

A Term Base Translator Over The Web

Mustafa Yaseen†, Bassam Haddad ‡, Harris Papageorgiou §, Stelios Piperidis §, Mamoun Hattab*,
Nick Theophilopoulos +, and Steven Krauwer #

† Computer Science Dept.
Amman University
Amman, Jordan
myaseen@cbj.gov.jo
‡ bh@go.com.jo

§ Language Technology Applications Department
Institute for Language and Speech Processing-ILSP
Athens, Greece
xaris@ilisp.gr, spip@ilisp.gr

* Arabic Textware
Arabic Language Core Technologies Dept.
Amman, Jordan
Arabtext@go.com.jo

+ Impetus Engineering
Athens Greece
impetus@otenet.gr

Utrecht University / ELSNET
The Netherlands
steven.krauwer@elsenet.org

Keywords: *Terms Based Translation, Morphological Analysis, Canonical Forms of Vocabularies.*

Abstract

This paper is an attempt towards producing a utility to process documents over the Web in diversified formats in a multilingual mode, and produce semi-translated documents from Arabic to English and the other way around, by using a term based approach. The process incorporates morphological analysis techniques of Arabic to handle the canonical forms of vocabularies in the terms stored in the *termsbase* database.

1 Introduction

The objective of research into natural language processing is to make computers deal

intelligently with the diversity, complexity and variation of human natural languages. The efforts of researchers in this area have produced core technologies that resulted in adopting methodologies, creating systems, many tools and utilities that enabled the realization and, to some extent achieving this goal (Euromap 1998).

The current Information Age is characterized with explosion of information and great demand for effective and natural communication requirements. There is a critical demand for the utilization of research in the natural language processing area, to be utilized and expanded to be an integral part of the relations in general, with the digital information as a whole (Euromap 1998).

The Internet is rapidly bringing to the foreground the need for people to be able to

access and manage information in many different languages (Hovy 1999). Research in Arabic NLP is very rich in areas such as morphology, moderate in syntax analysis, still not very mature in semantics and lexicon building. Selected references for work in various areas were given to demonstrate part of the work being developed in various areas.

Several projects have been launched recently, many have achieved significant results, and they are all centered around research in Arabic NLP with support from programs financed by the EC, they all fit within the scope outlined in the (Euromap 1998) policies and strategies concerning the multilingual directions within the funded projects; e.g. (Pease 1996), (Yaseen 1998), (Belhadij Kacem 1998) (Dichy1998).

The overall objective of this effort is to provide a multilingual utility for semi-automatic, terms based translation of documents over the Web. The current system works for English to Arabic and Arabic to English translation, as a second phase the Greek component will be added to achieve the multilingual capabilities.

A specific domain is chosen as a pilot for the testing of the utility, and it is started with the creation of a multilingual *termsbase* in the financial domain. This *termsbase* has started with an initial set of approximately 3000 terms and their equivalents in the two languages Arabic, and English; Greek to be added later (*AR-EN-EL*); to be processed by the *Term Base Translator* component of the *NAPLUS* system. There has been an incremental increase in the number of terms in the *termsbase* in the subsequent cycles of the system.

2 Overview of the *NAPLUS* Project

The *NAPLUS* project fits within the scope of the European vision for building a Multilingual, cross Cultural, Information Society (Yaseen 1998). The original proposal was to develop a prototype system for processing and understanding the Arabic language, through producing core technologies, utilities and tools that could be used in many areas such as automatic translation, text abstraction, question answering interface to databases, or any other related areas of application.

Moreover, study and analysis of semantic issues is under going starting with a suitable knowledge representation for Arabic, and on the lexicon structure to support such requirements. The original proposal was based on several components that constitutes an understanding of the language; they could be summarized in the following modules:

- ⇒ **Morphological Analyzer.**
- ⇒ **Syntactic Analyzer.**
- ⇒ **Semantic Analyzer.**
- ⇒ **Building a Lexicon.**

including building the *Roots*, *Patterns* and the language matrix defining the Arabic language by linking Roots with Patterns, identifying each part of the speech; i.e. *Verbs*, *Nouns*, and *Particles*; and finally to determine if a vocabulary is a member of the language or not; i.e. if it is an inflection of a known Root and a valid Pattern. This constituted the morphological analyser that is enhanced with linguistic rules to speed up the search and minimize the ambiguities. Currently the syntactic and semantic analysis is under investigation.

As a first cycle it was decided to address the issue of multilinguality from the prospective of a Term Based Processing tool for a specific domain of knowledge, e.g. Financial Domain. Thus the issues of syntactic analysis and semantic analysis will not be addressed in the current paper.

3 The Concept

The application scenario envisaged for this part of the *NAPLUS* project could be considered to serve a well-defined goal (uniformity and terminological translation quality and consistency) by providing a set of tools aiming at terminological database management during the translation process. Specifically, our focus will be the processing of MS WORD, Text format or HTML documents in Arabic and/or in English in order to spot and highlight terms (single or multi-word terms), and subsequently to substitute them with their English or Arabic equivalent, which will be retrieved from the term base. Later the Greek language will be added and handled similarly.

4 The Scenario – Term Spotting and Term Substitution:

The first cycle of this application scenario consists of the following phases:

- I. The creation of a terminological Database: A multilingual Terminological Database (Arabic - English -Greek) in the financial domain is developed. This phase deals with the structure and content of the terminological database, which we will call *termsbase*. Issues that have to be addressed here are among others: the design of the *termsbase*, the representation of the terminological data, and the database engine to process the *termsbase*.
- II. Term level translation: Automatic terminology lookup could be thought of as the term level equivalent of machine translation. Based on the previously created *Termsbase*, conceptually this phase will provide the functionality to automatically spot “possible terms” in an Arabic document and retrieve their translation equivalents in the predefined target (Greek or English) language. A relational database approach is utilized for retrieval and maintenance of the *termsbase*. The utility for this sub-module is built using *MS ACCESS*.
- III. Morphological analysis of the vocabularies in the spotted terms to link variations

(different inflected forms of the word) with the same term in the *termsbase*.

The importance of infrastructure for such a translation exercise becomes more evident in multilingual situations. Elements and modules of the infrastructure need to be integrated, both among them and within the application framework. These modules of the infrastructure are the following:

- The Arabic morphological database,
- The Arabic morphological analyzer,
- The multilingual *termsbase* (GR-EN-AR)

Furthermore, in order to enrich the *termsbase* with additional entries in the same or other domain, a term candidate extraction process could be foreseen as the next cycle in the application workflow of the *NAPLUS* project.

Term Extraction research can draw on many resources, including morphological analysis, partial syntactic processing and some sort of statistical filtering. As an example, in a bi-text we may assume that the term extraction process provides us with a set of term candidates in the source language. The same process could also be applied to the target language part of the bi-text.

A consequent human introspection and verification of the results would provide additional bilingual terminological data that could be integrated in the *termsbase* for further exploitation.

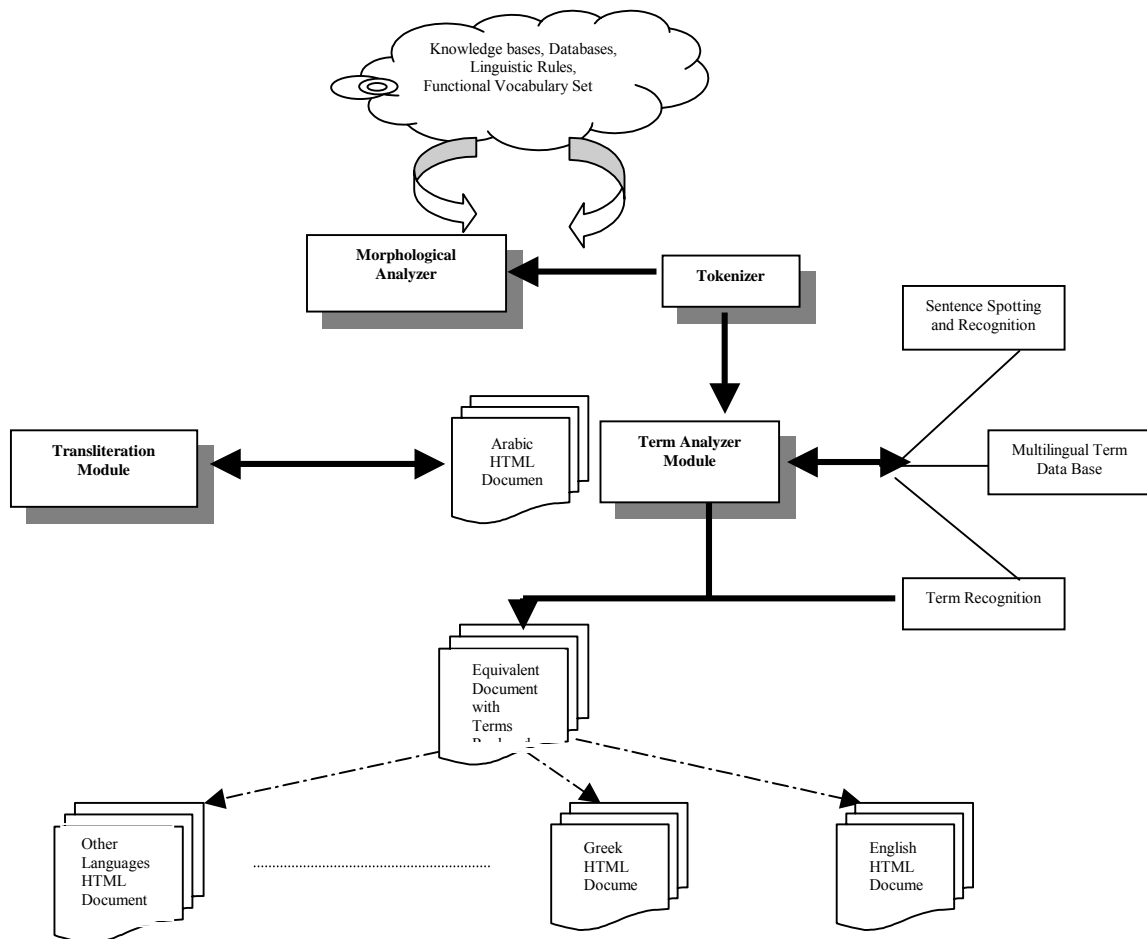


Figure 1: A Schematic Diagram of the Main Components along with the Flow of The System

5 1st Cycle System Components

The major components used for this first cycle of the prototype, are the following, as illustrated in Figure 1:

1. A MS-WORD, HTML and/or Text interface providing users an API to our *lingware* (morphological analyser, tokenizer, *termsbase*, ...etc.).
2. Arabic tokenization of the HTML or Text Arabic document in order to identify sentences and word boundaries.
3. Morphological processing of the Arabic sentence(s) in order to represent the word forms by their canonical forms.
4. An automatic terminology lookup of these canonical forms of words in the *termsbase*. As a result, all actual terms including the multi-word terms will be identified in the Arabic document and presented to the user.

As an example, suppose that the user input consists of the following Arabic sentence $\langle aw_1 aw_2 aw_3 \dots aw_n \rangle$. This word sequence is converted to an equivalent canonical-form-based sequence $\langle acf_1 acf_2 acf_3 \dots acf_n \rangle$ after morphological analysis takes place. Subsequently, each canonical form acf_i that could be a candidate term (e.g. nouns, adjectives, verbs, ...) is matched against the *termsbase* and all terms that this canonical form is a member of are retrieved. Human inspection and intervention are facilitated through the interface in order to select the appropriate terms for the specific Arabic sentence. Alternatively, this term selection process can be done automatically. At the end, all selected terms are translated to their equivalents in the predefined target language.

It is worth mentioning that the same logic applies to processing documents in English

giving the equivalent translated terms in Arabic, with one exception that the morphological analysis is not done to the English vocabularies, and thus no canonical forms in the *termsbase* are considered.

In what follows, we provide a detailed proposal of a draft design of the NAPLUS *termsbase* schema grounded on the specific characteristics of our application scenario which call for efficient storage of monolingual as well as bilingual metadata allowing for fast and accurate retrieval of terms.

5.1 The specific running components

The following components have been developed within the system and are running in the 1st Cycle of Terms Translator:

- Build terms data files manually
- Database structure design
- Build Database tables
- Build Data conversion utility to upload the database
- Convert data created manually to database
- Remove data vocalization since at this stage vocalization is not considered, but will be kept for future extension of the system
- Include Arabic Roots tables created manually from basic traditional dictionaries, currently there are more than 5000 identified Arabic roots.
- Create the Arabic language matrix consisting from inflections of Roots with Patterns, to be used in the Morphological Analysis. Currently there are approximately 1200 different Patterns (Vocalization, i.e. *Diacritics* are considered but not used) resulting in approximately 70,000 entries in the matrix.
- Enhancing the morphological analyser with linguistic rules to reduce the ambiguities associated with

morphological analysis of Arabic vocabularies.

- Cross reference the *termsbase* with the language matrix by adding Arabic roots Matrix to the related terms tables, to allow identification of inflected words within the terms.
- Build relations on Arabic roots data and the related terms. Adding intelligence in the identification process to go beyond the term spotting (recognition) based on exact matching utilizing the morphological analysis in the process.
- Production of the final Terms database with all the relationships with the (*Root, Pattern*) Matrix.

5.2 Building the databases

5.2.1 Building the Initial Language Matrix

The following steps were implemented in building the Arabic Language matrix, which constitutes the nucleus for the morphological analyzer. The *Matrix* consists of Rows representing the basic *Arabic Roots*, and columns representing the basic traditionally accepted *Patterns* taking vocalization into consideration.

At the current initial stage vocalization will not be considered but in future when dealing with syntax and semantics this will be a major part to be considered, so at this stage we will keep this information and use it later, it is a by-product of the current phase. The intersection of the *Roots* and *Patterns* represents the inflections of the roots, i.e. generating the whole language, algorithms for building morphological analysers and for finding inflections of vocabularies were given in (Yaseen 1998, and Feddagh).

Building the language matrix involves manual and automatic procedures. The manual process basically relies on scanning traditional dictionaries and lexicons of Arabic language creating a text file of entries. The automatic procedure is a process where the text file is processed and a database is created. Concentration has been done on Modern Arabic, thus a process of intersection between various dictionaries is done to eliminate the unused roots

in the current modern language, keeping advantage of the inflections found in the old dictionaries.

The following are the processes involved in building the matrix.

- Build Arabic language text files containing roots and patterns manually from traditional Arabic language dictionaries and lexicons such as *Taj Al Arous*, *Lisan Al-Arab*, *Al Qamous Al Muhaet*, and *Al Mujam Al Waseet*. The first three lexicons are traditional basic lexicons that were developed in the thirteen and fourteen century, containing more than 12000 roots. While the fourth is a modern lexicon that eliminated many of the old deserted roots that are not currently used in Modern Arabic. This process reduced the number of used roots to approximately 5000 roots.
- A process of building the intersection between the various roots and patterns from the different dictionaries to eliminate the unused roots and to build the inflected words. This process resulted in approximately 1200 patterns.
- Scan the text file to extract and build the Arabic language components:
 - o Roots table.
 - o Patterns Table.
 - o Cross reference table linking Roots with Patterns to create the Language matrix.
- As a result a matrix of approximately 70,000 different vocabularies is created.
- Build utility to fill the database tables from the text files.

5.2.2 Building The Terms Database

As before the process of building the terms database involves manual and automatic process. The manual process is to extract and identify terms from a specific domain of knowledge. The *Financial Domain* was selected. A text file is created for the terms (consisting of

one or more words along with functional words). In the text file, each term is entered in one line, to be used at a later stage in loading the *termsbase* database with the terms and the related information. The process of loading the database with the terms is done automatically, where each line is processed and information related to the term is collected, identifying the words, their order, their roots, their types, and any other relevant information as required in the design of the structure of the term extractor database requirements.

The following is an overview of the steps and processes involved in this stage:

5.2.3 Terms Initial Data Construction

- Receive initial terms data structure from 1st cycle documentation.
- Modify the data structure to enable it to handle the applications requirements.
- Build terms data file manually. This file contains terms in Arabic and its equivalent in English and Greek.
- Build functional words tables, information in those tables is basically linguistic information: Arabic and English functional words.
- Build utilities to transfer from text file to database tables containing words of terms in Arabic and English language. Also information related to the words in a specific term in Arabic, and the same for English language case. The detailed information related to the terms is built and gathered too. Finally, the *termsbase* is built which relates Arabic, English and Greek terms.

The first cycle of this application scenario has terminated. The structure and content of the terminological database, (*termbase*) has been finalized. Two bilingual Terminological Databases (Arabic-English, English-Greek) have been developed. The first *termbase* (AR-EN) consists of 16000 terms in the financial domain. The second *termbase* (EN-EL) consists of about 3000

multi-word terms, in the boarder financial domain.

5.2.4 Applying Language Matrix on Termsbase

To allow the system to cater for inflected words (canonical forms with same root origin), and not only to relate just to the root base; this will allow matching not to be based only on exact match of terms. A Building utility is designed to fill the terms related tables with data from the language tables. Allowing linkages between the terms and the different forms of valid inflections within the terms. Some utilities were developed and were used for this purpose.

6 Conclusion and Final Results

The system that is produced at this stage is a step towards handling documents over the Web in Arabic English and Greek, in the future the methodology could be extended to incorporate other languages, This is a utility that will help interacting with the internet in a multilingual mode.

It is worth mentioning to note that in order to handle any other language (i.e. in a multilingual mode), some modifications are needed to incorporate the new languages, and the same processes will apply. The system can handle documents in various formats among those is HTML, MS-Word and Text. The output could be in any format that the users like to specify.

The *termbase* database is rich, in the financial domain, this exercise started with 3000 terms, then grew to 16000 terms, and recently it is enriched with a set of another 2500 terms in the area of *Payment Systems, collected from publications of the Bank of International Settlements BIS, in Switzerland*. A term collection utility will be developed to allow for automatic collection of terms without specifying the domain.

This is a step towards machine translation, it could be used in many applications such as indexing and parallel text processing. The work

in the other areas of syntax, semantics and the lexicons will definitely enhance the behavior of the system and will complement its role in achieving the understanding of the Arabic language. Currently, work is underway to enhance the performance of the morphological analyzer to achieve less than 35 microseconds response, currently available, for analyzing a vocabulary. This will facilitate using the system over the Web and making the response instantaneous. Figure (2) represents a sample of processing a document giving the output of equivalent terms.



Figure 2: Sample output from the system

References

Al-Fedagi, S.1989, Al-Anzi, F., "A New Algorithm to Generate Arabic Root-Pattern Forms", Proceedings of the 11th National Computer Conference, Saudi Arabia, 1989, pp. 391-400.

Al-Hanash, M. 1992, "A Computational Linguistic Approach for Building an Arabic Lexicon", (In Arabic), Proceedings of the Conference on Using Arabic Language in IT,

King AbdulAziz Library, Riyadh, Saudi Arabia, 1992, pp. 363-395

Ali, N. 1998 "New Paradigm for Arabic Computation". The First Conference on Language Engineering, Cairo, Egypt, March 1998, pp. 24-28.

Ali, N. 1992, "Parsing and Automatic Diacritization of Written Arabic: A Breakthru", Proceedings of the 13th National Computer Conference, Riyadh, King AbdulAziz City for Sc. & Technology, Saudi Arabia, 1992, pp. 794-812.

Ali, N., 1989 "Formalization and Computation of Arabic Syntax", Proceedings of the 11th National Computer Conference, Saudi Arabia, 1989, pp. 309-320.

BelhadijKacem 1998, IRS-based document localisation (idol) co-ordinator: epos Etudes et Programmation en Optimisation et Software, France (Rafik BelhadijKacem); 1998-1999.

Dichy 1998, short term achievement of a corpus-based multilingual basic Arabic lexical db and related resource-productive tool-box (DIINAR-MBC) Co-ordinator: Université Lumière-Lyon 2, France (Joseph Dichy) ; 1998-2000.

Euromap 1998, The EUROMAP Guide to Multimedia Content & Language Technologies; in the Information Society Technologies Programme: 1998-2002; Release 3.0; December 1998.

Feddag, A.1992, "Arabic Morpho-Syntax and Semantic Parsing", Proceedings of the 13th National Computer Conference, Riyadh, King AbdulAziz City for Sc. & Technology, Saudi Arabia, 1992, pp. 717-770.

Higazi, M.F. 1992, "Computers and Arabic Lexicons", (In Arabic), Proceedings of the Conference on Using Arabic Language in IT, King AbdulAziz Library, Riyadh, Saudi Arabia, 1992, pp. 43-52.

Hilal, Y.1985, "Arabic Morphological Analysis" {In Arabic}, Proc. Of the Computer Processing of Arabic Language, Kuwait, 1985.

Hovy 1999, Multilingual Information Management: Current Levels and Future Abilities; Report Commissioned by the US National Science Foundation and also delivered to the European Commission's Language Engineering Office and the US Defense Advanced Research Projects Agency; editors: Eduard Hovy, (co-chair) and Nancy Ide (co-chair); April 1999.

Kareh, S. 1990, Cote, P., Maamouri, M., "Development of Computer Tools to Analyze Arabic Questions", Proceedings of the 12th National Computer Conference, Riyadh, King Saud University, Riyadh, Saudi Arabia, 1990, pp. 693-705.

Khayat, M.G., 1994 "Arabic Syntax Analysis and Generation", (In Arabic), Proceedings 2nd Computer Arabization Symposium, College of Computer and Information Science, King Saud Univ., Riyadh, Saudi Arabia, Vol. 2, 1994, pp. 1-14.

Pease 1996, ARAMED: Extension and integration of Arabic lingware components in a unification-based MT system for the field of medical terminology and classification, Co-ordinator: Universität des Saarlandes, Saarbrücken, Germany (Catherine Pease); 1996-1997.

Thalouth, B.1986, Al-Danan, A., "A Comprehensive Arabic Morphological Analyzer- Generator", Syria Proc. Of Arab Summer School, Syria, 1986.

Yaseen 1998, NAPLUS: natural arabic processing for language understanding systems ; inco ec project co-financed by DG XIII; Project No. 973133, Co-ordinator: Impetus Engineering, Athens, Greece (Nick Theophilopoulos); and Amman University, Amman, Jordan (Mustafa Yaseen), 1989-2001.

Yaseen, M.1990, Al-Fedaghi, S., "An Etymological Theory for Non-Diacritized Arabic Text", (In Arabic), Proceedings of the 12th National Computer Conference, Riyadh, King Saud University, Riyadh, Saudi Arabia, 1990, pp. 660-674.