A SYNTAXICALLY-BASED PREPROCESSOR
FOR A LIMITED EXPERIMENTAL ARABIC DOCUMENT RETRIEVAL SYSTEM

by
Farid Mohammed Salim Ibrahim
B.Sc., H.Dip., M.A.

A Doctoral Thesis
submitted in partial fulfilment of the requirements for the award of

The Degree of Doctor of Philosophy

of the

Loughborough University of Technology

March 1988

Supervisor: Gwyneth Tseng
Department of Library and Information Studies

© F.M.S. Ibrahim
A SYNTACTICALLY-BASED PREPROCESSOR

FOR A LIMITED EXPERIMENTAL ARABIC DOCUMENT RETRIEVAL SYSTEM

SYNOPSIS

The research reported in this thesis is about the description and discussion of an experimental document retrieval system for Arabic texts, using linguistic methods of analysis. Specifically, Arabic presents difficulties for the efficient retrieval of information because it is an agglutinative language, thus rendering the stop list method (as commonly used for English texts) near to useless.

The system has two stages: the creation of the retrieval lexicon and the search program. The latter is done using a limited on-line searching which allows for partial matching. The former has four stages. Texts in the form of abstracts are processed by morphological analysis, syntactic analysis, term extraction and term manipulation modules. Each stage produces a new representation of the source text. The morphological analyser attempts to recognise any prefixes and/or suffixes attached to the words in the corpus being processed. It also assigns grammatical labels specifying the part of speech using a contextual analysis of individual words (assuming that the inflectional features of a word are indicative of its syntactic role). An augmented transition network grammar and parser have been built for this purpose. The same parser has
been developed and used in the second stage which is syntactic analysis.

It takes as its input the representation of the text created by the morphological analysis, and uses a separate grammar file defined as a recursive transition network. The aim of syntactic analysis is the definition of the relations of the different constituents in the individual sentences being processed. The information added by the morphological and syntactic analysers is used in the term extraction module. This module uses a traversal algorithm to negotiate the structure built by syntax, utilising a set of rules, kept on a file, specifying the type of constructs needing to be selected. The manipulative module generates new entries for each term selected by rotating its elements.

The system has been implemented using the Hull V-mode Pascal compiler available on the L.U.T. Prime System. It has been tested using 40 abstracts selected from a conference proceedings in the field of computer applications. The results obtained were encouraging, particularly in the identification of affixes (a success rate of 89% for suffixes and 85% for prefixes) and in the identification of syntactic categories (a success rate of 98% for nouns and 79% for verbs). These figures have highly influenced the identification of the syntagmatic relations underlying the word order in the sentence.

The research concludes that although successful results have been obtained without the aid of a pre-constructed Arabic lexicon, many errors can be avoided by the inclusion of small dictionaries. This research names these particularities.