface localizable, and (less straightforwardly) by storing knowledge of the output language in a language independent grammatical database.

An initial, limited, prototype of the system is now available and work has started on a second prototype with more extensive functionality. This is due for delivery for user evaluation in mid 1997.

The academic partners initially saw the project as an opportunity to move their existing knowledge and, more particularly, linguistic resources into practical use, the major challenges being in computational linguistics. In practice the commercial partners have identified many problems of a less intellectual nature which must be overcome if the system produced is both to be accepted by users and to operate within their environment. This paper will concentrate on these practical issues and hopefully provide a useful basis for those undertaking applications orientated research which they hope to move into practical use in the medium term.

The remainder of the paper is structured as follows: Section 2 reviews the development of the project from its conception following a series of conversations between academics. Section 3 moves on to describe the way the initial project was reformulated to become user-requirements orientated rather than technology driven, once we began to work with our industrial partners. Section 4 describes the architecture of the system as we are now constructing it and Section 5 covers some of the User Interface issues the MABLe project and its architecture have thrown up. Section 6 covers the issue of delivery platform which was of great importance to the commercial partners and, once selected, drove us towards the architecture described in Section 4, while Section 7 briefly reviews some of the consequences of this approach for the development and representation of the grammatical formalisms. Section 8 moves on to cover the issues of facilitating the ultimate exploitation of the software. Section 9 draws some conclusions from our work to date.

2. History

The MABLe project arose from a series of conversations between John Tait of the University of Sunderland and Peter Hellwig and Heinz-Detlev

Koch of the University of Heidelberg during 1994 and 1995. At that time the concept was very much an interactively controlled natural language generator utilising Heidelberg's Dependency Unification Grammar resources (Hellwig, 1986, 1993) combined with a newly developed text grammar for business letters. During 1995 interested commercial partners and potential users were identified, leading to the successful submission of an EU Framework IV Language Engineering proposal.

The aim, from the academic's point of view, was always to try to constrain the problem so that existing, well-developed, even old-fashioned, computational linguistic technology would suffice. Our intention was to explore the limits, in a practical setting, of what the technology could do. Our main concerns were issues like whether a user interface could be built which was not excessively cumbersome, and whether sufficient coverage could be produced within a reasonable amount of time and effort, given available human and linguistic resources.

Once the project was started, in late 1995, it become clear that there were a number of other issues of great concern to our users and commercial partners which we had not considered previously. Foremost amongst these was the need to integrate MABLe with the users' existing and likely future working environment, a requirement which created several low and high level technical problems.

3. User motivation

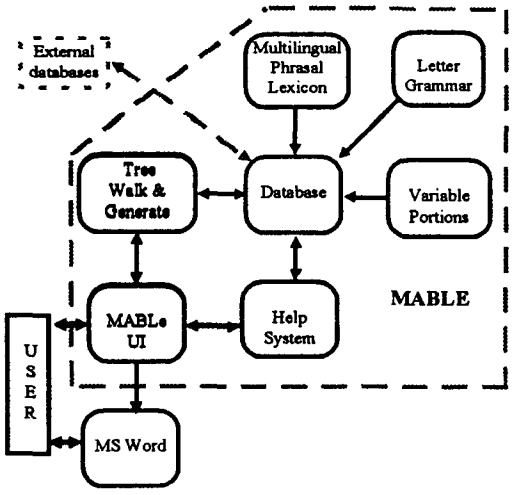
The MABLe project as it is currently formulated has as its core the belief that language technology now exists which can satisfy real user needs and requirements. Writing letters in a foreign language is one of the most common problems faced by organisations engaging in international trade. This is especially true for SMEs who are unlikely to have specialist translation teams or expensive multilingual staff. Before the project started in earnest, the MABLe partners carried out a survey to identify the problems potential users encountered when writing letters in foreign languages. This survey was continued during the early stages of the project in order to create a set of user requirements to serve as a basis for the MABLe system specification and design. The results of the survey showed the need for an application which provided:

- Compatibility with all commonly used word-processors
- Friendly and self-contained user interface
- Access to user databases (with client data, product data etc.)
- A set of common 'template' letters
- A good thesaurus of phrases
- Textual cohesion and good linkage between phrases.
- Search by expressions and keywords
- A good electronic dictionary

The application should run under MS-Windows and have an affordable price, not higher than 200 ECU.

MABLe is being developed with all these user requirements in mind, and in order to make user-requirements elicitation a dynamic process, user groups have been created in Spain and Greece to review MABLe progress and provide development feedback. The Spanish group is composed of organisations from different market sectors in which foreign business letters writing is an important task. This group will, consequently, also play an important role in the market analysis tasks to be carried out within the project, providing information on market expectations, target users' willingness, barriers to market entry, and potential competitors. The Greek user group is composed of import & export organisations.

Figure 1 MABLe architecture



4. MABLe Architecture

Following our initial analysis of user requirements we arrived at a system architecture consisting of seven main components:

1) a text grammar of English for business letters
2) a multilingual phrasal lexicon
3) a fragmentary grammar of English for variable portions within phrases
4) a database in which both static data (multilingual phrasal lexicons, grammars etc.) and intermediate results of processing are stored
5) a tree walk and generate programme which walks over the tree represented by the grammar, making decisions in consultation with the user about which path to follow through the grammar
6) a user interface through which the text is generated
7) a help system

Communication between these components is outlined in Figure 1. (The view of the architecture provided by Figure 1 is somewhat simplified, omitting for example, morphological processing)

Perhaps one of the most noteworthy features of our architecture is the central role accorded to the database. Our reasons for adopting this architecture are discussed in Section 6.

We are considering a number of changes and extensions to this model: for example the substitution of MS Word output by other formats, but these matters of detail are not really the concern of this paper.

5. MABLe User Interface

User Interface design, in language engineering especially, is often thought to be about the superficial ergonomic aspects of button and window layout on the users screen. Of course, getting this right is essential to a users basic ability to read and interpret the screen correctly. However, sophisticated products need to consider the nature of the user interaction required, and how this may be provided to the user in a controlled and controllable manner.

MABLe presents some unique opportunities in this area, since it is understood that the program is designed for users who know that their knowledge of a foreign language is less than perfect. They do not need reminding of this constantly, rather, they need guidance and support whilst being made to feel it is they that are in control of the program, and not vice versa.

At the same time, MABLe needs to fit tidily into the user's work routine, or it will not be accepted. These requirements have lead us to design several user interfaces with different philosophies.

The first interface is based on the well known "Wizard" interaction style found on PCs. Here a user is guided through an interaction, and the program selects the best options based on user input. For MABLe this implies following a rigid tree walk of the letter grammar to produce grammatical output. Although the user may track both his input and the target language output, there is a rigid feel to the program, with the user feeling out of control. Thus, even though this interface is integrated with Microsoft Word it has not found instant popularity with the users in the second phase of our user requirements work.

Our second interface is based on representing the grammar as a clickable tree in a standalone application. This has the advantage of the full screen area being exploited with users choosing phrases as they like. Output is subsequently exported to a Word Processor.

Here one may choose phrases at will, and not necessarily following a coherent form. However, one could say from work in text, planning and Rhetorical Structure Theory (de Beaugrande and Dressler, 1981, Appelt, 1985, Mann and Thompson, 1987) that discourse coherence is enforced by the user, and is language independent, so that in practice this matters little.

Clearly we are currently looking at merging both approaches. That is, we need to advise the user of grammatical or foreign language conventions that we may derive from our letter grammar, whilst leaving him free to make errors if he wishes.

As we have seen, MABLe's letter grammar gives users the possibility of creating coherent business letters. It is the task of the User Interface designer to ensure that the user takes advantage of these possibilities.

6. Delivery Vehicles

As noted earlier, the initial targets of MABLe are those Small-to-Medium sized Enterprises (SMEs) undertaking international trade. Our initial user survey work indicated that for delivery and evaluation on Spanish and Greek user sites within the time frame of the project (i.e. for a December 1998 completion) earlier versions of the MABLe system would need to integrate with Microsoft Word 6.0 running under Windows 3.1. This combination was still likely to be the predominant operating environment at user sites and user PCs with more than 8 or 12 Mbytes of RAM would be unusual. In fact when the user survey was undertaken the majority of Greek sites were still operating with DOS and WordPerfect and the like, though they expected to change shortly. Although Microsoft would lead developers to expect predominance of Windows '95 and Word 7.0 (or even Windows NT) at much earlier dates, to rely on this occuring in our target market seemed to produce significant additional commercial and technical risks for the project.

From the NLE point of view this decision has profound implications. Most research computational linguistic systems are built to run on UNIX workstations with very large virtual address spaces, supported by 64Mbytes or more of real RAM. Windows 3.1 is a 16 bit operating system, implying a much smaller address space, and although 32 bit programming is possible it is by no means completely reliable and satisfactory.

Typically the products of computational linguistic research require large amounts of RAM, tending to operate with their grammars and even their dictionaries as in-store data structures. Taking into account the restrictions on address space and RAM,

along with the need to build a well engineered system which could be easily modernised, we chose to store our grammar, lexicon (mainly phrasal) and intermediate results in a relational database on disc.

Having adopted a database-central architecture with comparatively modest demands on virtual memory and real RAM, we can easily move the system on to more modern systems with fewer restrictions. Specifically, systems with full 32-bit addressing support (like Windows 95 or Windows NT).

Early technical feasibility work has led us to adopt the MS Access database, because of its wide adoption in the commercial marketplace and the comparative ease with which it may be integrated with other Microsoft products. However, MABLe could easily be adapted to work with most common alternative relational database systems.

7. Grammatical Representation

It is important that the user be able to find the letter or letter fragment they require, and this 'navigation issue' has implications for the design of the grammar. There are many ways to sort types of letter: by general subject, e.g. 'complaints', or by transaction grouping, e.g. all those letters involved in the process of buying goods.

The overlapping of such sets of letters calls for internal representation as a directed graph rather than a tree, a design which is also appropriate for the internal structure of letters, allowing re-use of commonly recurring sub-structures such as formulaic openings etc.

Similarly, a 'direct-search' approach to reaching a desired letter/letter-fragment imposes a need for systematic naming of non-terminals, especially if incremental search of the type offered in standard 'help' systems is made available to the user

The hierarchy and internal structure of the business letters is represented by a grammar with a context free phrase structure skeleton. However, in order to enforce consistency of register, lexical cohesion and subject matter, a set of attributes of varying complexity can be associated with relevant constituents.
The formalism is described as an SGML Document Type Definition (DTD) both to facilitate eventual re-use of linguistic resources and to define and constrain the power of the grammar precisely.

UNIX tools have been created to allow automatic conversion from the SGML format to an agreed Access database format accessible to the Tree Walking algorithm that creates the final letter in conjunction with the user.

Additionally, a database browsing program has been written to allow convenient construction of the grammar via Access. For the linguist constructing a large grammar, this browser allows tree structure to be easily viewed and modified. Furthermore, it is envisaged that end users may wish to make simple additions and changes to the letter grammar, and a method more portable and user friendly than SGML editing was required. Envisaged customisation of the grammar includes the addition of new slot-fillers (products, companies, etc.) and the addition of simple letter templates. Database integrity is vital, so in order to ensure grammar extensions introduced with the browser follow the given formalism, a set of database queries are employed to regulate the Access tables.

It is hoped the approach adopted will support the required level of textual cohesion, but it would be premature to conclude that it does at this stage.

8. Exploitability

Commercial exploitation is the main aim of the MABLe consortium. In addition to the creation of the MABLe user groups, which complement the work of the user partners in ensuring that product development follows the evolution of user needs and expectations, MABLe is undertaking a series of dissemination activities to ensure that the project is internationally promoted. These include production of leaflets, organisation of conferences, and publications.

Commercial success hinges on the elaboration of a good exploitation plan. This entails starting with a suitable product definition, continuously monitoring competition and analysing the benefits and risks inherent to the potential market. MABLe has a clearly defined target market: writers of foreign business letters with some knowledge of the target language, but who are not sufficiently fluent to produce output of the necessary quality.

The essential requirements of these writers appear to be clear and the consortium is progressing with the development to satisfy these under the understanding that MABLe will be an attractive product which will:

- help organisations (in particular SMEs) to overcome language problems
- enable the creation of high quality unambiguous business letters
- provide consistent letter authoring standards
- provide stylistic suitability of content

With these characteristics it is thought that MABLe will help to improve productivity, reduce overheads by eliminating unnecessary waste of time and resources, and, consequently, be a useful tool which organisations will be prepared to buy.

The consortium is aware of the risks inherent to such a development: the need to match development time and market maturity, choose the right technologies and platforms, and to control costs and obtain scale economies. With the ability to control these risks, the development driven by the user requirements to ensure position as high performance, high value, and long lead time product, and the appropriate marketing and promotion activities, the consortium is confident of reaching the target market in a short time after the project concludes.

9. Conclusions

Even at this comparatively early stage in the project there are two initial conclusions which may be drawn.

First it is has been difficult to re-use existing linguistic resources. The demand for different delivery platforms, the constraints of a new task and so on have driven us to primarily re-use our language engineering *knowledge*, rather than re-use existing resources and software, although it is hoped that indirect re-use of existing resources in radically re-processed forms may be possible in later parts of the project.

The second conclusion is that issues like real user requirements, in terms of interfaces, delivery platforms and so on are perhaps too little considered on many Language Engineering projects.

Acknowledgements

The authors acknowledge the financial support for this work provided by the European Union under the Telematics Applications of Common Interest Programme, through Language Engineering Project LE1-1203.

References

D.E. Appelt. 1985. *Planning Natural Language Utterances.* CUP, Cambridge.

R de Beaugrandeand and W. Dressler. 1981. *Introduction to Text Linguistics.* Longman, London.

Peter Hellwig. 1986. Dependency Unification Grammar . In *Proceedings of the 11th International Conference on Computational Linguistics COLING 86*, pages 195-198.

Peter Hellwig. 1993. Extended Dependency Unification Grammar. In Eva Hajicova, editor, *Functional Description of Language*, pages 67-84. Faculty of Mathematics and Physics, Charles University, Prague.

W.C. Mann and S.A. Thompson. 1987. Rhetorical Structure Theory: A Theory of Text Organisation. *ISI Reprint Series ISI/RS-87-190.* Marina del Ray (CA): Information Sciences Institute.

Telematics. 1996. WWW URL: http://www2.echo.lu/telematics/off-docs/brochure.html inspected on 4 March 1997.