3. History, General Aims and Interests of the Cambridge Language Research Unit

The Cambridge Language Research Unit is a Registered Research Charity. It was founded in 1956 and over the years people from a fairly broad range of diciplines have been involved in developing various theoretical approaches to the topic of main interest to the Unit, that of natural language.

Since its founding and up until recently the Unit's principle activities have centred method of handling natural language (specifically in text form) mechanically, with a view to developing sophisticated machine translation systems genuinely capable of producing outputs acceptable to native users of a particular target language. This has involved carrying out many linguistic, lexicographical and morphological investigations for British and American government and private agencies. It has experience of the compilation of computerise dictionaries, specialist in both content and application. more recently its activities in this domain have been largely concerned with the EEC Systran Machine translation system in Luxembourg.

This work has provided both experience and insights into the difficulties involved in, and the nature of the optimal methods for, handling natural language textual material. the Unit has recently used this expertise to explore natural language handling techniques for the British Library. This project is specifically geared to information retrieval systems, again with textual material. This has meant that, in the recent past, aided by the availability of cheap, reliable and sophisticated micro-computer systems, the Unit has expanded its work to deal with most aspects of the handling of natural language textual material (such as information retrieval, text structuring and paragraph processing), although Machine Translation is still a major interest.

The work currently underway has involved a concerted and sustained effort to correlate and integrate the accumulation of data and experience of the previous years (which was sometimes, of necessity, somewhat erratic and inconsistent) to produce an integrated handling system and coherent programme of research which takes account of modern developments in linguistic, psycholinguistic and technical approaches in this field.

The Unit itself has developed the concept of the Cognitive-linguistic Units (CLU) which is defined as part of a text string (usually part of a sentence) which, in isolation, maintains an element of meaning and comprehensibility to the human 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 to, and readily memorizable by, the English speaker.

This concept has been implemented on a micro-computer and has been operated on a fairly large and diverse body of text. Currently the system can produce approximately 90% acceptability of the CLUs obtained. In order to implement this approach, it has been necessary to develop various sophisticated dictionaries, to utilise syntactic information, and to develop fairly sophisticated software to cope with the vagaries of the English language. All these developments had necessitated detailed considerations of most aspects of English language textual material. The Unit is currently extending this work with a view to integrating into the present system a degree of semantics to allow higher level structuring of text.

There are always capabilities for grammar checking, creation of specialist dictionaries involving lexical analysis, cross-referencing and text structuring. We are also developing thesauri for our system which should enable more sophisticated language handling techniques to be used.

It should be noted, however, that research in the Unit was, and is, developed from a consideration of the human use of language, and then implemented on a computer to reflect those needs. Any approach which started from the capabilities of the computer would encounter insurmountable difficulties. It is essential to keep in mind the basic aims of any such system; that is, to analyse information produced by humans in such a way that the results can be used easily by humans.