مدينة الملك عبدالعزيز
للعلوم والتقنية KACST
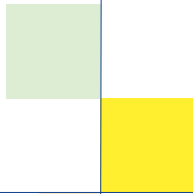
Computer Research Institute

Towards a Digital
Life

مدينة الملك عبدالعزيز
KACST للعلوم والتقنية

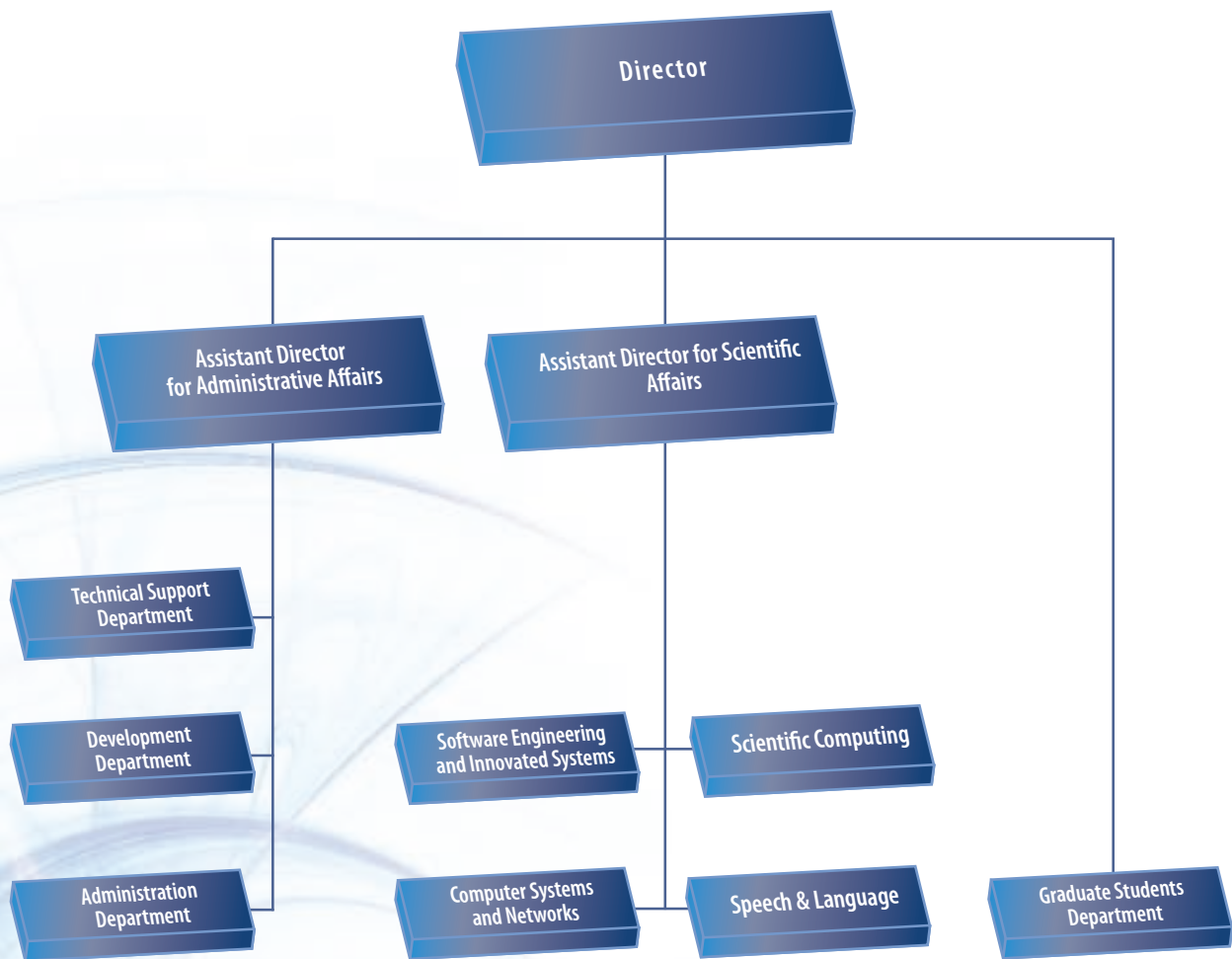**Computer Research**
Institute

Table of Contents

The Computer Research Institute (CRI) was established in 1992, with the aim of being one of the leading institutes serving the Kingdom for its science and technology development plans particularly in the area of Information Technology (IT).

The institute conducts high quality research leading to advanced technologies and innovative solutions. In addition to collaboration with international R&D centers, the institute offers consultation services, organizes scientific conferences. It also contributes to the training and development of Saudi HR through its postgraduate education scholarships.

Director

Assistant Director for Administrative Affairs

Assistant Director for Scientific Affairs

Technical Support Department

Development Department

Administration Department

Software Engineering and Innovated Systems

Scientific Computing

Computer Systems and Networks

Speech & Language

Graduate Students Department

Research Programs

CRI four main research programs are speech and language, scientific computing, computer systems and networks, and software engineering and innovated systems.

Each one of these programs covers a specific area of the IT R&D priorities outlined by the national plan which was developed jointly by CRI and stakeholders in the kingdom.

Speech and Language Program

The objective of this program is to conduct research and also develop software and databases that can be utilized for Arabic language processing.

Areas of interest include:

- Morphological analysis, generation & databases
- Search engines
- Language statistical analysis
- Language corpora
- Electronic dictionaries and glossaries
- Grammatical analysis
- Language understanding and translation
- Automated composition of Arabic texts
- Arabic diacritization
- Arabic phonetic databases
- Speech synthesizer
- Speaker identification
- Speech principles
- Speech recognition
- Speech signal processing

Scientific Computing Program

This program specializes in providing High Performance Computing (HPC) services, typically used in advanced scientific and engineering applications which require complex, long and high-speed computing such as simulation and modeling.

Computer Systems and Networks Program

This program conducts applied research studies including consultation on issues related to computer networks, information security, privacy, operating systems and database systems.

Software Engineering and Innovated Systems Program

This program specializes in the localization of computer applications, open source software and information systems analysis and design.

National IT Research and Development Plan

Information Technology (IT) is crucial to the Kingdom's development since it is one of the key drivers for productivity and economic growth in many world countries. In 2008, CRI developed jointly with stakeholders a comprehensive national plan to identify the key IT R&D priorities in the Kingdom. The plan was based on needs of IT users and stakeholders in the Kingdom that include government organizations, the private sectors and universities. The plan outlined the IT R&D vision, mission, goals and priorities in the Kingdom as follows:

Vision

The Kingdom of Saudi Arabia will be recognized as an innovator in IT, with local companies and universities that conduct leading edge research and produce world class graduates; a national laboratory system that serves government, universities and industry, and effective linkages between all elements of the system.

The Kingdom will be a regional leader across a broad range of IT areas and will be a world leader in areas of special importance to the Kingdom.

Mission

To enhance the Kingdom's scientific and technological competence in information technology and the application of information technology to the Kingdom's economic and social needs through a coordinated program of research, development, localization, and transfer of technology.

IT Goals

- To support an expanding and innovative KSA IT industry
- Advance IT capabilities to meet critical needs in the Kingdom, in areas such as computer networking, and security
- To develop innovative high quality IT applications to meet specialized needs in the Kingdom, such as the oil and gas industry
- To adapt and localize IT applications for e-services and e-business
- To contribute to the global open source software trend
- To develop world class capabilities in language technologies especially tailored to serve the Arabic language
- To improve scientific high performance computing in order to expand the Kingdom's capabilities in science and engineering through modeling, simulation, and visualization

IT R&D Priority Areas

The plan identified the following technology areas for IT R&D:

- Speech and Language (speech recognition technologies, search engines, etc.)
- High Performance Computing(simulation and modeling)
- Computer Systems and Networks (information security and operating systems)
- Software Engineering and Innovated Systems (localization of open source software and applications)

Local Cooperation

In terms of science and technology cooperation among relevant organizations and research institutes, KACST with its research institutes, participates in different national science committees and events. It also collaborates with technology incubators and innovation centers.

International Cooperation

In the global context, KACST collaborates with Gulf, Arab and world countries through joint committees. It also collaborates with a number of Arab and International scientific commissions and organizations. This allows KACST to exchange IT expertise with regional and international bodies.

مبادرة الملك عبدالله للمحتوى العربي

King Abdullah Initiative for Arabic Content

Information Technology (IT) nowadays plays an important role especially in developing countries. In addition, the extent of IT possession and proper usage has become an indicator of progress in these countries.

IT has changed the means of exchanging information, gaining knowledge and providing services.

Being the most common form of IT, the internet has become the means for people around the world to preserve their culture and enrich their knowledge by providing large digital content covering various subjects in the form of text and multimedia. The web digital content has recently become an underlying foundation for different activities and industries.

Although many languages have a large share of the web content, Arabic digital content remains limited. Its share is far less than that of the number of Arabic speakers. King Abdullah Initiative for Arabic Content was launched in light

of these facts. Its main objective is to expand the Arabic content on the web quantitatively and qualitatively.

The Initiative Projects

Implementation of the initiative has been assigned to KACST. It is currently working with other institutions within and outside the Kingdom to develop tools that help enrich the content such as dictionaries, search engines, morphological analyzers, speech recognition systems, and optical character recognition systems. KACST is also seeking to increase Arabic content itself by cooperating with different organizations and companies to create more Arabic content and to enhance its quality.

KACST has already started implementing some projects that include: a strategic and roadmap plans for the initiative, the translation of dozens of scientific books into Arabic, an Arabic interactive dictionary, Arabic open content and a multimedia center. Other activities have been also organized including content enrichment competition, symposiums, meetings, and workshops. KACST is working towards starting other projects with related partners since Arabic digital content is the responsibility of all Arabic speakers. KACST will take the lead and pave the road for any tools or methods that would ultimately enrich Arabic content so that other parties can carry on and fill in the missing elements.

Arabic TTS

A Text-to-Speech (TTS) system is an important element in any human-machine communication system. A TTS system is a complex system with many programs and algorithms set to produce a sound that can be heard, understood and as close as possible to human natural voice. During the last 4 decades, significant progress has been made in TTS technology of the European languages.

Only recently have efforts on Arabic TTS been initiated, and the resulting systems remain closed software with limited usage and capabilities. Jointly with King Fahd University for Petroleum and Minerals, KACST has completed the first open source Arabic TTS system. The system is called KACST Arabic Text-to-Speech (KATTS).

Contents:

KATTS contains:

- A system that can convert any diacritized Arabic text into speech. The quality of the output speech can be controlled in terms of:

- Fundamental frequency
- Speech tempo
- Allophonic method
- Diaphonic method

Beneficiaries:

KATTS can be licensed to:

- Companies with call centers or companies which deal with ATTS
- Research centers
- Speech departments at universities

Applications:

KATTS can be used in many applications including:

- Interactive voice response systems
- Communication centers
- Talking machines and tools
- Warning systems
- Information systems
- Services related to individuals with special needs

Arabic Language Morphological Analyzer

This project aims is to develop algorithms for the analysis of the Arabic language vocabulary according to morphological and grammatical properties. The algorithm is based on the morphological and grammatical characteristics extracted from a large linguistic corpus. It is developed according to the following steps:

- Comprehensive morphological and grammatical properties are defined to encode the characteristics of Arabic vocabularies

- A morphological and grammatical database is built and derived from a corpus of carefully selected articles
- Expert system is outlined to assist in encoding grammatical and morphological characteristics for Arabic vocabularies- the system will also benefit from experiences and self-learning
- Characteristics of vocabularies records are encoded
- Finally, a language morphological analyzer is developed

Morphological and linguistic characteristics:

The system specifies the following characteristics for an Arabic vocabulary entry:

Nouns

- The origin of the word (formal or informal language, Arabized or foreign)
- Defective and derivative
- Definite and indefinite articles
- Plural and singular
- Diminutive

Verbs

- Verbal noun
- Pattern
- Verb type
- Syntax and conjugation

Components:

This system consists of the following components:

- Text tokenizer tools
- Storage and retrieval tools of grammatical and morphological characteristics
- Morphological and grammatical characteristics entry system
- System teaching tools

Beneficiaries:

- Linguists
- Researchers in the field of language properties
- Developers of automated systems

Applications:

- Analyzing and encoding of Arabic language properties
- Grammatical and morphological analysis systems
- Recognition of the grammatical and morphological properties of the Arabic language
- Statistical linguistic systems

Morphological Characteristic for Arabic Words

The goal of this project is to create a database for the morphological characteristics of Arabic language lexis, and to produce the necessary morphological, syntactic and phonological algorithms associated with the computer systems generation of lexis. The database entries include lists of Arabic alphabet, word roots and their features. In addition, the entries include morphological measures used in generating Arabic lexis such as inert nouns and verbs, derivatives and infinitive forms of verbs and nouns. It also includes morphological features of generated lexis such as the plural and diminutive forms, relation and feminine cases, the tense and aspect of verbs and the nominal forms generated from infinitive and derivative nouns.

Importance of this research:

This research lays out the basis for Arabic language computerized processing applications.

These applications cover text generation, analysis, text compression and compilation, information retrieval, data encoding, machine translation, and Arabic language machine teaching and understanding.

Research methods:

The project employed linguists specialized in the morphology of the Arabic language, where codes are collected or created for data entries. Computer engineers assisted researchers in developing suitable algorithms and publication methods for the database. Computerized programs were used to facilitate the data entry process.

Results:

Linguists and computer engineers collected, revised and uploaded almost all morphological rules entries and encoded the necessary algorithms to generate an Arabic lexis according to morphological and grammatical rules of the Arabic language such as roots and morphological features.

Additions:

In this project, new algorithms were approved to facilitate the process of constant and regular generation of new lexis in Arabic language and to collect and compile rare and infrequent forms in small groups.
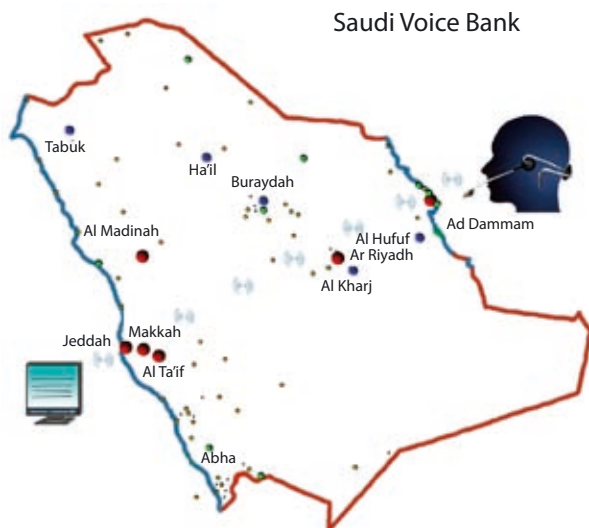
Saudi Voice Bank

Information technology is increasingly incorporating speech applications in almost all of our daily activities.In light of this global trend, and given the importance of this technology to Arabic speakers, KACST has been working on several projects related to this concept. One of them is the Saudi Voice Bank on which Arabic speech recognition systems are trained.

The Saudi Voice Bank contains speech data that is phonetically rich and balanced to empower speech recognition systems to recognize Arabic speech regardless of speaker's gender, dialect and age.

One of the users of the Saudi Voice Bank is IBM which signed an agreement with KACST in 2002 to use the Saudi Voice Bank in developing a telephony human machine communication system for Arabic.

Saudi Voice Bank

Contents:

The Saudi Voice Bank contains the following:
- More than 1000 Arabic speakers (50% males; 50% females)
- More than 60.000 wave files
- More than 60.000 Arabic text files
- More than 60.000 transcription files

Beneficiaries:

The Saudi Voice Bank is available at KACST and can be licensed to:

- Research Centers
- Companies that develop computer speech systems
- Researchers interested in speech recognition and speaker verification

Applications:

The Saudi Voice Bank can be used in speech recognition systems which have many applications including:

- Telephony translation
- E-learning and education
- Access to information and call centers
- Speaker verification, language and dialect identification call centers
- Banking and finance
- Flight reservations
- Telephone control of home appliances

## Prayer Time and the Islamic Calendar Software

This software is designed to calculate and determine prayer times all over the world, particularly in main cities. They also calculate the beginning and ending of the lunar months according to Umm Al Qura Calendar. Working with the Advanced Electronics Company, KACST produced the " Azzan Clocz " which is used in several mosques in Saudi Arabia.

Speech Recognition for Telephony Applications

The use of speech recognition systems over fixed-line and mobile networks has significantly increased lately. With speech technology, users can, for example, call airlines or travel agents and make reservations, or check flight details verbally. Although speech technology is being increasingly applied to a range of languages, little effort has been devoted to Arabic speech recognition.

KACST has recently worked with IBM to develop a speech recognition system for Saudi speakers. This system is based on the Saudi Voice Bank. KACST has also built and evaluated a speaker verification system using the Saudi Voice Bank.
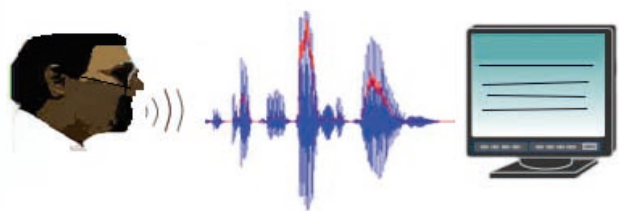
KACST is currently developing speech recognition platforms using Hidden Markov Model Toolkit (HTK) and Carnegie Mellon university (CMU) speech recognition tools. As part of our efforts in this area, we are investigating different techniques for language and pronunciation modeling. The goal is to have a robust speaker-independent speech recognition system of telephony applications for Saudi speakers.

Contents:
The Speech Recognition System contains the following:

- Acoustic models
- Language models
- Pronunciation dictionary

Beneficiaries:
Customer Services and call centers including:

- Airlines and travel agencies
- Banks
- Directory assistance

Arabic Phonetic Database
Arabic is the first language to have its sounds fullycategorized and described over 12 centuries ago. Unfortunately, this information is not suitable for modern information technology.

Therefore, KACST started in the mid 90's to build an Arabic phonetic database. By the turn of the century, the KACST Arabic Phonetic Database (KAPD) was already available for researchers and research centers. The most sophisticated equipments and tools were used to compile the database with data that can easily be utilized by researchers and parties interested in speech and phonetics.
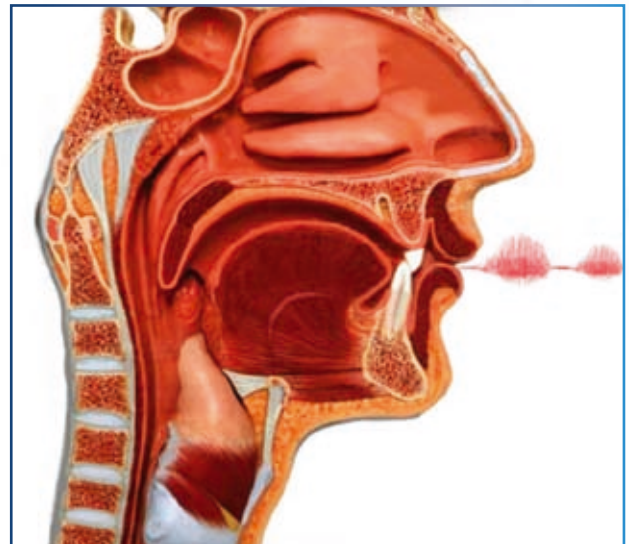
Contents:

KAPD contains the following:

- More than 12,000 wave files for all Arabic sounds in different word positions
- 7 subjects participated in KAPD experiments
- Aerodynamic data during the production of Arabic sounds
- Electropalatographic data
- Facial movements
- Images of the vocal folds, epiglottis and velopharyngeal port
- Degree of nasalance

Beneficiaries:

The following specialists can utilize KAPD:

- Speech therapists
- Researchers on speech synthesis and speech recognition
- Researchers on speaker identification and verification

Applications:

KAPD can be used in the following applications:

- Vocal tract modeling
- Speech recognition systems
- Speech synthesis systems
- Speech therapy
- Voiceprint

Arabic Diacritizer

Modern Arabic writing includes only letters representing consonants, which means that Arabic vowels and geminates are not represented in the daily writing of Arabic.

The absence of the vocalic and geminate symbols does not allow full usage of other computational systems such as text-to-speech and automatic speech recognition systems and search engines.

Therefore, CRI started conducting experiments on Automatic Arabic Diacritization to develop a system that can be integrated into other related computer systems. The result is KACST Arabic Diacritizer (KAD). All the components of KAD are innovated and solely owned by KACST.
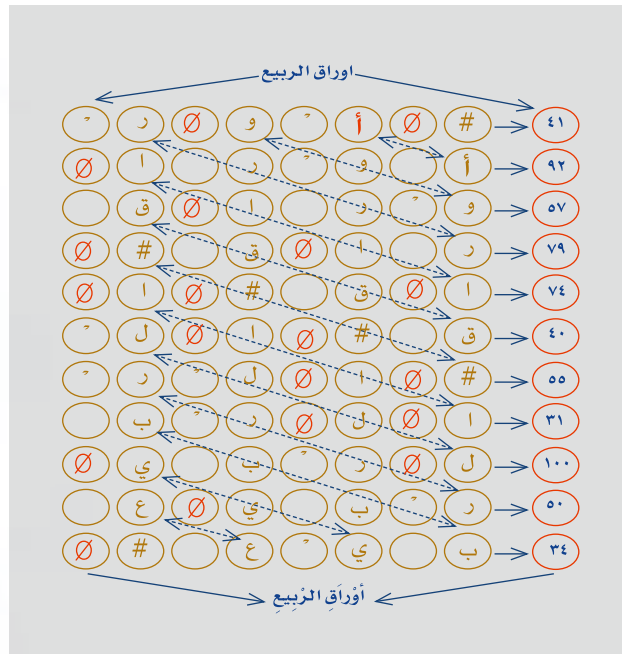
Contents:

KAD contains the following:

- Algorithms to calculate probability
- Algorithms to select the diacritic of the highest probability
- 68,378 quad-grams of the Arabic letters and diacritics

Specifications:

KAD technical specifications:

- Diacritization accuracy is 87% on the letter level including word-final letters
- Diacritization speed is more than 500 words/second
- KAD size is less than 3 MB

Beneficiaries:

KAD is available at CRI and can be used by:

- Research centers
- Educational institutions
- Translation centers

Applications:

KAD can be used in:

- Text-to-speech systems
- Automatic speech recognition systems
- Search engines
- Automatic translation systems

Arabic Names Romanization

The Arabic Names Romanization System is an automated system that converts a given name written in Arabic letters to the corresponding Latin letters. The system can handle complete and long Arabic names. The system does not require the name to be vowelized in order to romanize it, the process essentially depends on small databases with few transformation rules. It has the ability to analyze compound names, recognize all of their parts, romanize and finally merge all components to form a correct final Romanized name.
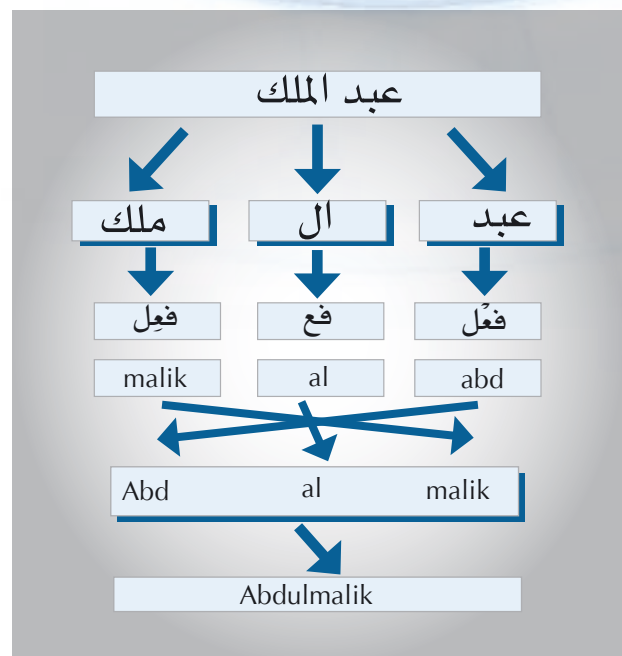
Contents:

The Romanization system consists of the following components:

- Databases with morphological patterns used for name generation with their frequencies
- Small databases of irregular names
- Algorithms analyzing compound names
- Algorithms for names Romanization, including all parts
- Algorithms for the merger and incorporation of the Romanized name parts

Beneficiaries:

Organizations that require registering Arabic names in the corresponding Roman letters such as:

- Security authorities
- Travel and tourism agencies
- Educational institutions
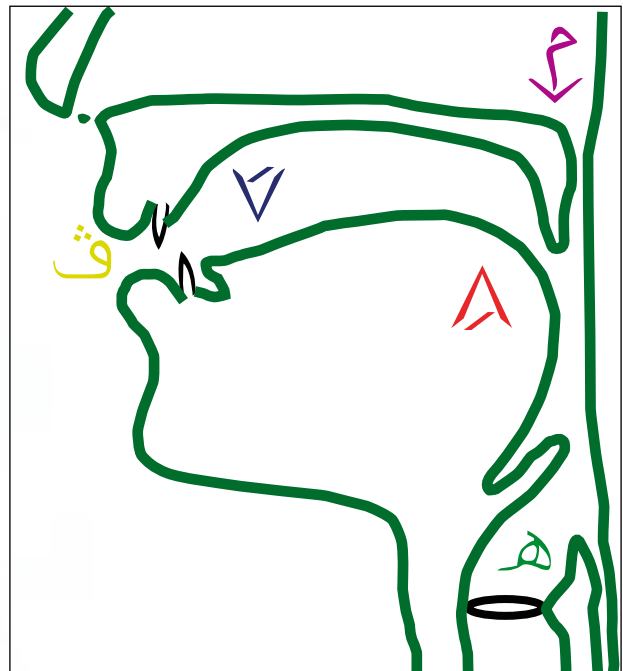- Hospitals
- Employment agencies

Applications:

- Systems requiring registration of Arabic names in Roman letters
- Machine translation systems
- Arabic language processing systems and applications

Arabic International Phonetic Alphabet

The Arabic alphabet can record almost all the sounds of Arabic, as is the case for other languages and their writing systems. It does not, however, possess symbols to record all the sounds of other languages, including certain sounds of the Arabic language itself which are part of either the standard or dialect inventories. Therefore, Arabic speakers find it impossible to transcribe speech sounds using Arabic alphabet. They often either describe the sound in words or use the International Phonetic Alphabet.

Thus, CRI designed a phonetic alphabet based on the Arabic alphabet to represent all speech sounds for all languages. The new phonetic alphabet is called Arabic International Phonetic Alphabet (AIPA). AIPA was further developed as 2 fonts that can be installed onto one's PC and used in different word processors.

Contents:

AIPA contains the following:

- 186 sound symbols
- The kacstaipa1.ttf. font
- The kacstaipa2.ttf. font

Beneficiaries:

AIPA is available at CRI and can be used by:

- Researchers and students of linguistics and languages
- Transcribers
- Researchers in automatic speech recognition and speech synthesis
- Speech therapists

Applications:

AIPA can be used in:

- Transcription
- Speech pathology assessment
- Automatic speech recognition and speech synthesis
- Language and linguistic studies

Arabic Stemmer

The system is used to analyze Arabic vocabulary and linguistically recognize and isolate suffixes, prefixes and infixes and then extract the corresponding stem (verbal noun). The system depends on regular expressions to serves the users and developers of applications that support Arabic language. It also helps in expanding the research process of Arabic source words when used with search systems, retrieval systems and search engines.

Beneficiaries:

- Developers of applications dealing with the Arabic language
- Linguists
- Researchers in the field of Arabic properties

Components:

The morphological analyzer consists of the following:

- Text segmentation tools
- Regular expressions analysis tools
- Group of regular expressions (more than 1400 expressions)

System Mechanism:

The system depends on the apparent similarity of Arabic words generated from different roots using the same pattern. A group of regular expressions can be formalized to express the generation process. Accordingly, the regular expression can then be used to analyze a given Arabic word to a verbal noun, suffixes and prefixes with the ability to isolate infixes.

Optical Recognition of Arabic Braille System

The Braille system has enabled the blind to read and write, a luxury they were deprived of before the invention of the Braille system. With the Braille system, blind people are able to acquire knowledge and communicate with others. The growing use of the Braille system, almost all over the world, has created a need to develop an automated system for Braille. Many languages have produced optical systems to recognize Braille documents such as (OBR) system. Nevertheless, an optical recognition system for Braille in Arabic has not been developed yet.
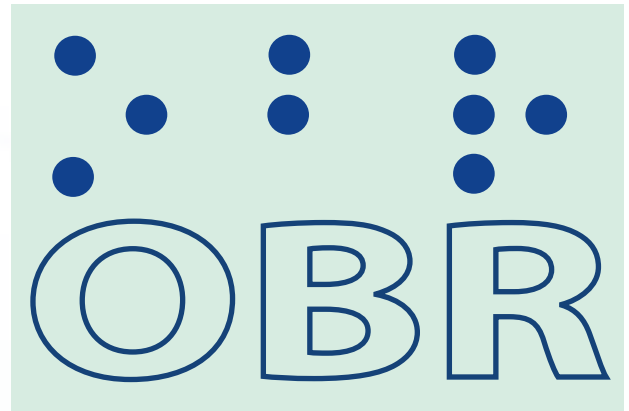
The optical recognition system is done by collecting documents produced in Braille. The documents are then scanned, processed and transferred into digital forms similar to Braille codes. Finally, the document can be transferred into an Arabic text that can be used in many fields.

| Applied Rule	Word	Prefix	Stem	Suffix
3.	على		على	
<ال>3.<ية>	النارية	ال	نار	ية
<ال>2.<أ>.<ا^>.<ة>	الأتربة	ال	تراب	
<>.<ت>.<و ا^>.	تموت		مات	
<ب>.<ه>	به	ب		ه
<وال>.<أ>.<ئ^ؤا>.<ة>.<>	والأفئدة	وال	فؤاد	
<>.<ا>.<أ^أ>.<ها>	أمالها		أمل	ها
<ال>4.<ية>	السريرية	ال	سرير	ية
<ولل>. 3.<ات^ة>	للمسرات	و	مسرة	ات
2.<^ى>.<ائئ><هم>	وزائهم		وزير	هم
<><ة>.<ا^>2.<أ>.<>	أجهزة		جهاز	
<ال>2.<و>	الرموز	ال	رمز	
5.<ب>	بزراعة	ب	زراعة	
6.<ة>	متكاملة		متكامل	ة
<وس>.<ي>.<و ا^>.	وسيقول	وس	قال	
<وال>2.<ى^ية><>	والرؤي	وال	رؤية	
2.<أ>.<^اء><ية><>	أوعية		وعاء	

CRI Projects and Products

CRI implemented an Optical Recognition of Arabic Braille system project to serve the blind.
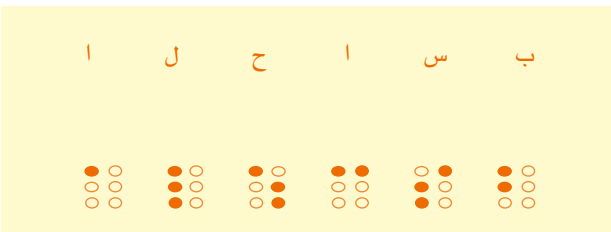
Braille System:

The Braille system was invented by Luis Braille in 1809. Luis became blind in the early years of his life. The system spread all over the world due to its ease for reading and writing, as it mainly depends on the touch sense. The system is made up of a number of cells; each cell consisting of six dots in three rows and two columns as shown below.
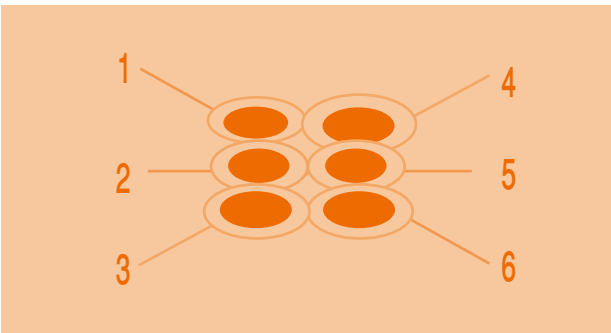
Thus, 63 cells could be built. The Braille system reads all languages from left to right.
An example of writing the word 'computer' in Braille

Components:

The Braille Arabic system includes the following:

- A scanner for collecting images of texts
- Algorithms for the processing of Braille images and their transfer into digital forms according to the existing codes
- Algorithms to translate Braille codes into readable Arabic text

Applications:

The Braille system helps the blind and other people to:

- Save electronic copies of any document produced in Braille
- Reprint old copies produced by the Braille system
- Help people communicate with the blind without a translator specialized in the translation of Braille texts

Open Source Software

Many countries, organizations and large companies have adopted the open source programs policy. 'Open source program' is the terminology used to refer to open source software, that is subject to GPL license conditions or similar licensing conditions. These licenses provide the user with several advantages such as:

- The high quality of open source software due to source code revising done by other programs
- Rapid application development
- High security measures and rapid tackling of shortcomings
- Learning, copying and distribution freedom
- Low cost
- Expanding knowledge

- A convenient and positive competition environment
- Availability of other sources for programs

Achievements:

In 2000, King Abdulaziz City for Science and Technology completed its first project on open source software programs. The project delivered an electronic mail program and Arabic texts editor. KACST then arabized other open source software programs such as open Office, Fedora and KDE. KACST now also provides Linux users with Arabic open font. Currently, KACST is working on the development of an Arabic spelling checker program for (OpenOffice) in addition to an Arabic Math editor.

Beneficiaries:

Open sources software programs meet the requirements of public institutions, individuals, specialists and developers of the following programs:

البرامج
مفتوحة المصدر

php MyAdmin

fedora

OpenOffice.org

|	
Operating Systems	Linux,OpenBSD , FreeDOS , OpenBeOS
Another Systems	PHP , MySQL , BIND , netatalk
Server Applications	XFree86 , Apache, samba, mailman, postfix, sendmail
Users Applications	Gnome , KDE , FireFox , OpenOffice

Arabic Name Translator

Today, people travel around the world more than they did at any other time. Thus, it is necessary that their names be written on their travel documents in the Roman alphabet. Proper names are written in the Roman alphabet on passports, airline tickets, credit cards, driving licenses and certificates.

Transliteration of Arabic names has not been consistent because Arabic orthography is different from that of the Roman alphabet. The result is that the same name is romanized in different ways, where such inconsistencies can have negative effects in terms of security and property rights.

To standardize the method of Romanizing Arabic names and make it available to others, CRI has developed KACST Arabic Name Romanizer (KANR).
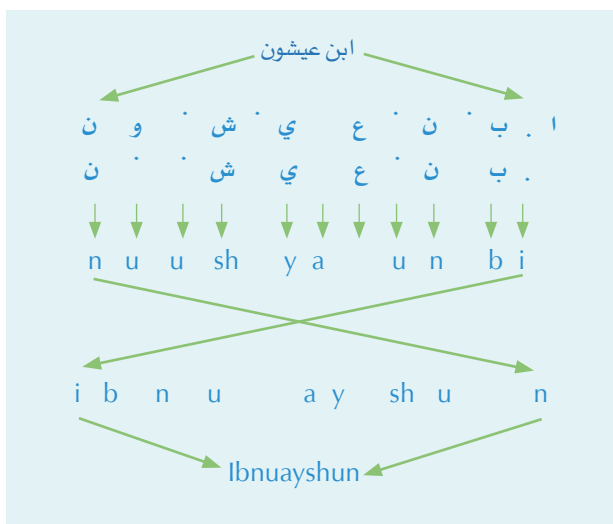
Contents:

KANR contains the following:

- Algorithms to process Arabic letters
- Algorithms to transliterate the Arabic letters into the Roman alphabet
- Algorithms to process Roman letters
- More than 50,000 Arabic names written in the Arabic orthography

Beneficiaries:

KANR is available at CRI and can be used by:

- Passports agencies
- Banks
- Universities
- Hospitals
- Travel agents and airlines
- Motor Vehicle Offices (driver licenses)

Applications:

KANR can be used in:

- Translation and automatic translation systems
- Transliteration of Arabic names in official documents and identification cards

Acoustic Detection of the Red Date Palm Weevil

The Red Date Palm Weevil (Rhynchophorus ferrugineus is responsible for one of the most serious infections effecting date palm trees in the Kingdom. CRI is working on this serious problem, given the importance of early stage detection of hidden insect infestations of date palm trees (Phoenix dactylifera L.). In particular, an acoustic instrument has been developed for the early detection of the Red Date Palm Weevil (RDPW). This type of insect attacks date palm trees and causes irreversible damages at late stages where the infected tree has to be ultimately destroyed.

So far the acoustic system was tested and evaluated in terms of its early discovery ability of hidden pests inside the palm tree. Signal acquisition was completed for a number of date palms, and then signal processing techniques were evaluated in terms of their ability to provide an estimate that can be visually used to recognize the acoustic signature of the RDPW.

The testing instrument was tested in the laboratory first; then it was used on suspected or infested trees in the field. The final results indicate that the acoustic monitoring approach along with signal processing techniques are very promising for the early detection of the larva as well as the adult pest in the date palms.

Contents:
- Study the characteristics of RDPW acoustics signature
- Understanding noise and interferences
- Study the characteristics of acoustics associated with RDPW colonies
- Determining the impact of location, direction and specifications of recording instruments on the recorded signals
- Building reliable algorithms to detect acoustic signals inside palm tree in noisy environments
- Developing an instrument for RDPW detection - DBEEB

Beneficiaries:
Some of the parties that can benefit from this innovation include:

- The Ministry of Agriculture
- Farmers
- Researchers in the following areas: agriculture, entomology and signal processing

Applications:
The RDPW detection instrument can be used in:
- Controlling the spread and economic damage resulting from this type of infestation.
- Monitoring as well as early detection of the RDPW in order to take appropriate measures such as isolating or treating the infected trees.

Arabic Numerical Analyzer

The Arabic Numerical Analyzer is a series of algorithms that analyze and transfer numbers from numerical forms into words, following correct syntactic and morphological rules of the Arabic language. This program is a useful tool for banks, retail stores and academic institutions as it saves time and guarantees accuracy.

Research methods:

The project basically compiled and revised all written Arabic numerical forms in different grammatical cases; nominative, accusative, dative or genitive cases. Singular, dual and plural cases are also considered. Several algorithms were developed for analyzing numbers from zero to hundreds of billions. Sentences were formed from data saved in the Arabic language numerical database.

Results:

The system provides: algorithms for generating and writing Arabic numbers in different grammatical cases, a database of Arabic numerical forms, software and a user interface.

Beneficiaries:

This program is practical for individuals, private and public institutions. It is helpful for people working in commerce, banking, accounting and auditing offices, as well as schools and universities. The system helps accountants, cashiers and engineers.

Applications:

The system can be used in all programs requiring writing numbers such as bills, checks, accounting records, report cards and other types of documents.

Arabic Text Recognition Using HMM

The Arabic text recognition is one of the projects currently supported by the institute, which supports different applications of the Arabic language. The main objective of the project is to recognize a scanned printed Arabic text and produce a corresponding editable text.

This project will present a cursive Arabic script recognition system. After acquiring the document, the system decomposes the document image into text line images and divides each text line image into smaller overlapped frames. The system extracts a set of simple statistical features from each frame and then injects the sequence of the feature vectors to the Hidden Markov Model Toolkit (HTK). HTK is a portable toolkit for speech recognition system. The proposed system is applied to sophisticated cursive Arabic fonts, which requires building a corpus of omni-font documents.

Implementation Objectives:
- Building Arabic corpus including multiple computer generated fonts and typeset text (books, reports, newspapers, magazines, conferences minutes, etc.)
- Building an Arabic OCR System

Application:
- Arabic text recognition
- Producing electronic versions of books and documents
- Supporting the Arabic content over the internet
- Supporting the Arabic e-learning applications

Automatic Classification of Arabic text

This research objective is to study the various stages for the classification of Arabic text starting from the collection of several texts representing various fields, the investigation and selection of the linguistic features chosen to represent the texts. Texts are first chosen on a linguistic basis. The second stage is to study the application of several methods of text classification based on machine learning or data mining techniques. The performance of methods used in the classification of both general and specialized texts are studied. Finally, the main aim of the project is to master these techniques and exploit them to serve Arabic language.

Principle objective:
- Develop algorithms for Arabic texts machine classification
- Construct a corpus database for the Arabic text machine classification system

Applications:
- Classify various Arabic texts
- Determine harmful e-mail messages
- Classify the contents of internet pages
- Machine directing of messages

New Projects

CRI also supports and conducts a number of new projects which are aligned with the IT R&D priorities of the Kingdom. These projects have been implemented either by CRI or in collaboration with local and international research centers. Some of these projects include:

Arabic IVR System (interactive voice recognition systems)
Sound Source Recognition
Machine Translation (from and to the Arabic Language)
Arabic Syntactic Parser
Open Source Program
Multimedia Lab
High Performance Computing Lab
Online Arabic Handwritten OCR
2D Barcode & Steganography
Trusted Platform Module (Information Security Systems)
Computer Forensics
Language Identification
Several Micro IT Projects
Wireless Sensor Networks
Biometrics

Computer Research Institute

King Abdulaziz City for Science and Technology

PO BOX 6086 RIYADH 11442

Kingdom of Saudi Arabia

Tel.: +966-1-481-3765

FAX: +966-1-481-3764

Email: ceri@kacst.edu.sa

Visit our webpage at: http://ceri.kacst.edu.sa

www.dakkin.com