9 CLRU, Cambridge

by Bill Williams, Director and Trustee

The Cambridge Language Research Unit - a Registered Research Charity - was founded in 1956 by a group of notable Cambridge academics. From its inception, the Unit's underlying consistent interest has been directed to a better understanding of the complex convoluted structures, social evolution and idiomatic usage of natural language in both textual and oral forms. At an early stage it was realised that natural language - essentially the product of those who daily use it - was influenced by the entire gamut of human experience and activity, and was not the sole prerogative of the relevant professional disciplines. Over the years, therefore, representatives from a broad range of disparate disciplines have been involved with Unit members in developing theoretical and practical approaches to the understanding and handling of natural language.

Since its founding, the Unit's core activities have centred on systems and methods of handling natural language (until recently confined to textual inputs) by mechanical (i.e.computer) means, with the intention of developing sophisticated machine translation systems actually capable of producing outputs acceptable to the native users of the target language. Considerable experience has been gained from the many linguistic, lexicographical and morphological investigations carried out for British, American and European government and private agencies. Particularly useful to the unit has been the practical experience gained in compiling computerised dictionaries, specialist in both content and application. The Unit has also carried out consultation, design and implementation of 'Add-on' facilities and performed advisory functions for a number of commercial MAT and MT systems.

These varied areas of work have provided the Unit with considerable insights into the difficulties involved in, and the nature of the optimal methods for handling natural language textual material. This expertise enabled the Unit to explore natural language handling techniques for the British Library. The BL project was specifically geared to information retrieval systems involving textual material. The Unit has thus over the years accumulated a comprehensive experience in most aspects of natural language handling: information retrieval, text structuring, paragraph processing, machine translation, etc..

Since 1988, a concerted and sustained effort has been made to correlate and integrate the mass of data and experience of the previous years and to produce an integrated handling system and coherent programme of research which takes into account the latest developments in linguistic, psycholinguistic and technical knowledge in this field.

The concept of the Cognitive-Linguistic Unit (CLU) has been developed by the Unit. The CLU is defined as part of a text string (usually part of a sentence) which, in isolation, maintains a unique element of meaning and comprehensibility to the user. The CLU, whilst generally grammatically acceptable, more significantly is close to the word groupings that linguistic and psycholinguistic research would indicate to be the most easily understandable and readily memorizable by the user.

The Unit has implemented the CLU concept on a micro-computer and it is being operated on a large, and growing, diverse body of text. Currently the system produces approximately 92% acceptability of the CLU's obtained. It is interesting to note that the basic algorithm developed to handle English is usable, with only minor adjustments, for some twenty-five other languages to which, so far, it has been applied. In order to implement this approach it has been necessary to develop various sophisticated dictionaries, to utilise and extrapolate from syntactic information in unique ways and to develop specialised software to cope with the vagaries of the various languages under examination. However, it has become increasingly clear that to achieve the high level of accuracy in translation of natural language aimed for by the CLRU, there are limitations in the dependence on syntactically derived operating information - to achieve the target of 99.8% translation accuracy, a major semantic contribution is required.

In 1992 a much larger and more ambitious computer system was installed with a capacity adequate to accomodate the thesauri necessary to complete the envisaged CLRU Translation operating system. An extensive thesaurus is already in place and work is proceeding on the further necessary thesauri required for the first two target languages, namely German and Portuguese.

Apart from satisfying the requirements of CLRU's own research and development programmes, the new computer equipment now greatly enhances the Unit's capabilities for offering comprehensive research and development facilities in translation and linguistics to outside organisations.

Note. For further information contact Mr.Williams at CLRU, 11 Millington Road, Cambridge CB3 9HW, Tel. 0223 359625, Fax. 0223 359613. Editor