4.2 *Frequently Asked Questions*

1. What is Language Technology?

Language technology researches computer systems, which understand and/or synthesize spoken and written human languages. Included in this area are speech processing (recognition, understanding, and synthesis), information extraction, handwriting recognition, machine translation, text summarization, and language generation.

2. What is Computational Linguistics?

Computational linguistics (CL) is a discipline between linguistics and computer science which is concerned with the computational aspects of the human language faculty. It belongs to the cognitive sciences and overlaps with the field of artificial intelligence (AI), a branch of computer science that is aiming at computational models of human cognition. There are two components of CL: applied and theoretical. The applied component of CL is more interested in the practical outcome of modelling human language use. The goal is to create software products that have some knowledge of human language.

3. What do you mean by bilingual software?

The software supports two languages. One is English and the other is any regional language.

4. What is a script?

A script is the set of symbols required to represent a single writing system, which may in turn be used to represent several languages. Latin, Arabic and Thai are examples of scripts. English, French, German and Latin are all languages written using the Latin script.

5. What is speech synthesis?

Speech synthesis programs convert written input to spoken output by automatically generating synthetic speech. Speech synthesis is often referred to a "Text-to-Speech" conversion (TTS).

6. What is ISCII?

Bureau of Indian Standards formed a standard known as ISCII (Indian Script Code for Information Interchange) for the use in all computer and communication media, which allows usage of 7 or 8 bit characters. In an 8 bit environment, the lower 128 characters are the same as defined in IS10315:1982 (ISO 646 IRV) 7 bit coded character set for information interchange also known as ASCII character set. The top 128 characters cater to all the Indian Scripts based on the ancient Brahmi script. In a 7-bit environment the control code SI can be used for invocation of the ISCII code set and control code can be used for reselection of the ASCII code set.

There are 15 officially recognized languages in India. Apart from Perso-Arabic scripts, all the other 10 scripts used for Indian languages have evolved from the ancient Brahmi script and have a common phonetic structure, making a common character set possible. An attribute mechanism has been provided for selection of different Indian script font and display attributes. An extension mechanism allows use of more characters along with the ISCII code. The ISCII Code table is a super set of all the characters required in the Brahmi based Indian scripts. For convenience, the alphabet of the official script Devnagari has been used in the standard. The standard number IS1319:1991 issued by Bureau of Indian Standards is the latest Indian Standard for Information Interchange, and is being widely used for development of IT products in Indian Languages.

7. What is ACII Script Code?

Alphabetic Code for Information Interchange (Pronounced as Ae-Kee). It is a new name given to ISCII code which now encompasses national scripts of SAARC countries also. This is a 8-bit code, containing the ASCII character set in the bottom half. The top half contains the ACII characters. PC-ACII Script code is the version of ACII script code where the characters are split in the upper-half for compatibility with IBM PC.

8. How is text represented through ACII?

ACII (Alphabet code for Information Interchange) code contains all the basic characters available on the ACII keyboard. For example, The ACII Indian code and keyboard accommodates the requirements

for the 10 Indian scripts: Assamese, Bengali, Devanagri, Gujrati, Kannada, Malayalam, Oriya, Punjabi, Tamil and Telugu. The basic characters are ordered such that direct sorting gives results, which are almost the same as that for any of the scripts. The ACII codes have to be converted to ISFOC for display purpose. This is done through an ISFA algorithm for the selected script. An ACII text can be displayed in any of the scripts. Transliteration to another script can be achieved by merely selecting that script. ACII code is used in communication media, like telex, for optimal transfer of text. ALP word processor uses the ACII code internally to allow proper editing at alphabetic level and unique representation of spellings.

The existing window applications are unable to handle ACII directly, as it requires an intelligent algorithm for handling the display. They can, however handle the ISFOC codes, which were made for this purpose. Thus, conversion is necessary between ACII and ISFOC whenever text has to be transferred from ALP to a window application. It is possible to type ISFOC text directly within a windows application using the ACII keyboard. This is done through a custom keyboard driver who does ACII to ISFOC conversion internally.

9. Are there any new entities required for ensuring proper representation of complex scripts?

Following are the entities required for ensuring proper representation of complex scripts:

ACII- Alphabetic code for Information Interchange This is a computer code by which the basic alphabet of a script is represented. The basic letters and signs needed in most of scripts (leaving aside ideographic scripts like Chinese) are less than 96. All the possible shapes in a script can be expressed through combinations of these basic letters. The ACII code can be typed through an ACII keyboard overlay. The ACII keyboard overlay fits on a standard English keyboard. Each ASCII character has a unique position on the keyboard overlay.

ISFOC- Intelligence Based Script Font Code ISFOC is a coded character set containing all the basic shapes required for rendering a script. These shapes can be overlapped linearly to compose any word in the script. Each of the ISFOC characters is like a piece of a jigsaw puzzle; it may not be a complete letter by itself. Each ISFOC set can contain a maximum of 188 characters. This is adequate for most of the scripts. However, some require more.

ISFA- Intelligence Based scripts to Font Algorithm A word is always typed in terms of its basic ACII characters. It however, has to be displayed using the basic ISFOC shapes. An algorithm is required for converting the ACII codes to the appropriate ISFOC code. This is the ISFA algorithm.

10. What is UNICODE?

Unicode is increasing being accepted as a standard for Information Interchange worldwide Unicode for Indian Languages use ISCII-88 and not ISCII-91 which is the latest official standard.

Unicode standard is the 16 Bit (2 Byte) Universal character encoding standard, used for representation of text for Computer Processing. Unicode standard provides the capacity to encode all of the characters used for the written languages of the world. The Unicode standards provide information about the character and their use. Unicode Standards are very useful for Computer users who deal with multilingual text, Business people, Linguists, Researchers, Scientists, Mathematicians and Technicians. Unicode uses a 16 bit encoding that provides code point for more than 65000 characters (65536). Unicode Standards assigns each character a unique numeric value and name. The Unicode standard and ISO10646 Standard provide an extension mechanism called UTF-16 that allows for encoding as many as a million characters. Presently Unicode Standard provide codes for 49194 characters.

11. What is a font?

A font, as far as a computer is concerned, is the file or files necessary to display and print a particular typeface. Dv-TTYogesh, for example, is a typeface.

They are also referred to as fonts. Each font comprises one or more files, depending on the font technology used.

12. What is a font family?

Font families are collections of fonts which look similar but have slightly different attributes. Dv-TTYogesh Regular and Dv-TTYogesh Bold are two different fonts, but in the same family.

13. What is a bitmapped font? What is a screen font?

A bitmapped font is also referred to as a screen font. They are files which contain pixel information your computer uses to display the font on the screen. Bitmapped font files are for a particular point size. If you have bitmapped fonts for Helvetica at 12 point and 14 point, Helvetica at 13 point will look slightly pixelated on the screen. If all you have installed is bitmapped fonts, your printer will print fonts which don't look very smooth. Font sizes which are physically installed will look better, but there are problems with having too many fonts open simultaneously. There is font technology to avoid that problem these days, and you are likely using some of that technology.

14. What is a printer font? What is a PostScript font?

When you talk about printer fonts, you are usually talking about PostScript. PostScript fonts come in pairs (there may be more than two files involved): one or more screen (bitmapped) fonts, and one printer font.The printer font is scalable, meaning that whatever font size you are using will be scaled properly by a PostScript-capable printer, and will look smooth on the paper. The printer font is used for printers, the screen font is normally only used for on-screen display. You may have multiple bitmapped fonts which are all linked to the same PostScript font. For example, you may have Helvetica Bold 12 pt, Helvetica Bold 14 pt, and Helvetica Bold 24 pt, but they will all use the same printer font, HelveBol.

15. What is an OpenType font?

Open Type is a cross-platform font file format developed jointly by Adobe and Microsoft. The two main benefits of the Open Type format are its cross-platform compatibility (the same font file works on Macintosh and Windows computers), and its ability to support widely expanded character sets and layout features, which provide richer linguistic support and advanced typographic control.

The Open Type format is an extension of the TrueType SFNT format that also can support Adobe PostScript font data and new typographic features. Open Type fonts containing PostScript data, such as those in the Adobe Type Library, have an .otf suffix in the font file name, while TrueType-based Open Type fonts have a .ttf file name suffix.

Open Type fonts can include an expanded character set and layout features, providing broader linguistic support and more precise typographic control. OpenType fonts can be installed and used alongside PostScript Type 1 and TrueType fonts.

16. What is a dfont?

A dfont is a special version of a Macintosh TrueType font. All the information that is normally stored in a TrueType font's resource fork has been moved to the data fork. Typically, the only dfonts you will run into will come with Mac OS X.

17. What is TrueType font?

TrueType is a font technology from Apple which allows you to have smooth font screen displays and printing without needing extra screen font sizes or PostScript. TrueType fonts will print smoothly to non-PostScript printers. TrueType fonts consist of one scalable TrueType file, and possibly one or more bitmapped screen fonts. Although TrueType technology is very efficient, and removes the need for Adobe Type Manager to smooth your fonts, some PostScript printers have problems with TrueType fonts.

18. What are dynamic fonts?

Dynamic fonts are the technology used for delivering windows true type fonts on the client side in transparent way. If the user needs to provide a facility of viewing the pages in Indian Languages then fonts

can be delivered to the client in EOT and PFR format.

19. What are EOT (Embedded Open Type) & PFR (Portable Font Resource) format?

EOT (Embedded Open Type) format of fonts is Microsoft's way of sending encoded fonts to the clients. Only Internet Explorer, (version 4.0 Onwards) can use EOTs. EOTs have specific URL. If the web designer provides a link to an EOT the browser uses these EOTs to display the page. This means that only particular websites with links to the specific URL can use EOTs made for them. PFR (Portable Font Resource) is another way to send fonts dynamically to the user. It can be used both in Netscape (4.03 and above) and IE (4.0 and above). In IE however there is a one time download of a control on the clients machine. PFRs also have the URL security and can be locked to particular URLs. PFRs are more stable than EOTs but sometimes need Encoding changes in IE 5.0.

Usually a JavaScript is used to query the browser and accordingly PFRs or EOTs are given to the client so that a particular font can be displayed without user intervention.

20. How do fonts get activated?

Normally your fonts reside in the Fonts folder in the windows directory. The computer boots up, looks in the folder, and turns the fonts on. They are then available to all applications. If you put lots of fonts into your windows fonts directory, your system will slow down drastically, and you could run into stability problems as well.

21. What is font manager?

It helps you to have more fonts without causing system problems. When you use a font manager you keep your fonts elsewhere on your system than in the Fonts folder or directory. The font manager keeps a list of all your fonts, and you can turn on the ones you need and turn off the ones you are done with. You will have to restart most applications before the fonts will be available.

22. What are different Keyboard Layouts for typing in Indian Languages?

There are 4 different keyboard layouts.

1. Romanised Layout : In Romanised layout, phonetic English mappings are used to compose the Hindi Text. For example, the key raamaa (or rAmA) can be used to type 'Rama'.

2. Typewriter Layout : This layout is similar to the Hindi typewriter layout & useful for Hindi typists & other people familiar with Hindi Typewriter layout. Typewriter Layout & Key Sequence Charts

3. Phonetic Layout : This layout is standardized by the erstwhile Department Of Electronics (DOE), Govt. of India. The advantage of this layout is that the layout remains identical for all Indian Languages. For example, the key 'k' is used to represent the letter 'ka' in all Indian Languages. The Keyboard Layout and the Key Sequence Charts can be used to find the correct key combinations.

4. Consonant Keyboard : The phonetic division of Indian alphabets into Vowels and Consonants serves as a common base for all Indian scripts. Vowels are called soul and consonants are called body. The combination of these two become animated body. Without the addition of a vowel (soul) the consonant (body) is like a 'dead letter'. The dead consonant can also be termed as 'Pure consonants'.

The keyboard which accommodates the phonetic peculiarities efficiently and taking into account the inherent logic built in Indian scripts, is named as DESHA (consonant) keyboard (DESHA meaning Country). DESHA is based on the Barahkhandi' concept. Through DESHA Keyboard, all possible glyphs combinations as used in linguistic/ information environment, are produced using only 36 consonant keys and 12 vowel keys. In Desha Keyboard there are no separate keys for vowel signs and vowel matra signs. The keys showing the vowel signs produce vowel signs as well as vowel matra signs as per the inbuilt logic.