5.0 Software Quality Engineering in Indian Language Products

1.0 Introduction

In order to address the issues of quality in the developed language technology products, DIT initiated a project entitled- Software Quality Engineering in Indian Language Products. The objective of this project is "To facilitate the process of integrating Quality Engineering practices in the Indian Language Technology Products".

This report describes a general framework for language product evaluation. It details the various activities undertaken by the project in testing and standardisation of the Indian language products. A study of a similar initiative taken in Europe called the EAGLES (Expert Advisory Group on Language Engineering Standards) is briefly described to compare with other efforts in the area of Language technology evaluation.

2.0 A Framework for Language Engineering product evaluation:

The starting point for any evaluation is the International Standard ISO9126. This is primarily concerned with the definition of the quality characteristics to be used in the evaluation of software products. The Standard describes a hierarchy of attributes, sub-attributes and their metrics whereby a software product can be evaluated. The first level consists of defining a set of attributes for a software product. These are:

- Functionality:

  A set of attributes that bear on the existence of a set of functions and their specified properties. The functions are those that satisfy stated or implied needs. The sub-characteristics of functionality are suitability, accuracy, interoperability, security and functional compliance.

- Usability:

  A set of attributes that bear on the effort needed for use, and on the individual assessment of such use, by a stated or implied set of users. The sub-characteristics of usability are understandability, learn ability, operability, attractiveness and usability compliance

- Efficiency:

  A set of attributes that bear on the relationship between the level of performance of the software and the amount of resources used, under stated conditions.

- Reliability:

  A set of attributes that bear on the capability of software to maintain its level of performance under stated conditions for a stated period of time.

- Maintainability :

  A set of attributes that bear on the effort needed to make specified modifications.

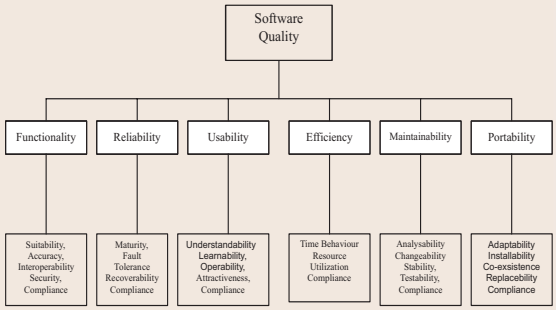
- Portability :

A set of attributes that bear on the ability of software to be transferred from one environment to another.

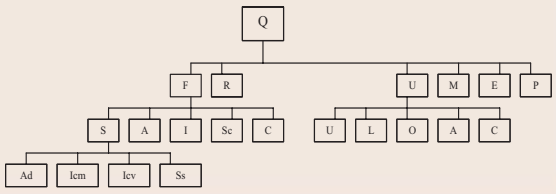
For the evaluation of software products one more characteristic (not in the Standard ISO9126) has been included: Documentation quality. This has no sub-characteristics but can be evaluated using a checklist that scores the product on the quality and completeness of its specification document and user manual.

Placed below is the model showing characteristics and sub-characteristics.

Software Quality Characteristics
(ISO 9126-1)



Quality Model



Q = fn (F,R,U,E,M,P)
Q = wfF+wrR+wuU+weE+wmM+wpP
F = fn (S,A,I,Se,C)
F = w'Ss+w'aA+w'iI+w'scSe+w'cC
S = fn (Ad+Icm+Icv+Ss)

2.1 Steps in Language product evaluation :

Step 0: Software testing as per ISO 12119:

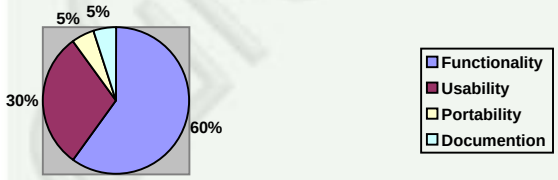
The products are required to be tested as per ISO 12119. This is essentially black box testing. The software requirements Specifications, user manual, product description (wrapper) and program and data (if any) are to be submitted for testing.

Step 1 : Assessment of required quality characteristics:

In case of Language engineering products the first two characteristics ie functionality and usability are of prime importance. The richer the feature set of a product, the higher it scores on functionality. Usability is important,, as this will encourage the use of the products in the Indian language. The characteristic portability is next in importance as many technologies have to be portable on Windows and Linux. Documentation quality is important for any product more so for language products, where a number of complex mathematical algorithms are used.

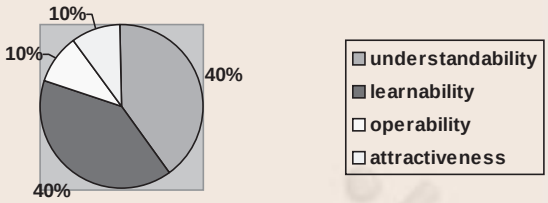
Step 2 : Weight distribution

Suitable weights are attributed to the characteristics and sub-characteristics. A sample model is given below with suitable weights for the characteristics:

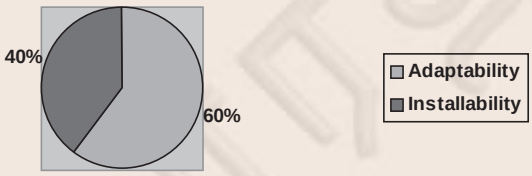


The sub-characteristics for functionality could be weighted as shown in the chart below:

The sub-characteristics for usability could be weighted as shown in the chart below:



The sub-characteristics for portability could be weighted as shown in the chart below:



The above weights for the characteristics and sub characteristics have been drawn from the experience of the evaluators but could be modified by general consensus amongst linguistic experts.

Step 3: Establish compliance level of metrics:

After defining the quality model at characteristic level and sub-characteristic level the corresponding metrics and their level of required compliance are established (could be 75%, 50%, 25% or High/Medium/Low).

Step 4: Measurement Plan:

A measurement plan is generated for gathering the metrics and the actual compliance is established. The measures include data such as number of functions evaluated, size of software, number of functions correctly implemented, number of functions that have anomalies etc.

Step 5: Product evaluation:

- Each sub-characteristic is evaluated as the ratio of number of metrics that are greater than or equal to the desired compliance level to the number of applicable metrics.
- Each characteristic is evaluated as the weighted sum of sub-characteristics.
- The software product quality Q is evaluated as a weighted sum of characteristics.

- The quantitative value obtained above is converted to a scale where the product is described as Excellent (Q> 90%), Good (89<=Q<=75%), Fair (74<=Q<=65%) and Q< 65% as unsatisfactory.

3.0 Present Status of Indian Language Product Evaluation:

In order to bring the technologies developed under the TDIL program to the masses, these have to be tested for functionality, usability and portability as many of these technologies are developed both for Windows and LINUX environments.

The products developed have been classified as follows:

I. Office aids (Text editors, spell checkers, presentation software etc)

II. Optical character recognition software (OCRs)

III. Text to Speech (TTS)

IV. Machine Aided Translation (MAT)

The details of products tested under each category, a generic specification that could be considered as a benchmark is included in the Appendix for each category of product.

3.1 Office Aids

A number of RCILTS have been involved in the development of Office Aids in the regional languages. Office Aids include text editors, spell checkers and other such software. Other RCILTS have developed products that address one or more of the aspects of Office Aids.

Briefly let us discuss the quality attributes desired in software such as Office Aids.

- Functionality:

  Functionality is the most important attribute. The more features the software has the higher the user satisfaction. The instructions to the user must be as simple as possible and help must be extensive. In cases of 'Wrapped' software or otherwise, the seamless integration of a suitable keyboard is important. One cannot expect users to go through elaborate combination of keys to write a letter. A floating Inscript keyboard is a must for the language addressed. The transliteration keyboard is an interesting feature but not in line with our target users. A native user of a text editor will find the output of transliteration keyboard at variance with familiar spellings of the words. While testing an Office aid the feature wise comparison with a benchmark like MS Office would be useful. Each feature not implemented must be justified on the basis of how such a requirement will not find usage in the Indian languages. Example, change of case: in all Indian languages, there is no concept of upper case and lower case, so also the use of subscript and superscript.

- Usability :

  The importance of usability cannot be over emphasized. The user interface should have a form that makes the system easy and efficient to use eg: dialogue and windows based. Before a technology can be released in the public domains extensive usability testing is a must.

  The Appendix I gives a sample generic specifications for a text editor.

  The following products have been tested or are in the process of testing:

Akshara (Advanced Multilingual text editor)

This package is developed by University of Hyderabad, Hyderabad. It is a powerful Advanced Multilingual Text Processor that works on various platforms like Windows and Linux, and is based on the ISCII Standard. It supports several Indian Languages like Gujarati, Hindi, Kannada, Malayalam, Marathi, Oriya, Punjabi, Sanskrit, Tamil and Telugu. It facilitates typing in Indian languages through two keyboard layouts: INSCRIPT and Phonetic. It has Email send and receive facilities.

Oriya Word Processor (Trilingual)

Oriya Word Processor is developed by Utkal University, Bhubaneswar on Windows platform.

It supports phonetic typing. Oriya spell Checker and Grammar Checker has been integrated with it. It supports English, Hindi, and Oriya. It has help facility.

Open Office (Kannada)

This package consists of a suite of writer's aids: text editor, drawing tool (paint), presentation software (power point), spread sheet (Excel) .It works on Windows XP.

Office Suite (Tamil)

This consists of a suite of products consisting of Tamil text editor (Palagai), presentation tool (Arangam), Spell checker (Annam), Tamil Database (Thenkootu), Document Visualization (Chitrapavvai). The text editor is not integrated with the text editor but a separate facility.

3.2 Optical Character Recognition (OCR) Software

The development of Optical Character Recognition (OCR) Software is one of the important successes of the TDIL project. A number of OCR's have been developed in the different Indian languages. The development of OCRs in native languages will in the development of a number of applications such as content creation in local languages from books and newspapers, database population from forms etc. In all cases of testing, text from newspapers, books and laser printed output were used as test inputs to be scanned and then presented to the OCR for converting to computer editable text. In the context of testing both accuracy (character level, word level and sentence level) and speed of conversion would be important. Next, the ability to take almost any kind of input format and convert to text format of a wide variety would be desirable. In the context of evaluation, functionality and usability would be most important.

The OCRs tested by STQC are now elaborated:

OCR Bangla:

Developed by ISI, Kolkota this OCR works in DOS. It can take character size 14-36 pts. The accuracy is between 30-96% depending on type of inputs. Its speed (CPS) is 13-47. The skew angle correction is +/- 5 degree. Its input format is TIFF and its output file format is PC. It supports single column. It supports bold characters but not symbols. The resolution supported is 300 dpi.

OCR Gurmukhi :

Developed by TIET, Patiala, this OCR works on WIN 9X. It can take character size 12-20 pts. The accuracy is between 93-99% depending on type of inputs. Its speed (CPS) is 12-59. The skew angle correction is +/- 5 degree. Its input file format is BMP and the output file format is Punjabi font in Word. The column support is single column. It supports bold characters but not special symbols. The resolution supported is 300 dpi.

OCR Tamil:

This OCR is developed by IISC, Bangalore and works on WIN 95/98. It can take character size 12-18 pts. The accuracy is between 62-98% depending on type of inputs. Its speed (CPS) is 1-19. The skew angle correction is +/- 10 degree. Input file format is BMP, PGM and its output file format is TAB, RTF. The column support is single column. It supports bold characters and special symbols. The resolution supported is 300 dpi.

OCR Devnagari:

This OCR developed by ISI Kolkota, works on WIN 95 (Min)/Linux. It can take character size 10-24 pts. The accuracy is between 48-97% depending on type of inputs. Its speed (CPS) is 20-44. The skew angle correction is +/- 5 degree. Its input file format is TIFF and the output file format is pc. The column support is single. It recognizes bold characters but not special symbols.It supports resolution greater than or equal to 300 dpi.

OCR Hindi/ Marathi:

Developed by CDAC Pune this OCR works on WIN 95X/NT/2000. It can take character size 10-36pts. The accuracy is between 85-98% depending on type of inputs. Its speed (CPS) is 93-368. The skew angle correction is +/- 5 degree. Its input file format is BMP, JPG, TIFF and its output file format is RTF/UNICODE/ISCII. The column support is Multi column. It recognizes bold characters but not

special symbols. It supports a resolution of 270-320 dpi.

OCR Telugu:

Developed by DCIS, University of Hyderabad this OCR works on Linux. It can take character size 12-20pts. The accuracy is between 84-87% depending on type of inputs. Its speed (CPS) is 23-29. Its input file format is PGM. Its output file format is ACI. The column support is single column. It supports a resolution of 300 dpi.

OCR Hindi:

Developed by CDAC Noida, this OCR works on Win 98/2000. It can take character size 12-36 pts. The accuracy is between 5-96% depending on type of inputs. Its speed (CPS) is 9-188. The skew angle correction is +/-5 Degree. Its input file format is TIFF, BMP and output file format is RTF. The column support is Single column. It recognizes bold characters but not special symbols. The resolution supported is 280-320 dpi.

OCR Oriya :

This package developed by Utkal University, Bhubaneswar works on Windows. The input is BMP format and the output is in TXT format. The accuracy is between 74-86% depending on type of inputs It supports the fonts OR-Sarala, OR-TTMukta, TT-Sarala & Dharitrismall. It supports single column text and resolution of 300dpi.

OCR Malayalam :

This OCR is developed by CDAC, Thiruvanthapuram. It works on windows and recognizes characters in the new script of Malayalam language along with the compound characters from the old script. It supports BMP, TIFF, JPEG and GIF as input formats and creates editable text in RTF, HTML, ACI and TXT output formats. The accuracy is between 93-97% depending on type of inputs It has facilities to process multiple font sizes, formats and multicolumn layouts containing both text and images and supports UNICODE, ISCII and ISFOC format. It has a skew correction of +/-5 degrees. It supports resolution of 300 dpi.

3.3 Text to Speech

Text to Speech software has been developed by a few RCILTs. This is an important area with significant social relevance as in the education of the visually impaired. The important consideration in case of TTS is intelligibility and naturalness, i.e a sentence must be clearly heard and with the right intonations and modulations. The main application areas of TTS are in Interactive Voice Response Systems, reading system for speech and sight impaired persons, talking books and toys, vocal monitoring, man-machine communications, fundamental and applied research etc. Blind people widely benefit from TTS System when coupled with an OCR which gives them access to written information.

It also posses a peculiar feature which makes them wonderful tools for linguistics, they are completely under control so that repeated experiences provide identical results as is hardly the case with human beings. The TTS tested have got the intelligibility to a great extent but need to go a long way to be considered as true TTS and not merely as announcement systems. A generic specification is given in Appendix III for the TTS. A few of the TTS tested are briefly described below:

Malayalam TTS:

Subhashini is a Text-to Speech System for the Malayalam language. It intelligibly reads out Malayalam text input. It is integrated with a text editor and is capable of handling popular Indian font codes and standard formats like ISCII and UNICODE. It consists of an in-built keyboard driver conforming to the INSCRIPT keyboard layout. It reads out the data entered through a keyboard or from a text file intelligibly. It also does automatic reading of printed material scanned from books, magazines, newspaper etc in combination with an OCR. It has got automatic handling of numerals and standard abbreviations and it works on Windows98/NT/2000 platforms.

Oriya TTS

Oriya TTS16 is text to speech software developed by Utkal University, Bhubaneswar. It supports document

in .aci extension as input files. It generates the output sound file in .wav format. It has previously created wave file database. It reads out the data entered through a keyboard or from a text file intelligibly. It also does automatic reading of printed material scanned from books, magazines, newspaper etc in combination with an OCR. The software works on Windows 2000.

Tamil TTS

Ethiroli is a Tamil Text to Speech software. It works in Windows 2000. It produces sound of words intelligibly.

3.4 Machine Aided Translation

Machine aided Translation is an important area and has immense potential and requirement in our country. Due to multiplicity and complexity of different languages in our country the MAT becomes an important area. There being eighteen languages in the country, translation from one language to another would yield a large number of pairs. Keeping in view the maximum correspondence in the language pairs, and keeping in view the need for translation of official correspondence from English to Hindi, this pair has been identified as the priority area for MAT system. Because of similarity among Indian languages, the translation among Indian languages is easier than translation from English to Hindi.

Internationally it is a well accepted fact that it is impossible to develop General Purpose Machine Translation Systems however it is practical to achieve reasonable degree of success in development of Domain Specific Machine Translation Support Systems which will largely aid the translators in doing their job faster. Indian Languages are phonetic in nature and are close to each other whereas English stands apart from the linguistic considerations.

For the evaluation of MAT system we have identified some of the major parameters. The Intelligibility of translation in different domains, accuracy and speed are some of the important parameters. A traditional way of assessing the quality of translation is to assign scores to output sentences. A common aspect to score for is Intelligibility, where the intelligibility of a translated sentence is affected by grammatical errors, mistranslations and untranslated words. It is important to check whether the meaning of the source language sentence is preserved in the translation. This property is called Accuracy or Fidelity. To measure engine performance the speed of translation can be measured.

English to Hindi MAT :

The Angla Hindi MAT tested by STQC is developed by IIT, Kanpur. The accuracy of translation at word level is 88 to 93% and at sentence level 74 to 80%. The intelligibility is on around 60% and speed is 1 word/sec.

4.0 EAGLES Initiative:

The EAGLES (Expert Advisory Group on Language Engineering Standards) was initiated in the early 90s, in Europe. The work of the group was entirely concerned with natural language processing systems and products. The long-term objective of this group is to identify and specify the components of a compendium of evaluation criteria and associated techniques, together with guidelines for their use. An individual evaluator can select those techniques that are relevant for his purpose.

This initiative is based on the International Standard ISO 9126. In this manner the Group's recommendation was an attempt not to incline towards any particular user/ customer but look at a general evaluation methodology that can cater to a wide variety of different classes of users and can be tailored to reflect specific requirements. The group selected three types of systems: writer's aids, translator's aids and knowledge management systems for their detailed study. An interim report EAGLES Document EAG EWG PR.2 is available at their site. This report first attempts to recapitulate the ISO 9126 Standard. Next, an attempt to construct a parameterized test bed is attempted. Here the evaluation begins with the description of the object of evaluation, description of classes of users and the description of the attributes of systems potentially of interest to classes of users

along with the metrics that would provide value for that product for each attribute. This test bed can be used while evaluating a particular product on behalf of a particular class of users.

The fundamental basis for evaluation is:

O X U -> V

Where

O is a set of objects of evaluation

U is a set of users interested in O

V is a set of values representing the utility that drives any evaluation

As can be seen from the above equation there is an emphasis on users. There cannot be a generalized product without the user category spelt out. For example the user of a text editor for typing letters will have different requirements (read expectations) as compared to a content creator.

The equation can be looked at in two ways obtaining two perspectives on evaluation:

1.0 O -> (U -> V) meaning given some object, evaluation tells us who likes it

2.0 U -> (O -> V) meaning given some user, evaluation tells us what they like

The report elaborates on this theory and covers in some detail the evaluation of Writer's aids and translator's aids. A case study of a European Translation service is described.

5.0 Conclusions:

The area of Natural Language processing is vast and not very much explored. The efforts in the present project have been directed towards testing and documentation of the various technologies that have been developed in the TDIL project. The stage of evaluation will come after the users of the systems and products have been identified and the attributes of the systems have been established. An attempt has been made to prepare generic specifications for four categories of Indian language products. The specifications reflect the level of usage and the desired functionality and performance in each case. In some cases a definitive value of a parameter has not been established.

Appendix I

Generic Specifications of Text Editor

1.0 Title: Generic Specification for text editor

2.0 Scope: This specification covers the following family of products:

S No.	Text Editor Level	Application
1.0	3	Government notifications, multi column document preparation for language newspapers, magazines, multi column printed text books with inlay for tables, pictures
2.0	2	Has some application between those at 1.0 and 3.0
3.0	1	Letters, report preparation, routine documentation, for Facsimile transmission

3.0 Terminology& Abbreviations:

TIFF is an acronym for Tag(ged) Image File Format. It is one of the most popular and flexible of the current public domain raster file formats.

PICT is a file format that was developed by Apple Computer in 1984 as the native format for Macintosh graphics. The PICT format is a meta-format that can be used for both bitmap images and vector images.

BMP files are commonly used file format for the Windows operating system. BMP images can range from black and white (1 byte per pixel) up to 24 bit colour (16.7 million colours).

JPEG (pronounced "jay-peg") is a standardized image compression mechanism. JPEG stands for Joint Photographic Experts Group, the original name of the committee that wrote the standard. JPEG is designed for compressing either full-color or gray-scale images of natural, real-world scenes. It works well on photographs, naturalistic artwork, and similar material; not so well on lettering, simple cartoons, or line drawings.

GIF-The Graphics Interchange Format defines a protocol intended for the on-line transmission and interchange of raster graphic data in a way that is independent of the hardware used in their creation or display.

UNICODE-Unicode provides a unique number for every character, no matter at the platform, no matter what the program, no matter what the language. This is the Standard for NLPs.

ISCII - Indian Script Code for Information Interchange is Standard

4.0 Features:

S. No.	Parameter	Level 3 requirement	Level 1 requirement
1.0	Documents category	Newspaper, magazines and printed books, Fax, laser printed documents, forms	Letters, laser printed documents
2.0	Color document support	Yes	Nil
3.0	Create, edit and save a document in various storage devices such as hard disks, floppies or other media	Yes	Yes
4.0	Delete, move and search for files	Yes	Yes
5.0	Allow definition of page sizes and margins	Yes	Nil

6.0	Column support	Multiple	Single
7.0	Cut, paste, duplicate section of text from one place in a document to another	Yes	Yes
8.0	Text search and replace word string facility	Yes	Yes
9.0	Embed tables, graphs and charts in document	Yes	Nil
10.0	Image format supported to embed illustrations	BMP, TIFF, JPEG	Nil
11.0	Automatic document formatting for bullets and numbering	Yes	Yes
12.0	Allow placement of headers and footers	Yes	Yes
13.0	Allow automatic page numbering	Yes	Yes
14.0	Allow representation of short cut keystrokes that can represent text or commands by definition of macros	Yes	Nil
15.0	Font Size support	8-72	8-12
16.0	Font Support	Italics, Bold, Normal	Normal, Bold
17.0	Built in spell checker	Yes	Yes
18.0	Word wrap	Yes	Yes
19.0	Automatic grammar check with suggestions	Yes	Nil
20.0	Built in thesaurus	Yes	Nil
21.0	Output File Format	ISCII, Unicode, Doc, PDF, HTML, XLS etc.	ISCII, Unicode, HTML
22.0	Size of Paper	Letter, A4, A5	A4
23.0	Allow document to be printed on a printer	Laser printer, Dot matrix printer	Laser printer, Dot Matrix printer
24.0	WYSIWYG- document appears on display screen exactly as it will look when printed	Yes	Nil
25.0	Operating System	Linux & Windows	Any one

5.0 Ease of Use

S No.	Parameter	Level 3 Requirement	Level 1 Requirement
1	Menu Navigation	Simple and usable	Simple and usable
2.	Help and documentation	Yes	Yes
3.	Online help	Yes	Nil
4.	Wizards (creating letters, memo etc)	Yes	Nil

Appendix II

Generic Specifications of Optical Character Recognition (OCR) Software

1.0 Title: Generic Specification of optical Character Recognition software.

2.0 Scope: This specification covers the following family of products:

S No.	OCR Level	Application
1.0	3	News archiving from newspaper, web content development from multi column printed documents, form scanning
2.0	2	Has some application between those at 1.0 and 3.0
3.0	1	Content creation from simple documents, text book content creation, archiving letters and documents from single column printed material

3.0 Terminology& Abbreviations:

OCR- Optical Character Recognition OCR is an omni font optical text recognition system that recognizes texts set in practically any font.

TIFF is an acronym for Tag(ged) Image File Format. It is one of the most popular and flexible of the current public domain raster file formats.

PICT is a file format that was developed by Apple Computer in 1984 as the native format for Macintosh graphics. The PICT format is a meta-format that can be used for both bitmap images and vector images.

BMP files are commonly used file format for the Windows operating system. BMP images can range from black and white (1 byte per pixel) up to 24 bit colour (16.7 million colours).

JPEG (pronounced “jay-peg”) is a standardized image compression mechanism. JPEG stands for Joint Photographic Experts Group, the original name of the committee that wrote the standard.

JPEG is designed for compressing full-color or gray-scale images of natural, real-world scenes. It works well on photographs, naturalistic art work, and similar material; not so well on lettering, simple cartoons, or line drawings.

GIF-The Graphics Interchange Format defines a protocol intended for the on-line transmission and interchange of raster graphic data in a way that is independent of the hardware used in their creation or display.

UNICODE-Unicode provides a unique number for every character, no matter what the platform, no matter what the program, no matter what the language. This is the Standard for NLPs.

ISCII - Indian Script Code for Information Interchange is Standard

4.0 Features

4.1 The OCR software shall have TWAIN interface with Scanner. The OCR software shall support almost all the available make/models of scanners.

4.2.1 Specific Features:

S. No.	Parameter	Level 3 requirement	Level 1 requirement
1.0	Documents category	Newspaper, magazines and printed books, Fax, laser printed documents, forms, Dot matrix print output	Printed books, laser printed documents
2.0	Input Print Quality	Offset print document, Laser/ inkjet print document, Photocopy document, Dot Matrix Printout, Type written Document, spreadshee Newspaper clipping,Fax document,	Offset print document, Laser/inkjet print document
3.0	Color document support	Yes	Nil
4.0	Other supports e.g. table, excel sheet, logo, etc.	Yes\	Nil
5.0	Text search and Indexing	Yes	Yes
6.0	Image/text splitting tool	Yes	Yes
7.0	Size of Paper	Letter, A4, A5	A4
8.0	Image Resolution (DPI)	270-320	300
8.0	Font Size support	8-72	8-12
10.0	Font Support	Italics, Bold, Superscript, Subscript, normal	Normal, Bold
11.0	Column support	Multiple	Single
12.0	Highlighting of non recognized characters	Yes	Nil
13.0	Layout Analysis (picture, table etc)	Yes	Nil
14.0	Image format supported (Before OCR)	BMP, TIFF, JPEG	BMP
15.0	Skew detection & correction	+10 Deg to – 10 Deg	+5 Deg to -5 Deg
16.0	Scheduling	Yes	Nil
17.0	Batch processing	Yes	Nil
18.0	Noise Reduction	De-speckle, de-smudge	Nil
19.0	Output File Format	ISCII, Unicode, Doc, PDF, HTML, XLS etc.	ISCII, Unicode, HTML
20.0	Spellchecker	Yes	Yes
21.0	Text Editor	Editing, copy, paste, Save & formatting etc.	Editing, copy, paste, Save & formatting etc.
22.0	Learning/Training Capability	Yes	Nil
23.0	Operating System	Linux & Windows	Any one

4.2.2 Specific Parameter Values

S No.	Parameter	Level 3 Requirement	Level 1 Requirement
1.	Speed of Conversion (Characters/Sec)	>200cps	>60 cps
2.	Character Accuracy		
	Offset print document	98%	60-80%
	Laser/inkjet print document	99%	70-80%
	Photocopy document	95%	60-70%
	Dot Matrix Printout	95%	Nil
	Type written Document	95%	70-80%
	Spreadsheet	98%	Nil
	Newspaper clipping	90%	Nil

5.0 Ease of Use

S No.	Parameter	Level 3 Requirement	Level 1 Requirement
1.	Auto processing		
2.	Menu Navigation	Simple and usable	Simple and usable
3.	Help and documentation	Yes	Yes
4.	Online help	Yes	Nil
5.	Wizards	Yes	Nil

6.0 Evaluation of OCR

The OCR will be tested as per ISO 12119 with respect to quality characteristics. Functionality will be tested as per clause 5.0 above and usability as per clause 6.0 .The metrics considered for evaluation will be those applicable in ISO 9126-2.

Test samples will be Standardized as follows:

- Sample page from literary work
- Sample page from poetry
- Sample page from official communication
- Multi column report from Standard Newspaper
- Multi column document with picture, table and cartoon from Standard magazine
- Sample political Speech
- Sample laser/inkjet print with multi-font text
- Sample document with specified skew
- Sample photocopy and fax document
- Sample type-written document.

Appendix III

Generic Specifications of Text To Speech System

1.0 Title: Generic Specification for text to speech system

2.0 Scope: This specification covers the following family of products:

S No.	Text To Speech Level	Application
1.0	3	Automatic reading of text encoded in popular Indian fonts from word processor, e-mail, mobile text message, scanned text from newspaper, magazines etc or data entered through keyboard.
2.0	2	Has some application between those at 1.0 and 3.0
3.0	1	Automatic reading of text from data entered through keyboard, and text files

3.0 Terminology& Abbreviations:

- Intelligibility- The sentence is understandable, plain, lucid and logical in meaning
- Scale 1-4(scoring mechanism)

  Level4- The sentence is perfectly clear and intelligible.

  Level3- The sentence is generally clear and intelligible. Despite some inaccuracies of the sentence one can understand immediately what it means.

  Level2- The general idea of the sentence is intelligible only after considerable playback.

  Level1- The sentence is unintelligible. Understanding the meaning of the sentence is hopeless; even allowing for context, one feels that guessing would be too unreliable
- Scale 1-2(scoring mechanism)

  Level 1 Unintelligible

  Level 2 Intelligible

  UNICODE-Unicode provides a unique number for every character, no matter what the platform, no matter what the program, no matter what the language. This is the Standard for NLPs.

  ISCII - Indian Script Code for Information Interchange is Standard

4.0 Features

S. No.	Parameter	Level 3 requirement	Level 1 requirement
1.0	Input category	Newspaper, magazines and printed books, laser printed documents, word processor, e-mail, mobile text message and data entered through keyboard	data entered through keyboard, text from files
2.0	Format Support(Input)	ISCII,UNICODE	Any one
3.0	Format Support (Output)	Speech, Wave file	Speech
4.0	Intelligibility	High(3on a scale of 1-4)	Medium (2 on a scale of 1-2)

5.0	Expansion of Standard Abbreviations	Yes	Nil
6.0	Control Speed and frequency of speech	Yes	Nil
7.0	Allow export to Wave format	Yes	Nil
8.0	Inscript floating keyboard support	Yes	Nil
9.0	Distinguish words and sentences	Yes	Yes
10.0	Conversion of Digits to words	Yes	Yes
11.0	Phonetic conversion	Yes	Nil
12.0	GUI Interface	Yes	Yes
13.0	Mixed language capability	Yes	Nil
14.0	Multi-voice facility	Yes	Nil
15.0	Dynamic switching between multiple voices	Yes	Nil
16.0	Help Menu	Yes	Nil
17.0	Delete, move and search for files	Yes	Yes
18.0	Cut, copy, paste, section of text from one place in a document to another or from another document	Yes	Yes
19.0	Selective reading of a text from the document	Yes	Nil
20.0	Create, open, edit and save a document in various storage devices such as hard disks, floppies or other media	Yes	Yes
21.0	Facility for text Search	Yes	Yes
22.0	Allow document to be printed on a printer	Yes	Yes
23.0	Operating System	Linux Windows98/NT/2000/XP	Windows 9X

5.0 Ease of Use

S No.	Parameter	Level 3 Requirement	Level 1 Requirement
1	Menu Navigation	Simple and usable	Simple and usable
2.	Help and documentation	Yes	Yes
3.	Online help	Yes	Nil
4.	GUI Interface	Yes	Nil
5.	On-screen keyboard support	Yes	Nil

6.0 Evaluation of Text To Speech

The Text to Speech will be tested as per ISO 12119 with respect to quality characteristics. Functionality will be tested as per clause 5.0 above, usability as per clause 6.0. The metrics considered for evaluation will be those applicable in ISO 9126-2.The test bed will be standardized as text / file created from the following:

- Sample page from literary work
- Sample page from poetry with different metre lengths
- Sample of official communication
- Sample multi column report from Newspaper
- Sample multi column document from a standard magazine
- Sample political Speech .

Appendix IV

Generic Specification for Machine Aided Translation (MAT) Software

1. Title : Generic Specifications for Machine Aided Translation Software

2. Scope : This specifications covers the following family of products.

S.No	MAT Level	Application
1.0	3	Content translation, official notifications, literary translations from files
2.0	2	Applications between 3 and 1
3.0	1	HTML, simple sentences, letters

3. Terminology& Abbreviations:

- Intelligibility- The sentence is understandable, plain, lucid and logical in meaning
- Scale 1-4(scoring mechanism)

  Level 4 The sentence is perfectly clear and intelligible. It is grammatical and reads like ordinary text.

  Level 3- The sentence is generally clear and intelligible. Despite some inaccuracies of the sentence one can understand immediately what it means.

  Level 2- The general idea of the sentence is intelligible only after considerable study. It contains grammatical errors and / or poor word choice.

  Level 1- The sentence is unintelligible. Studying the meaning of the sentence is hopeless; even allowing for context, one feels that guessing would be too unreliable
- Scale 1-2(scoring mechanism)

  Level 1 Unintelligible

  Level 2 Intelligible

  Accuracy- the meaning of the source language sentence is preserved in the translation

  A similar scoring mechanism of Scale1-4 or Scale1-2to be adopted, but with reference to the source language.

4.0 Features

S.No	Parameter	Level 3Requirement	Level 1Requirement
1.	Intelligibility	High(3on a scale 1-4)	Medium (2 on a scale 1-2)
2.	Accuracy	High(3on a scale 1-4)	Medium (2 on a scale 1-2)
3.	Format support (input/output)	Doc, html	Text
4.	Input/output	Text, file, speech	Text
5.	Nos. of words, Sentences	Max. nos of word or bytes for translation (pages)	1 page
6.	Speed of Conversion		

7.	Spellchecker support	Yes	Yes
8.	Dictionary	Technical, Legal, Medical.	Normal
9.	Text Editor	Editing, copy, paste, save & formatting etc.	Editing, copy, paste, save & formatting etc
10.	Translation Memory		
11.	Font	Multiple font	One font
12.	Language/Script	Multiple languages	One language
13.	Operating System	Windows, Linux	Windows
14.	Speed of Conversion (Word/sec)		
15.	Accuracy (Word level)		
16.	Accuracy (Sentence level)		

5.0 Ease of use			
1.	Menu Navigation	Simple & usable	Simple & usable
2.	Help and documentation	Yes	Yes
3.	Online help	Yes	No
4.	Wizards	Yes	No

6.0 Method of Evaluation:

MAT systems are very difficult to evaluate and a particular system may do all translations with a score of 2-3 on the Scale 4 and another may do some sentences very well (4) and some poorly (Level 1) on intelligibility. It is therefore to be seen as to what domain the MAT is to be used for and accordingly the test suite developed. The test bed will be standardized as following

Official Communication

Sample test suites for a MAT to translate official communications must have the ability to translate the following:

- Official memos
- Office minutes of meeting
- Official correspondence
- Rules from Standard document like Swamy's handbook

Literary works

A MAT with ability to translate literary works may require to have a test suite consisting of the following:

- Sample text from a modern writer (An Indian Author like Vikram Seth, Arundhati Roy etc)
- Sample text from any writer of Classical literature

In most cases of literary translation there would be the necessity of a good deal of manual editing to make the text readable.

Education Applications

A MAT can be used for the purpose of:

- Translation of Science Texts from English to regional languages
- Translation of newspaper content, magazines etc for educational purposes.

In all of the cases mentioned above the MAT will have to be evaluated as per the features given in clause 5.0 for functionality and 6.0 for usability.

Related standards:

a. ISO 12119: 1994 Information technology - Software packages- Quality requirements and testing

b. ISO/IEC 9126-1:2001 Software engineering - Product quality - Part 1: Quality model

c. ISO/IEC TR 9126-2:2003 Software engineering – Product quality - Part 2: External metrics

d. ISO/IEC TR 9126-3:2003 Software engineering – Product quality - Part 3: Internal metrics

e. ISO/IEC TR 9126-4:2003 Software engineering – Product quality - Part 4: Quality in use metrics

Courtesy :

Dr Hema Khurana,
S. Beena,
E. Kamalakar Rao,
Suresh Chandra
STQC IT Services Centre,
Department of Information Technology
Electronics Test & Development Centre
Bangalore 560058
Email: hema@etdcb.ernet.in