STATEMENT OF REQUIREMENT

LSA AND HAL METHODS FOR THE SEMANTIC ANALYSIS OF TEXT

1. Background

1.1 The ability to handle large volumes of data is a central concern for intelligence analysts. This is particularly true in the case of text, where it is important to be able to find a few key documents within a large volume of available material. The obvious means of addressing this issue is to provide some level of automation in the semantic analysis of sets of textual documents. Algorithms that are capable of making rapid semantic judgments can assist analysts by focussing the search of available documents, and allowing closer attention to be given to those documents that are more likely to provide valuable information. For example, measures of semantic relatedness may allow an analyst to identify a cluster of documents that are similar to one just found to be relevant, and this cluster is likely to contain more relevant documents.

1.2 DSTO has recently developed a capability to represent and visualise the information provided by semantic similarity measures. In the text domain, some use has been made of the established n-gram technique to generate these measures. There are, however, two other widely used and apparently successful methods for the semantic analysis of text. These are known as Latent Semantic Analysis (LSA), and the Heuristic Analogue of Language (HAL). Each method provides a means for determining the similarity of meanings of words or passages by analysis of large text corpora.

1.3 LSA performs a mathematical manipulation known as a singular value decomposition on a matrix that records the co-occurrence of words or passages in a set of documents. This manipulation constructs a high-dimensional 'semantic space' in which words, passages and documents are represented by points in such a way that their semantic similarity is readily measured. HAL develops similarity measures by passing a 'window' across a set of documents, and recording the co-occurrence of words or passages within this window. This record of co-occurrence is then used to generate vectors from which semantic similarity may be measured.

1.4 This contract is intended to provide software that implements the LSA and HAL methods. This software is initially intended to facilitate the use of these two techniques in DSTO representation and visualisation research. Ultimately, however, this research effort should enhance the automatic text analysis capabilities available to intelligence analysts.

2. Deliverables

2.1 Software that implements the LSA method on text corpora provided by the user, together with associated documentation, December 1999. Software is required to be installed and executable on a PC platform, and be implemented in such a way that large text corpora can be processed in reasonable time on a high-end

PC. Manuals are required in hard copy and electronic copy, in either MSword or LaTeX formats.

2.2. Software that implements the HAL method on text corpora provided by the user, together with associated documentation, March 2000. Software is required to be installed and executable on a PC platform, and be implemented in such a way that large text corpora can be processed in reasonable time on a high-end PC. Manuals are required in hard copy and electronic copy, in either MSword or LaTeX formats.