	<b>Export HIS project – Localization</b> <b>Tekes grant nr 70062/04</b>	
	Writer	Hellevi Ruonamaa (hellevi.ruonamaa@uku.fi)
	Document status	Finalized
	Date	8.1.2006

## Software localization to China and Chinese

Introduction .....	3
1 A brief introduction to Chinese language .....	3
1.1 The spoken Chinese .....	4
1.2 The written Chinese .....	4
1.2.1 Characters .....	4
1.2.2 The simplification .....	5
1.2.3 The transliteration .....	6
1.3 The language summary .....	6
2 The software globalization .....	8
2.1 The internationalization.....	8
2.2 Linguistic tools in software localization .....	9
3 Localization problems .....	11
3.1 Characters .....	11
3.1.1 Character sets and encodings .....	11
3.1.2 Fonts .....	12
3.1.3 Sorting and indexing .....	15
3.2 Linguistic aspects .....	16
3.2.1 Text Input.....	16
3.2.2 Text orientation .....	20
3.2.3 Word separation, word wrapping and hyphenation.....	20
3.2.4 Punctuation .....	21
3.3 Cultural aspects .....	22
3.3.1 Numerals.....	22
3.3.2 Date and Time .....	23
3.3.3 Currency .....	25
3.3.4 Measurement Units.....	25
3.3.5 Icons, Symbols and Graphics .....	26
3.3.6 Colours.....	26
3.3.7 Local conventions .....	27
3.4 User interface .....	28
3.5 Hardware devices .....	28
3.6 Software services .....	29
3.7 Support .....	30
3.8 Other parts to be localized.....	31

4	Software development products .....	32
4.1	Operating systems.....	33
4.1.1	Windows .....	33
4.1.2	Sun Solaris.....	37
4.1.3	HP-UX.....	38
4.1.4	Asianux (Miracle Linux, Red Flag Linux, Haansoft) .....	39
4.2	Databases.....	39
4.2.1	Caché.....	39
4.2.2	Oracle .....	40
4.2.3	MS SQL Server.....	40
4.2.4	Sybase Adaptive Server Enterprise .....	41
4.3	Development environment.....	42
4.3.1	Java .....	42
4.3.2	.net.....	44
4.3.3	Qt (Trolltech AS) .....	44
5	Summary.....	45
	References .....	46

## Introduction

This document will focus on those issues of the software localization process that must be weight up specifically when you develop or localize a software product to China and Chinese. It is aimed to Western people who are used to work only with computer environment of one-byte ASCII or ISO8859 characters. This paper includes also general information about problems in Chinese information processing, not only in localization.

The most important problems in the information processing that will be discussed in this paper are the character sets, fonts, input methods and required national standards in the software development for the Chinese market. In addition to those also available tools for software development are studied.

Internationalization and localization as specific problems will be discussed in this paper very slightly. Also problems that will be faced in Chinese word-processing software, like word segmenting, are much more complicated and will not be studied here.

This is not a 'how to do localization' guide but it contains information that one should be aware of when localizing software into Chinese. This gives IT staff information about common requirements of Chinese software development and help them decide what tools can be used.

The first chapter has the basic facts about Chinese language to help to understand the diversity of problems related to the language. The second chapter has general information about software globalization. The chapter three describes problems that will be faced in Chinese software and in the chapter four there is information about products.

The Product section is not very detailed because features of products are changing all the time. Several products mentioned in this paper have added the required GB18030-2000 standard support during the past twelve months.

## 1 A brief introduction to Chinese language

The 'Chinese language' is a group of languages that are mostly spoken in Mainland China and Taiwan. It is spoken by more than one billion people and approximately 95 percent of the population of China speak Chinese. Other languages spoken in China are e.g. Tibetan, Mongolian, Lolo, Miao, and Tai which are spoken by minority groups.

The majority of the Chinese-speaking population is in Mainland China and Taiwan, but there are Chinese speaking groups also in Singapore, Indonesia, Malaysia, and Thailand.

When entering to Chinese market it is important to understand the diversity of the Chinese language. It has seven major language groups and each of them has several dialects. Fortunately there are only two writing forms if we don't speak about languages of minority groups. The combination of the spoken language and the writing form depends on the location. The fact that Mandarin is the official language of China and used character set is Simplified Chinese gives software companies an opportunity to use those and their software will cover almost the whole country.

In the following chapters there will be introduced some characteristics of spoken and written Chinese.

## 1.1 The spoken Chinese

The three biggest spoken Chinese language groups are Mandarin, Wu and Yue. Mandarin is based on a dialect used in Beijing area and it is the most widely spoken covering at least 80 % of the country area and 74 % of the population. Because the area where Mandarin is spoken is so huge, it varies in some amount. Mandarin is the official language of China and it is also one of the four official languages in Singapore.

The second biggest spoken language group is Wu (吳) whose dialects are spoken in Zhejiang and Jiangsu provinces and in Shanghai and Hong Kong and Taiwan by about 80 million people. Its major dialects are Shanghainese and Suzhounese.

The third biggest spoken language group is Yue (粵) which is also called Cantonese. Its dialects are spoken by about 65 million people in Guangdong, eastern part of Guangxi and Hong Kong and parts of Macau etc.

Because Mandarin, Wu and Cantonese etc. are spoken Chinese languages, those names should not be used when one needs a written translation but when one needs an interpreter.

In Mandarin there are only 1300-1400 differently pronounced syllables consisting of about 400 basic pronunciations and four or five tones for most of them. As a result, there are many homophone words which are distinguished in written Chinese by using a different character for almost each one.

In China there are also other official languages e.g. Tibetan, Mongolian, Lolo, Miao, and Tai which are spoken by minorities.

## 1.2 The written Chinese

### 1.2.1 Characters

Chinese does not use phonetic alphabets when creating words but characters that represent a whole word or only a part of it. There are altogether more than 80 000 characters but many of them have overlapping meanings or they are variants of other character. The commonly used 3000-5000 characters can easily create more than 50 000 words.

New words are built using existing characters. For example 'computer' is 电脑(dian4nao3) in Chinese where 电 (dian4) means 'electricity' and 脑 (nao3) means 'brain'. A number after the syllable expresses its pronunciation tone.

#### The four types of characters

There are four types of Chinese characters; pictographs, ideographs, compound characters and semantic-phonetic compounds.

*Pictographs* were originally pictures of things and sometimes the original shape can still be recognized. For example 川 (chuan1) means 'a river', and 山 (shan1) means 'a mountain'.

*Ideographs* are graphical representation of an abstract idea. For example 三 (san1) means 'three' and 中 (zhong4) means 'middle'.

*Compound ideographs* are built by combining one or more pictographs or ideographs to form a new character. All component parts contribute to the meaning of the compound character. For example 明 (ming2) consists of a sun and a moon and the meaning of it is 'bright'

*Phonetic ideographs* consist of two parts where the first one gives a hint of the meaning of the character and the second one gives an idea of the pronunciation. This is the biggest group, more than 90% of all characters. For example 钟 (zhong1) means 'a bell' and the character is a combination of 钅 (referring to metal) and 中 (zhong1) referring to its pronunciation.

Most words are created of two or sometimes more characters and characters are written in a sentence without a space or any other separator between words. Therefore, while reading, in addition to understanding what characters do mean and knowing how to pronounce them, you must also know which characters standing next to each other belong together.

### The structure of a character

Characters are composed of smaller units called radicals and other elements that cannot serve as radicals. One radical in a character can be used for indexing purposes. There are altogether some 200 radicals where amount depends on the radical system and several radicals also stand alone as a character.

As an example, 明 (ming2) consists of two elements, 日 and 月 where the first one is a radical that is used for indexing.

Radicals and other elements are composed of one or more strokes. Each character has a defined drawing order that is used while drawing strokes. That order of drawing can also be used, as well as the number of strokes, for indexing purposes. The 明 character is composed of eight strokes where the first stroke is a vertical bar and the second one is a hook.

When writing Chinese, every character is given exactly the same amount of space, no matter how many strokes it contains. It means that radicals and other elements are stretched and squeezed so that they fit to the square that is reserved for each character. For example character 一 (yi1) has only one stroke but 龘 (tang4) has 36 strokes and they both have to fit to the square of the same size. This document uses normally font size 10 but if we, who are not used to look at Chinese characters, really want to see the latter character the size must be raised up to 24:

一 (yi1), 龘 (tang4)

### Writing direction

In Simplified Chinese (see 1.2.2. The simplification) almost all publications and in Traditional Chinese most books and newspapers are printed horizontally from left to right. Vertical direction starting from the right side is used only in advertising and some special cases.

## **1.2.2 The simplification**

There are two main versions of written Chinese: Simplified Chinese and Traditional Chinese. The main objective of simplification was to raise the literacy rate. The simplified writing system reduces the number of strokes per character and the number of characters in common use. The most frequently used and complex characters are written with less strokes in Simplified Chinese. There are approximately 2500 characters that have a simplified form.

The Simplified Chinese was implemented in 1952 and it is used in China and Singapore, Traditional Chinese is used in Taiwan, Hong Kong, Macao and overseas Chinese communities.

### 1.2.3 The transliteration

Because Chinese writing system is based on logographic symbols, there is no way of knowing their pronunciation only by looking at the character. In order to handle the pronunciation there are a few transliteration systems used. Some of these systems use Latin alphabet, e.g., Pinyin and Wade-Giles. These systems are called "romanization systems". There are also other transliteration systems that do not use Latin alphabet e.g. Zhuyin Fuhao.

The Pinyin system was developed in China in 1958 and it is now the only transliteration system used in China. It consists of 25 Latin letters (not 'v'). Pinyin has been the United Nations Standard from 1977 and it is adopted also by ISO (the International Organization for Standardization). The Wade-Giles system was first published by Thomas Francis Wade in 1859 and it was earlier almost the only system in English-speaking countries. It is still popular in Taiwan. As an example that we meet quite often one could mention the name of the capital of China: if it is Peking it is Wade-Giles transliteration but if it is Beijing it is Pinyin transliteration.

Zhuyin system does not use Latin alphabets but Chinese character elements to express pronunciation. It is also known as "bopomofo" according to the pronunciation of the first four characters in its character set. It consists of 37 characters and it was first introduced in early 20th century and is still used at least in Taiwan.

## 1.3 The language summary

As a conclusion to the previous chapters, the Simplified Chinese is used in written language and Mandarin, Wu or Cantonese or some other dialect is used in spoken language in Mainland China. In Hong Kong Traditional Chinese is used in writing and Cantonese is mostly used in spoken language. In Taiwan Traditional Chinese is used in writing and Mandarin or Wu is used in spoken language.

## Software localization to China and Chinese

A rough table of the spoken language, transliteration and characters used in different areas:

Transliteration, spoken and written languages in different geographic areas	Spoken language			Transliteration			Written language	
	Mandarin	Wu	Cantonese	Pinyin	Wade-Giles	Zhuyin	Simplified	Traditional
<b>Mainland China</b>	yes	yes	yes	yes			yes	
<b>Hong Kong</b>	yes		yes		yes			yes
<b>Taiwan</b>	yes	yes			yes	yes		yes

The used written language has an influence on needed character set and the transliteration system has influence on input methods. Because Mandarin is the official spoken language, the Simplified Chinese is the written form of it and Pinyin is the only transliteration system used in Mainland China, software products that conform to those will cover most of China. But you must keep in your mind that this software will not cover Taiwan, HongKong or minorities of China.

## 2 The software globalization

A software globalization is a process where software is build for use in global market. The first part of building global software is to build an international version that can be localized to different languages and cultures with minimal effort. The localization is a process where features of a specific language and region are adapted to the internationalized software.

### 2.1 The internationalization

The internationalization is a phase of a software globalization where linguistic and cultural dependent parts are isolated from the core. This requires extracting all language, culturally and country dependent elements and while creating the core software there is no assumptions of the target language or region.

It is possible to localize a software product that has not been internationalised at all. This requires a lot of work that has to be repeated every time for a new version of the product. This is not recommended at any circumstances. Another option is to first internationalise the existing product and then localize it. Using this method must be thought through because multilinguality has so deep impact on the software architecture that it is difficult to avoid problems. If a product has originally been developed for international use, and cultural and linguistic parts have been isolated in the first step, it will be easier to build localized versions to different cultural and linguistic regions later.

Because internationalization leads to a more modular product, support and maintainance of local versions of the software will be simpler. Also getting new versions of internationalised software to different markets will be much faster and less expensive. If there is only one code to be used everywhere it saves money in all levels: corrections only to one place, one compiled version of the software, debugging only in one version and much less testing. These all means less hours spent with software and less provided hardware. Technical support is easier with one code and it is better service for international customer if the product has no differencies in different countries.

There are two models for internationalization: The locale model implements a set of attributes for specific locales and one of those attributes is a character set. The character set is specific to a given culture, region or locale and it cannot be changed if locale is not changed. In the multilingual model the system can use a character set that contains all characters that are necessary for several cultures or regions. A multilingual software can be carried out by using the Unicode character set.

In the internationalization process the following parts of the software product should be set isolated from the core:

- window titles
- text strings, also text that is embedded to graphics
- menus
- dialogues
- presentation information (position, size, font, colour, intensity, text orientation)
- error and confirmation messages
- help text
- report layouts

- symbols
- graphics
- icons
- sounds
- cursor shape
- hot keys
- function calls, input arguments and output data
- strings communicated between components
- folder names
- constants
- formulas (tax, pension, ...)
- functions to change data from internal to external form and vice versa
- all references to devices
- all references to operating system
- all software services that have user interaction

The code of a localizable product has three basic requirements:

- The code supports localization with no modification to executable code.
- The code operates properly with all target character systems e.g. double or multibyte systems.
- The code operates correctly with texts in several different languages.

Outside of the basic software product there is also a need for a localizable documentation and help system.

There are a couple of useful websites for checking the internationalization features:

Microsoft: <http://www.microsoft.com/globaldev/getWR/nwr/notworldready.mspx>

Sun Microsystems: <http://developers.sun.com/dev/gadc/i18ntesting/checklists/index.html>

## 2.2 Linguistic tools in software localization

In addition to software development products you can ease our localization work by using linguistic tools. Linguistic tools are planned to face challenges of translating textual parts of the software product.

### Terminology management systems

Terminology is the foundation of a good documentation and translation. The purpose of a terminology management system is to collect and organize terminological data by building a *terminological database* that contains entries in different languages. It can help communication between original writers and those who carry out the localization.

### Translation memories

Translation memory is a database where previous translations, the source and the target language text, are collected to a *translation database*. When the subsequent version of the

same source is compared with the original, the previous translation is inserted into the new target text.

### Machine translation

Machine translation system performs linguistic analysis on the source text and translates them into the target language. Because machines cannot deal with ambiguity in the way that humans can, the result is not comparable to human translation but it can be used after post-editing. Machine translation works best on unambiguous texts but also in that case the basic requirement is good dictionaries used in the translation system.

### Globalization Management systems

Globalization Management systems are tools for translation of large and constantly changing websites. It consists of an engine that monitors site content and a component that passes content to translators or other linguistic tools for further processing. It also manages the workflow and synchronisation of translated content with the source-language website.

### Translation workbenches

A translation workbench is an integrated set of tools that support localization.

There are a lot of companies that have products mentioned above. They have a whole translation workbench or a smaller part of it. Besides tools they often also offer globalization services.

## 3 Localization problems

This chapter will discuss problems that will be faced in Chinese software. The biggest issue to concern is the character set and problems following that. In addition, there are linguistic and cultural aspects to handle as well as technical problems to be solved.

### 3.1 Characters

The first question most western people ask when discussing Chinese software is the question of characters. It is a big question and leads to other questions e.g. character sets, encodings, fonts and sorting.

#### 3.1.1 Character sets and encodings

A character set is a bunch of characters that have been set to a group based on some reason, for example a language or a group of languages or some other reason. Some character sets are non-coded, some are coded. Characters in a coded character set have been mapped to a numeric value and can be used by computers.

An encoding is the process where numeric values are set to characters. Different encodings set usually different numeric values to each character. Sometimes when an encoding is defined as add-on to an existing encoding, old values have been used when possible.

One digital byte can represent only 256 characters and there are more than 80 000 Chinese characters. Though all of them do not belong to any coded character set it is obvious that more than one byte must be used for encoding. Though a Chinese character is composed of two or more bytes, it needs to be treated as one character during all operations. In other words, when user is treating one character, all bytes that belong to this character must be treated e.g. while deleting a character, all its bytes need to be deleted, otherwise data is corrupted.

From 1981 GB2312-1980 was the official encoded character set of China. GB in the name means GuoBiao, national standard. It includes almost 7000 Simplified Chinese characters in two sets. The first set, 3755 frequently used characters, is arranged by pronunciation (Pinyin transliteration). Another set, 3008 less frequently used characters, is arranged by radical and then by number of strokes. GB2312-1980 also includes Zhuyin symbols, Pinyin vowels with tone marks, Latin alphabets, numerals in various series, punctuation and Japanese kanas and Greek and Cyrillic alphabets as well as some other symbols.

GB13000-1993 is compatible with Unicode 2.1 and it has code point for 20902 CJK (Chinese, Japanese and Korean) characters.

GBK is an extension to GB2312-1980. It is not a standard but an encoding specification of Hanzi (Chinese character used in Chinese), implemented in 1995. Last alphabet 'K' is the first letter of Kuozhan that means extension. It consists of GB2312-1980 characters, GB13000.1-1993 characters and some other having altogether 23940 code points and 21886 characters. It has also mapping to Unicode 2.1.

The latest encoding standard GB18030-2000 was released on March, 2000 and updated on May 2001. Its official name is 'Chinese National Standard GB18030-2000: Information Technology - Chinese ideograms coded character set for information interchange - Extension

for the basic set' (资讯技术 - 资讯交换用汉字编码字元集 - 基本集的扩充). GB18030-2000 was created as an update to GB2312-1980 for Unicode 3.0.

GB18030-2000 has, among others, the following properties:

- It is backward compatible with the previous official encoding standard GB2312-1980 having the same numeric value for each character of GB2312-1980.
- It includes all GB 13000.1-1993 characters (about 20000 Chinese characters)
- It includes all characters that are in Unicode 3.0 (6582 Chinese characters more)
- It includes characters of Tibetan, Monogolian, Yi and Uyghur languages.
- Characters are encoded in one, two, or four-byte sequences.
- It has 1.6 million byte sequences, where about 500,000 are currently unassigned.
- It provides code space for all used and unused code points of Unicode's plane 0 (BMP) and its 16 additional planes.
- It has mapping table to Unicode 3.0 code points.

All language related products in China must be able to use all the characters in GB18030-2000 and any product released on or after the 1st of September 2001 must be certified by one of the following groups.

- A+ The product supports the input, output, edit and display of all characters in GB18030-2000, including minority scripts.
- A The product supports the input, output, edit and display of all characters in GB18030-2000, excluding minority scripts. The product must not corrupt minority characters even if it does not have fonts to display them.

GB18030-2000 applies to the processing, interchange, storage, transmission, display, input and output of graphical character information.

There is more discussion about GB18030-2000 in products in the chapter 4. 'Software development products'.

### 3.1.2 Fonts

It is not enough to define the character set and its enodings but there must also be a digital definition of the printable form for each character. Printable forms have different fonts. The following examples of Chinese fonts are from the web page of URW++ Design & Development Company from Germany.

Fang Song	啊阿埃挨哎唉哀皑癌藹矮艾
Hei	啊阿埃挨哎唉哀皑癌藹矮艾
Song	啊阿埃挨哎唉哀皑癌藹矮艾
Yuan	啊阿埃挨哎唉哀皑癌藹矮艾
Hupo	啊阿埃挨哎唉哀皑癌藹矮艾

Kai	啊阿埃挨哎唉哀皑癌藹矮艾
Lishu	啊阿埃挨哎唉哀皑癌藹矮艾
Wei Bei	啊阿埃挨哎唉哀皑癌藹矮艾
Zong	啊阿埃挨哎唉哀皑癌藹矮艾

The amount of characters in Chinese character sets is a constraint on font providers. Only a few companies have resources to create fonts for thousands of characters. To mention a few of them, there are at the moment Bitstream, Founder Group, Changzhou SinoType Technology Co and Zhong Yi Electronics.

**Bitmapped fonts**

If a character set has bitmapped font it means that every character is constructed as a dot-matrix. They are not user friendly to scale to a bigger size and if the provider wants to offer different size of readable characters, one should design a new set of fonts for each size.

**Outline fonts**

Outline font characters are constructed from outlines what means that each character is described mathematically as a sequence of line segments and curves. The outline is scaled to the requested size, then filled and converted to a bitmapped image to the output device. An outline font character can be used at any size and designer needs to design only a single point size characters.

*PostScript* is a page-description language developed by Adobe Systems. It supports both text and graphics and provides built-in support for fonts. It has several font formats of which Type 1 is the most widely used. In PostScript each glyph in a character set has a unique CID (character id) that is a numeric value independent of any encoding. Encoding is associated to CIDs in Cmap files, where an encoding range is associated with a CID range. The ranges can be short or long and a row in the Cmap file consists of three numbers, the first and the last values of encoding and the third one is the starting point of CID from which encoded values are associated with CIDs.

*TrueType* font format was originally developed in 1980s by Microsoft and Apple. Later they have developed TrueType separately and now the two font formats are incompatible with each another. All TrueType fonts contain "cmap" tables that map the glyphs to encodings.

In 1996 PostScript and TrueType were merged to *OpenType* standard that was developed by Adobe Systems and Microsoft.

The following companies have developed GB18030-2000 fonts that have been certified to use in software products in China.

**Agfa Monotype Corporation** has two fonts, Hei Bold and Sung Light that have been approved in 2002 by the Committee on Information Technology Standards (CITS) and the State Language Committee (SLC) for distribution within China. They are from Agfa Monotype's WorldType® multilingual font library and include full support for the Chinese character set standard GB18030-2000.

Agfa Monotype Hei Bold

# 蒙纳优质中文字库

Agfa Monotype Sung Light

# 蒙纳优质中文字库

Agfa Monotype specializes in fonts and font technologies for graphic professionals, software developers and manufacturers of printers and display devices. Agfa Monotype is a subsidiary of Agfa Corporation and is part of Agfa's Graphic Systems business unit. Agfa is the U.S. subsidiary of the Agfa-Gevaert Group, one of the world's leading imaging companies.

**Bitstream Incorporation** from the United States has GB18030-2000 font Hei approved on June 28, 2005 by the CITS and the SLC for distribution within China.

Bitstream Incorporation is a software development company that enables customers worldwide to render high-quality text, browse the Web on wireless devices, select from the largest collection of fonts online, and customize documents over the Internet. Its core competencies include browsing, font, and publishing technologies. Its library includes over 1,000 high-quality fonts in OpenType, TrueType, and PostScript Type 1 formats for Windows, the Macintosh, Unix and Linux.

**Beijing Founder Electronic Co., Ltd** has developed TrueType GB18030-2000 fonts ShuSong, Hei, Kai, FangSong and SongYi. In 2001 those fonts passed the national certification and were accepted to distribution in China.

Font samples:

ShuSong

于字里行间 显方正科技

SongYi

于字里行间 显方正科技

Kai

于字里行间 显方正科技

FangSong

于字里行间 显方正科技

Hei

## 于字里行间 显方正科技

Founder Group is a leading provider of advanced information technology, software products, collaborative business solutions, and other value-added services. Beijing Founder Electronics Co., Ltd. (Founder Electronics), a subsidiary of the Founder Group, has emerged from a small peripheral R&D department of Beijing University to one of today's largest technology companies in China. Founder Electronics provides its products and services to customers whose businesses cover a wide range of industries worldwide, including newspapers, commercial publishing, printing, broadcasting, TV, Internet, Libraries, and Government administration.

Beijing Founder Electronic Co., Ltd. has developed also Founder Super Font, which includes 70244 Chinese characters. Committed to promoting the technology of Chinese Fonts and developing fonts for Chinese character printing, Founder Group offers more than ten new fonts every year. Meanwhile, the company actively participates in formulating the relevant Chinese National Standard and International ISO Standard, e.g. GBK, GB18030-2000 and ISO/IEC-10646 etc.

**Changzhou SinoType Technology Co., Ltd** has developed STSong™ Light font that is certified in 2002 by the Press and Publication Administration of China and CSLC and the National Typeface Committee.

**Hunan Huatian Information Industry CO., Ltd** has four GB18030-2000 truetype fonts, Song, Fang Song, Kai and Hei. There is no information about the year of their approval by CITS or SLC for distribution in China.

Hunan Huatian Information Industry Co., Ltd. is a high-technology enterprise whose goal is in developing national information industry and researching and developing the international advanced information technology and products.

**ZhongYi Electronic Ltd** has Song, Hei, Kai and FangSong GB18030-2000 fonts that have been approved by China State Language Committee, China State Press Publication and National Printing Font Committee.

ZhongYi Electronic Ltd from Beijing, is the affiliated company of Chinese Standard Technology Ltd. The core technologies of the company include input method, full text search engine which supports super large character set. ZhongYi provides font customization services to generate Bitmap, TrueType, Postscript, OpenType fonts in accordance with customer's requirements.

### 3.1.3 Sorting and indexing

Chinese written text is often sorted by Pinyin reading. Because Pinyin transliteration uses Latin characters, it is therefore sorted according to the English alphabet. Sorting can also be done according to the character structure.

Indexing is an essential issue when trying to find the right character. It can be done in various orders e.g. phonetic, radical + number of remaining strokes and number of strokes.

In the phonetic order characters are ordered according to the pronunciation in Pinyin transliteration system. Because there are lot of homophones in Chinese languages every item in index has several entries. And because some Chinese characters have multiple readings, phonetic indexing has multiple entries for that character. There are also characters that do not have any well-known reading. In that case character does not have any entry in the phonetic index.

If a character does not have a phonetic index, it may have a radical index. Radicals and radical-like elements are the basic blocks of a character. Radical index has two levels where the first table consists of radicals with their identifying numbers. The other table consists of all radicals and there are, under each radical, a list of characters that have requested radical as indexing radical. Those characters under the radical are ordered according to the number of remaining strokes in the character. When using radical index one must first recognize the radical of the character and then count the number of remaining strokes.

Sometimes it is difficult to recognize the radical of a character but in that case one can use the third method of indexing. It uses the number of total strokes including also strokes of the radical. If characters were grouped only by this number there would be too many characters in a group. This is why there is a second level also used. The most common way is to order strokes in a stroke count by radical. The other method is to order them according to the shape of their first strokes where system offers five different strokes. It is usually enough to use the first stroke that gives sufficient small amount of characters to find the requested one. If some stroke count includes a large number of characters two first strokes are used.

There are also characters with ambiguous stroke counts because there are multiple ways to write their components. A good indexing system has cognisance of this and have multiple entries in the index for those characters.

## 3.2 Linguistic aspects

### 3.2.1 Text Input

Text input is quite complicated in Chinese. You must be able to enter both Chinese characters and Latin alphabet because there can be Latin alphabet in the middle of Chinese text, e.g. common abbreviations.

Because Chinese language has a huge number of characters it is not possible to have a keyboard with a separate key for each character. Therefore there are a number of methods for entering characters with a QWERTY keyboard. Some methods let also enter characters with a device with much less keys, e.g. cellphone while sending sms-messages.

Input software applications are called IMEs (input method editors) and they run as a separate process that is integrated to application.

Text is input in two phases: in the first phase user types keyboard input which the computer interprets, depending on the input method, to a list of candidates that refer to characters that are mapped to the input string. In the second phase user selects one candidate from the list. If there are more candidate characters than can be shown at a time, user can request for more candidates.

Input method can be based on pronunciation, character structure or encoding or it can use multiple criterias. Pinyin method is based on pronunciation and a person who is familiar with

pinyin will learn the pinyin method very quickly. Wubizixing method is based on a character structure and it is not easy to learn but, once learned, it is much faster than any phonetic method. There are also methods that use both pronunciation and structure. In this paper I will introduce only a few input methods.

### 3.2.1.1 Input by pronunciation

There are two units by which input can be converted to Chinese characters, a single character at a time or a string of two or more characters.

User enters text in two phases: the first phase is typing a keyboard input that the computer interprets to a list of candidates and the second phase is selecting one candidate from the list or requesting for more candidates.

If we convert input to characters one by one there will be a lot of candidates to choose for each character but if we enter input for two or more characters at a time there will be much less candidates and entering will be much faster. For example, if we want to write 汉字 (han4zi4) that means 'Chinese character', the difference is obvious.

Selecting characters one by one; first 'han', then 'zi'.

'han' gives you 21 candidate to choose:

撼 喊 酣 蚶 函 涵 汗 干 汗 罕 鼾 韩 翰 瀚 颌 寒 含 旱 捍 悍 焊

from where you choose the seventh character.

'zi' gives 27 candidates:

子 字 仔 籽 渍 自 滓 髭 齍 姿 咨 恣 资 贻 訾 紫 兹 滋 孳 缁 淄 淄 淄 姊 梓 孜 吱

from where you choose the second character. They both would have less choices if tone numbers were used after pinyin. 'han4' would give 12 candidates and 'zi4' would give 4.

If you are in a mode where you can enter input for two or more characters at a time, entering 'hanzi' gives the right choice 汉字 at once. This example is produced with NJStar Chinese WP 5.01 product.

### **Input methods by pronunciation**

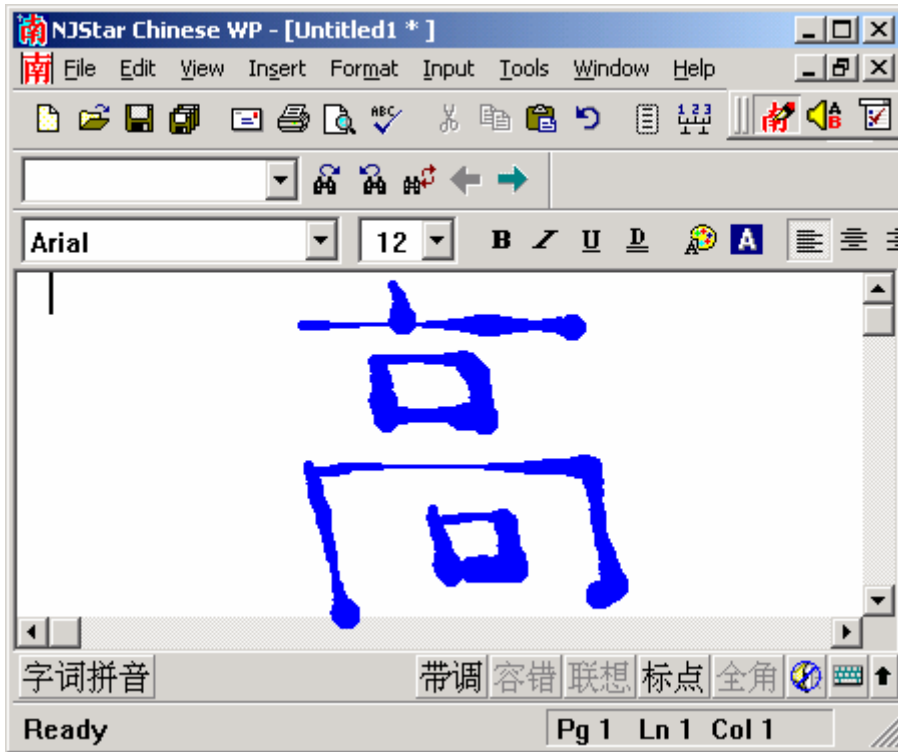
Input by pronunciation is most frequently used input method. It can use Pinyin or Zhuyin. Pinyin is based on Latin alphabet and Zhuyin has its own symbols that are elements of Chinese characters.

### **Pinyin Input method**

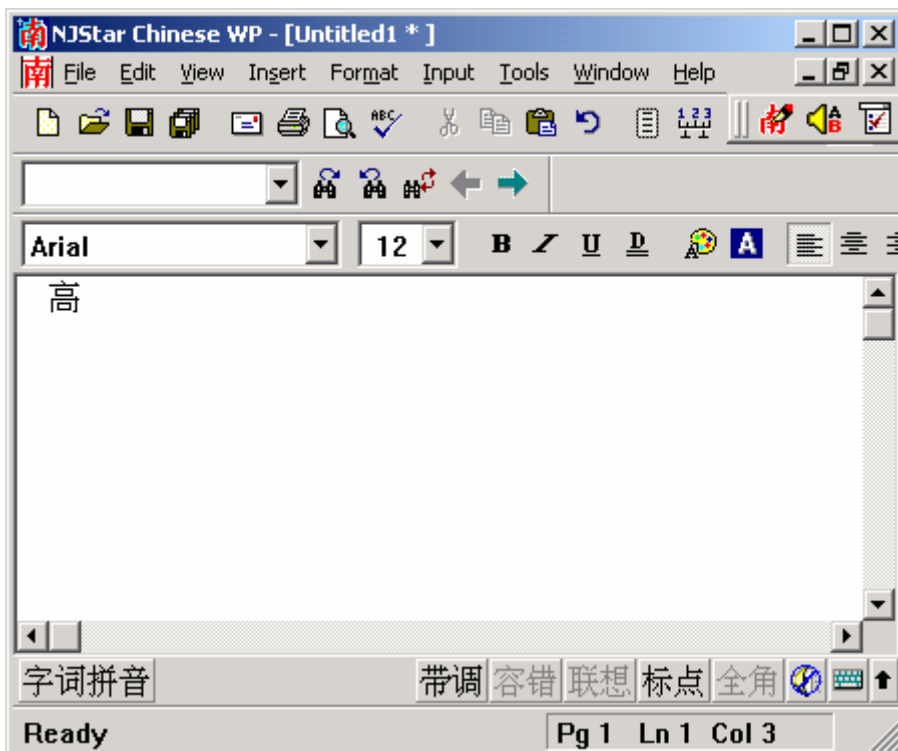
There are three types of Pinyin input: Full Pinyin, Double Pinyin and Half Pinyin. Full Pinyin means writing the pronunciation of a Chinese character with Pinyin just as it is. This requires one to six keystrokes per character.

In Double Pinyin certain letter combinations of Pinyin are replaced by one letter according to specific rules. Double Pinyin first divides Pinyin reading into two parts and then uses replacing characters of letter combinations. This requires one or two keystrokes per character. For example SHUANG in Full Pinyin is IH in Double Pinyin: SHUANG is divided into two parts, SH and UANG. In Double Pinyin letter combination definitions SH=I and UANG=H and so SHUANG can be entered by two keystrokes I and H.





As soon as I finish drawing the target character 高 will appear on the place where the cursor was standing.



### Input by radical

Characters are composed of smaller units called radicals and other elements that cannot serve as radicals. One radical in a character is used for indexing purposes. There are altogether 214 radicals and they all have names and numbers and a certain number of strokes.

Characters are entered in two phases: in the first phase one enters the name or number of the radical of the character and computer returns a list of characters that have the requested radical. In the second phase one selects one candidate from the list or requests for more candidates.

Sometimes indexing radical is not obvious. In that case one can count the number of strokes and use another method.

### **Input by number of strokes**

Except a few characters, they all have a unique number of strokes. In this method user is entering the number of strokes of a character and computer returns a list of candidates.

### **Input by stroke shapes**

Wubizixing method is based on a character's overall structure and its first, second and final stroke. It is not easy to learn but once learned, it is very fast method because every character can be written with at most 5 keystrokes.

Wubihua method is based on the first five strokes and it needs only five keys and can be used on a numeric keyboard. It is easy to learn but very slow to use because every input gives too many candidates to choose from.

### **3.2.1.3 Input by encoding**

Input by encoding-method is based on fixed encoded values for each character. It means that one can enter a character code and get the only matching character without candidates. This method is very fast if you remember codes, most often in hexadecimal values.

## **3.2.2 Text orientation**

In Simplified Chinese almost all publications are printed horizontally from left to right. Vertical direction starting from the right side is used only in advertising and some special cases. Software user interface has text in its titles, menus, input property names and input area, action buttons and user messages. Besides user interface, product has also printed or digital documents like installation guide, system and maintenance documents, training material etc. Both user interface and other material of the software products can be written in horizontal direction.

## **3.2.3 Word separation, word wrapping and hyphenation**

In western languages words are separated by a space between. Chinese language does not separate words at all but it is up to a reader to understand what characters standing next to each other belong to one word. One can only trust that characters in different sides of punctuation marks do not form a word. Hyphenation is not a problem in Chinese because the language does not have inflections as the case is in Western languages and e.g. in Japanese.

There are also a few characters that should not begin or terminate a line. Most of them are punctuation and enclosing characters. Characters that may not begin a line are punctuation marks (e.g. 、 。 , . : ; - ? !), symbols that need to be close to previous text (e.g. % ‰ ° ℃), closing quotes (‘ ” ] } > »), closing parenthesis, closing bracket etc. (} ) ) ] 】). Characters that may not terminate a line are opening quotes (‘ ” [ [ [ < 《), symbols that need to be

connected to the following text (e.g. ¥ @ \$ £ #), closing parenthesis, closing bracket etc. ( { ( [ ] ).

### 3.2.4 Punctuation

In Chinese some punctuation marks are the same that are used in Western languages but they occupy the same size square that is reserved for characters around it. The following list has punctuation marks that are used both in Chinese and for example in English.

- a comma ,
- a semi-colon ;
- a colon :
- an exclamation mark !
- a question mark ?
- parenthesis and brackets ( ) [ ] { } 【 】

There are also some differences:

A period: In Chinese there is usually no period (.) used in the end of a sentence but a small circle (。). The period is used sometimes in technical texts.

Quotation marks: There are four kinds of quotation marks: "" and "" as in English and besides those also 『 』 and 「 」.

A caesura: It expresses a short pause and it is marked (、). It is used as a serial comma and used between equal items in a list.

An ellipsis: In Chinese ellipsis is expressed with six dots, taking size of two characters (two groups of three dots) (……).

A dash: A dash is expressed with (—) and it takes size of two characters.

Hyphens: There are four kinds of hyphens that take space of 0.5 to 2 characters. (—) takes one character space, (——) takes two characters space, (-) takes half character space and (-~) takes one character space.

Separating dot: A separating dot (·) is used to separate characters in a name of a foreigner or volume and chapter names of a book. For example 列奧納多·達·芬奇 is 'Leonardo da Vinci' and 中国大百科全书·物理学 is 'Chinese big encyclopedia, Physics'.

Book title marks: While Western languages use quotes around book titles, Chinese has special marks to that: 《 》 and 〈 〉.

A proper noun mark: If a noun word is underlined, it means that it is a proper noun. It is used occasionally in teaching material.

### 3.3 Cultural aspects

#### 3.3.1 Numerals

Both Arabic numerals and Chinese characters meaning numerals are used in China. Arabic numerals are used in mathematics and many other occasions but Chinese characters are mostly used if there are numerals among text. If Chinese characters are used they do not have spaces separating numbers from the text. There are also specific characters that are used for numerals in financial documents.

'Zero' is expressed with an Arabic numeral 0, a Chinese zero 〇 (ling2) or a Chinese character 零 (ling2). 'Two' can be expressed with an Arabic numeral 2 or two different Chinese characters, 二 (er4) or 两 (liang3). Their usage has specific rules. The most difficult difference in Chinese numerals compared to Western languages is the way how units are grouped. In Western languages a unit has a new name after each thousand, in Chinese the unit changes after ten thousand: after 千 (qian1)=thousand comes 万 (wan4)=ten thousand 10E4 , 亿 (yi4)=a hundred million 10E8, 兆 (zhao4)=a million million 10E12.

Cardinal numbers in Chinese characters:

〇	ling2	zero
零	ling2	zero
一	yi1	one
二	er4	two
两	liang3	two
三	san1	three
四	si4	four
五	wu3	five
六	liu4	six
七	qi1	seven
八	ba1	eight
九	jiu3	nine
十	shi2	ten
百	bai3	hundred
千	qian1	thousand
万	wan4	ten thousand
亿	yi4	a hundred million (wan4 wan4)
兆	zhao4	a million million (wan4yi4)

Ordinal numbers are formed by adding a character 第 (di4) before the numeral, e.g. 第三 (di4san1) means the third. There are also nouns with which one uses ordinals in English and Finnish but cardinals in Chinese, e.g. the 3<sup>rd</sup> floor is in Chinese 三楼 (san1lou3).

Twenty and thirty are normally written 二十 and 三十 but they also have shorthand characters

廿	nian4	twenty
卅	sa4	thirty

There are also different characters used for numbers in financial documents.

零	ling2	zero
壹	yi1	one
贰	er4	two
叁	san1	three
参	san1	three
肆	si4	four
伍	wu3	five
陆	liu4	six
柒	qi1	seven
捌	ba1	eight
玖	jiu3	nine
拾	shi2	ten
佰	bai3	hundred
仟	qian1	thousand

Digits in telephone numbers are in groups of three.

If it is a question of a year the number can be expressed in several different ways: For example a year 2005 have the following forms: 2005, 两〇〇五, 二零零五, 两零零五

If it is not a question of a year, 2005 is 两千零五 and 2000 is 二千 or 两千.

A negative mark is a hyphen (-) and a decimal separator is a period (.). In case Arabic numerals are used the thousand's separator is a comma (,). Fractions and percentages can also be expressed in Chinese characters.

### 3.3.2 Date and Time

In the Chinese culture things are usually expressed from a bigger unit to a smaller one. That is why dates are always in form year-month-day. A date can be expressed in Arabic numerals or Chinese characters. If Arabic numerals are used the separator can always be a hyphen (-). If Chinese characters are used year, month and day can be separated by Chinese characters that mean a year 年 (nian2), a month 月 (yue4) and a day 日 (ri4) or 号(hao4).

A year in Arabic numerals can be expressed in two or four digits. With Chinese characters all four characters are expressed. If a year has zeros, the year expressed in Chinese characters use Chinese zero, not Arabic one. Year 2005 can be expressed in several ways e.g.

05, 2005, 2005 年, 二〇〇五年, 二零零五年, 两零零五年.

A month in Arabic numerals can be expressed in one or two digits. Months can also be expressed with their Chinese names that consist of a number of the month and character that means a month 月 (yue4):

一月, 二月, 三月, 四月, 五月, 六月, 七月, 八月, 九月, 十月, 十一月, 十二月

A day in Arabic numerals can be expressed in one or two digits. Days can also be expressed with Chinese characters that consist of a number of the day and a character that means a day 日 (ri4) or 号 (hao4), e.g.

一日, 二日, ..., 三十日, 三十一日

Morning and afternoon, if they want to be expressed, are expressed with 上午 (shang4wu3) for morning and 下午 (xia4wu3) for afternoon. They can be a part of an expression of time or as themselves expressing the whole morning or afternoon.

As a conclusion, the date of September the 17<sup>th</sup> in 2005, can be in Chinese in the following ways:

2005-09-17

05-09-17

2005-9-17

05-9-17

2005年9月17日

二〇〇五年九月十七日

二零零五年九月十七日

两零零五年九月十七日

Time can be expressed with Arabic numerals or Chinese character, depending on the situation. If time is expressed in Arabic numerals, the separator between hours, minutes and seconds is always a colon (:). If it is expressed with Chinese characters, the hour separator is 点 (dian3), the minute separator is 分 (fen1) and the second separator is 秒 (miao3).

Also expressions 半 (ban4) that means 'a half' and 刻 (ke4) that means 'a quarter' can be used. As an example, time 18:12 can be expressed in several different ways e.g.

18:12

6:12 下午 (xia4wu3)

六点十二 (liu4dian3shi2er4)

六点十二分 (liu4dian3shi2er4fen1)

下午六点十二 (xia4wu3liu4dian3shi2er4)

十八点十二分 (shi2ba1dian3shi2er4fen1)

But you can never say 6:12 PM.

If all units are expressed they must be in the following order: Year – month – date - day of the week – morning or afternoon – hour:minutes:seconds.

### 3.3.3 Currency

The Chinese currency is Renminbi, RMB that is called 元 (yuan2). Its symbol is ¥. It has two smaller units called 角 (jiao3) and 分 (fen1). One Yuan is ten Jiaos and one Jiao is ten Fens. In everyday speech Yuan is called 块 (kuai4) and Jiao is called 毛 (mao2).

A period (.) separates smaller units from Yuan and number of desimal digits is two. ¥ is set before the amount of money e.g. ¥143.22.

A negative sum of a money is expressed with a hyphen (–) in front of ¥ e.g. -¥143.22.

If the amount is bigger than 999 Yuan, the thousand-separator is a comma (,) as the case is with other numbers.

The amount of money can also be expressed with Chinese characters representing numerals, e.g. 100 元 is the same as 百元.

### 3.3.4 Measurement Units

The SI (metric) system is used in China. The mandatory GB 3100-1993 standard is equivalent to ISO 1000:1992 and consists of SI units and recommendations for their usage. GB 3101-93 standard is equivalent to the ISO 31-0:1992 and has general principles concerning quantities, units and symbols in scientific and educational documents.

There exists also Chinese measurement units and if they are used, there might be a need for conversion tables. In the following there are conversion tables for units of length, area, weight and capacity.

#### Units of Length:

1 寸 (cun4)	= 3,333 cm	
1 尺 (chi3)	= 33,33 cm	= 10 cun
1 丈 (zhang4)	= 3,333 m	= 10 chi
1 引 (yin3)	= 33.333 m	
1 里 (li3)	= 500 m	
1 公里 (gong1li3)	= 1 km	

#### Units of Area:

1 平方英尺 ping2fang1ying1chi3	= square chi	= 1/9 m <sup>2</sup>	= 11,111 dm <sup>2</sup>
1 亩 (mu3)	= 60 square zhang	= 1/15 hm <sup>2</sup>	= 6.666 ares = 2000/3 m <sup>2</sup>
1 顷 (qing3)	= 100 mu	= 6.6667 hectares	
1 平方里 (ping2fang1li3)	= 1 square li	= 25 ha	

#### Units of Weight

1 钱 (qian2)	= 5 g	
1 两 (liang3)	= 50 g	= 10 qian
1 斤 (jin1)	= 500 g	= 10 liang

### Units of Capacity

1 升 (sheng1)	= 1 liter
1 斗 (dou3)	= 10 liter
1 石 (shi2, dan4)	= 100 liter

### 3.3.5 Icons, Symbols and Graphics

If a software is meant to be easily localizable one must be aware that if there are graphics in messages or documentation or any part of the product one must be very careful when it is a question of people in the picture: there must not be any reference to race, colour of the skin, ethnic background, gender, clothing or any visible parts of a body.

There are several hand signs that are used in China. For example the thumb and the second finger forming a circle means in China 'zero', not 'ok' as it does e.g. in the US. The second and the third fingers crossed means number 10 and an open palm faced to another means number 5 not 'no' or 'stop'. If one holds up one's little finger it expresses dissatisfaction. The second and the third fingers up mean 'two' not e.g. 'victory'. So, to avoid confusion, it is better to avoid using hands in any picture.

Also some political symbols are sensitive in China. The flags of Taiwan and Japan are not welcome because they remind of historical or current political issues of the country. Also if there is a map of China and Taiwan is not included it is an offensive gesture to China.

Pictures of flowers, animals and fruit are sometimes used as decorations for example on CD-material and Internet portals. So many flowers, animals and fruit have a strong symbolic meaning that they should be used only when it is really a question of that particular flower, animal or fruit. Also dragons fall into animals. If you really have to use some animals, fishes and bats are harmless. But never use them in groups of four because numbers have also symbolic meanings and number four refers to death. It is better to use even numbers but never number four.

Any abbreviation that is used in Western countries and 'are known by everyone' should be forgotten. Each abbreviation that one wants to use must be opened first.

In China a tick can be used as a 'checked' symbol though in Finland it means 'error'. Traffic signs are harmless symbols and if one wants to express a 'stop', the red octagon is used also in China.

### 3.3.6 Colours

Red and white colours have meaning that one should be aware of. Red is a colour of happiness and a good luck but don't write text on a card with red colour! In some parts of China it refers to death.

White is a traditional colour of mourning but is nowadays also used in western style wedding instead of red colour. A small picture with pure red and white colours together refers to the flag of Japan and should be avoided.

Yellow colour has been devoted to the emperors of China and it still connotes property and good luck.

To avoid unwanted effect it is better to let a native Chinese person check all used symbols and colours one by one.

### 3.3.7 Local conventions

The heritage of the Confucian thinking has reflections in everyday life in China. The bigger unit is more important than the smaller one and public is more important than private. This is conveyed in many occasions. For example, person name is always written family name first, then the forename. Also the way how addresses are expressed in China: first the country, then province, city, street, last name, first name and the title.

The paper and printed matter size standard GB/T 148-1997 is recommended in China. It is not totally equivalent to ISO 216 standard that is the most common in the world and is used also in Finland. Both A and B series are used and while drawing up a layout of the output in the application one must keep in mind that the most often used paper size can be A4 (210x297) but also B5 (176x250).

The Gregorian calendar has been the official calendar from 1912 but all traditional Chinese festivals are based on the lunar calendar. The most important festival is the New Year festival that is celebrated between the 21<sup>st</sup> of January and the 20<sup>th</sup> of February. During the New Year festival almost all people have about one week vacation. Also the 1<sup>st</sup> of May celebration and the anniversary of the founding of the PRC (People's Republic of China) celebration in the 1<sup>st</sup> of October are big festivals when people have several days off. The autumn semester in schools does not end in the end of year but in January before Chinese New Year.

If some part of an application refers to geographic regions of the country one should notice that there are three kinds of regions directly under the central government and they are not hierarchically within each other. The first group is the five autonomous regions Inner Mongolia, Ningxia, Xinjiang, Guangxi and Tibet. The second group is the 23 provinces and the third one is the areas of Beijing, Shanghai, Tianjin and now also ChongQing which is the most populous area with more than 32 million inhabitants. Therefore the big cities of Beijing, Shanghai and Tianjin belong to two groups, the second level governmental areas and also to cities.

Rules, methods, procedures and formulas of daily life also differ and must be found out according to the target subject. This has also effect on output forms to be constructed. Taxes, pension and even health insurances and other formulas dealing with money may change in different parts of the country.

Applications must follow laws, regulations and standards of national and local levels. To promote an integration of applications all related existing code systems of the target field should be studied and the highest standard ones should be used in the application.

According to 'The basic functional norm of Hospital information system (HIS)' that is determined in 2002 by the Ministry of Health, the Office of the Leading Group for Informatization Work there are three basic laws and regulations that must be followed when designing a database:

- The security protecting regulations of the computer information systems of the People's Republic of China.
- The confidential law of the People's Republic of China
- The standard of the computer security law in China

There are also other mandatory standards e.g. for security techniques (entity authentication, digital signature and data integrity mechanism using a cryptographic check function) and information processing vocabulary for fundamental terminology. More about standards in Jiechen Jiang's document 'Export HIS Project – Standardization'.

## 3.4 User interface

### Geometry

While building a user interface one must consider its position and size on the display. Written Chinese characters can never be as small as written Latin alphabet and this relates both to the height and the width of characters. The width of the text is not so critical in localization because Chinese language needs much less characters than Western language needs alphabet to express a same thing. This relates to all parts of the application that are called by the user interface like pop-up windows, dialog-boxes, message boxes and other parts of the application if they are opened in a separate window. This is why presentation information (position, size, font, colour, intensity, text orientation) is so important to isolate from the code.

For the same reason also property names and text input controls on data input form must be higher than in Western languages. If it is a question of a date input the order must be changed to year-month-day. The length of the input control that is for a person name can be shorter in a Chinese application. Most Chinese person names, including both family name and first name, have only two or three characters.

In brief, if there are one or two lines of text in Finnish or English the length is sufficient for Chinese but there is a need for a higher space. If the text is more than two lines in Finnish or English, the space is enough for Chinese because there will be less lines. If there is space reserved for a Japanese text then there is always space enough for a Chinese text.

### Menus and buttons

The direction of menus can be the same what we have in Western applications. But the space reserved for the text in menu title and menu items must be higher.

The text size has impact to the size of a button but also the colour and icon of the button must be carefully considered.

Hot keys can be used with menu items and buttons but they must be separated in a very early phase because the same hot key cannot be used in e.g. Finnish and Chinese application.

### Messages

Message lines in Chinese need higher space and like any localization the text of the message must be in the repository as a whole sentence. It must not be gathered from short pieces of text.

### Other parts of the user interface

Colours and shapes of pointers and cursors may vary. While defining available pointers and cursors it is good to remember what colours and symbols are sensitive in the Chinese culture.

## 3.5 Hardware devices

Whatever hardware devices an application needs one must make sure that there are available drivers that support use of Chinese characters and multibyte character system.

### **Keyboards**

With a Chinese application one must be able to input data that may include Chinese characters, Latin alphabet and Arabic numerals. An English keyboard and keyboard driver can be used if data input method is based on Pinyin transliteration system. Whatever is the input method there must also be an input editor software to convert the inserted data to the sought Chinese character.

If the input method is Zhuyin or Wubizixing or some other that is not available with Latin alphabets it requires its specific keyboard and keyboard driver.

### **Displays and printers**

The resolution power of the display must be sufficient to enable building applications with reasonable size forms. As told in the Geometry chapter, Chinese characters need higher space compared to Latin alphabet. If anti-aliasing technique can be used, requirements are not so hard.

Displays and printers must be able to show used characters and Postscript fonts. The DPS (Display PostScript) uses PostScript imaging model and language to generate on-screen graphics and it can provide characters as WYSIWYG on the screen.

Fonts can be downloaded statically to the device memory but there must be space enough to load new fonts as a dynamic download or incremental load if only requested new characters are loaded to the device memory.

It is also possible to keep fonts only on the hard disk and build bitmaps of appropriate resolution in the host and send them to the Post Script device.

## **3.6 Software services**

Every application uses at least operating system and file system, nowadays also networking system. An application for Chinese market needs also an input method editor. They all must be Unicode capable and be able to handle characters in multibyte character system and have required conversion tools between character encodings. This relates also to third-party components that are used in the product. When there is data exchange between different parts of the product one must ensure that the contents of the data does not change. To ensure this data should be exchanged in the internal form and all parts should have functions available to change data from external to internal form and vice versa.

### **Operating System**

When developing a Chinese application it is not necessary to use a Chinese operating system but use an appropriate locale that assures language support. Operating systems have various locales for different languages and regions. If an application uses operating system level function calls whose reply depends on the locale it must be capable to handle the reply before it can be used as input data to the next function call. A good example of this is a date that has several different forms. There must be functions available to change data from external to internal form and vice versa.

### **File Systems**

All file systems that are a part of a software product at any position must be Unicode enabled. There must be conversion tools and support also for four-byte characters. If application data is transferred from one platform to another the endian issue must also be noticed. An endianness

depends on the computer architecture and means the order in which the byte of a multi-byte string is presented. Typically Windows uses Little Endian and most Unix systems use Big endian. If a name of a file is given by a user, it must be prevented to give a name that the target software cannot use.

### **Interprocess Communication**

Interprocess communication means data exchanging between separate processes. To allow this data storage format must be known by both processes to be interpreted unambiguously. If the transferred data is locale specific it has to be converted to universal format that the receiving process can understand. This conversion can occur in the transmitting process before sending data out or in the receiving process while reading it.

### **Input method editor**

To be able to input Chinese characters there must be an input method editor. Some operating systems have built-in input method editors for various methods but there exist also external software for that purpose.

## **3.7 Support**

Building an internationalized software product and a localized version of it is just the beginning. The localized product needs also marketing and packing. It must be installed before it can be used and users may also need some training. These all layers need support that can be e.g. printed material, on-line help or web pages. The most of it must also be localized and it does not mean only translating texts but also changing all culturally sensitive parts.

In the second section 'Globalization' there was a list of parts that should be isolated from the software core to ease its localization. This list should be kept in mind also when planning all supporting material, particularly when you are planning on-line help, web pages or tutorials. While producing any material, printed or on-line available, one must pay attention to what was said previously about cultural aspects.

### **Installation wizard**

Sometimes it is enough to have installation and configuration instructions only in English but that is not always the case. If an installation wizard is used it must also be localizable and localized or the product must have a separate version of the wizard for the Chinese installation.

### **Printed documentation**

A software product may have printed material like User documents, System documents, Configuration guides, Installation guides, Quick reference guides, 'Getting started' material, other training material etc.

The used writing style must be neutral polite written language. A localized document is not only a document written in Chinese but the documenting system must also allow replacing symbols, graphics and other sensitive parts with localized versions.

There must be separate layouts for paper sizes A4 and B5. The writing direction is horizontal from left to right and when the cover sheet is on the top, the spine is in the left side.

### **On-line help, web pages and training material**

If a software product has a built-in help system, its triggering and executing system must also be localized. Like with a separate on-line help, the geometry of its pop-up windows must be customized to Chinese language and all used examples, graphics and symbols must respect Chinese culture and conventions. This also relates to web pages. If the software product has on-line training material its sample data must be in Chinese.

### **Marketing material**

A localized product needs also marketing and packaging. All marketing brochures and other marketing material must be localized or designed separately to Chinese markets. It is very a sensitive thing because advertising may be visible also to those who are not in the target group.

## **3.8 Other parts to be localized**

Though all parts mentioned above are localized there is still something that must be localized to Chinese. As soon as the product is developed further, it will have new versions and the updating system must also be localized or it must have been build locale-independent.

Official documents like licences, terms of delivery and copyright information must also be translated to Chinese.

And, though this is not a software issue, if the software product is sold with a device and the device has text or pictures painted on, texts and pictures must also be changed to Chinese ones.

## 4 Software development products

Before you start building up a software product that would be used in China you must be sure that tools you are going to use will support all specific needs that Chinese language used in the China will bring with. In this chapter I'm going to discuss some operating systems, a few databases and a couple of development tools.

The locale model is a widely used way to build international software. A locale is used to define used attributes that are specific to language and region like character set, number of bytes per character, date and time formats, currency, numeric formatting etc.

A name of a locale is usually composed of a language identifier and a region identifier. The used character set may be the third identifier but it may also be included into the language and region information. For example zh\_CN refers to Simplified Chinese as a language and the other attributes specific to China and en\_US refers to English language and other attributes specific the USA.

Abbreviations of languages and regions in locales use often standard names. There are two ISO standards for language name codes. The ISO 639-1:2002 has language names in two characters and the ISO 639-2:1998 has them in three characters. The ISO 639-3 is now under development and its aim is to cover all known languages. The Chinese has the following language codes:

- ISO 639-1:2002 Simplified Chinese = zh
- ISO 639-2:1998 Simplified Chinese = chi/zho

Country codes are defined in the ISO 3166 where there are also two and three character abbreviations:

- ISO 3166 Alpha-2: China (People's Republic of China) = CN
- ISO 3166 Alpha-3: China (People's Republic of China) = CHN

In this document we have a few examples of products that can be used in Chinese software. These are not the only available ones but examples of those which have adequate support. Before you decide what products you are going to use, make sure that those tools have required capabilities e.g. they are Unicode-enabled and are able to handle four byte characters, they have fonts for GB18030-2000 character set, they have needed conversion tools and that they have input method editor available.

Operating systems and databases must support GB18030-2000 standard and the easiest way to conform this is to enable using Unicode. Because GB18030-2000 has double and four-byte characters, also four-byte support must be available.

**The information in the following chapters in this section 4 is gathered from producers' web pages and some sections are just copy-pasted.** Many products support Simplified Chinese but there is no mention about GB18030-2000 standard or four-byte support but it is mentioned in a 'how to fix it'-list or somewhere else that might give an idea of supporting GB18030-2000. Therefore the GB18030-2000 support of every product must be made sure before selecting that product to development work. Products that are mentioned in the following chapters are those whose information related to this topic was able to find from their web-pages with reasonable efforts.

This section is not very detailed because features of products are changing all the time. New products with adequate support are coming and when it is time to make a decision about development tools, available products should be studied again.

I also want to emphasize that the following products are not the only ones with adequate support at the moment, they are just examples.

### 4.1 Operating systems

The requirement to build up a Chinese system needs a localized operating system or there is a Chinese support added on. If there is a Chinese operating system it has the multi-byte character handling in the system level and menus and dialogs of the operating system are written in Chinese.

It is able to install a supporting Chinese package on the current operating system. In this case I would recommend using the original version of the operating system and add a Chinese package on that. If you add a Chinese package to an operating system e.g. with a Finnish locale you might meet more problems that cannot be known in advance.

#### 4.1.1 Windows

The following Microsoft Platforms have been approved for use in the PRC:

Windows XP	Approved by the Chinese testing agency.
Windows 2000	Approved by the Chinese testing agency.
Windows NT4	Exempt from the law because released before the standard was published.
Windows 98	Exempt from the law because released before the standard was published.
Windows 95	Exempt from the law because released before the standard was published.
Windows Millennium	Does not conform and may no longer be sold in China.

The following Windows system attributes are related to the Peoples Republic of China:

Locale code id:	2052 (hex 804)
Country code:	86
Country name:	CTRY_PRCHINA
Language name:	LANG_CHINESE (SUBLANG_CHINESE_SIMPLIFIED)

#### **Windows locale system**

In Windows there are a number of different locale settings: the user locale, input locales and the system locale. The supported and installed languages are known as "language groups" in Windows 2000. A language group is a set of all the keyboard layouts, IMEs (input method editors), fonts, language packs and codepage translation tables needed by the system to support the given group of languages. The list of installed language groups controls which user, input, and system locales can be selected by a user.

While installing Windows 2000 the Western Europe and United States language group is the default and it cannot be removed. There are sixteen additional language groups to be installed where Simplified Chinese is one of them.

In Windows XP language groups are cathered into three language collections where Chinese, Japanese and Korean belong to the East Asian language group collection.

**The user locale** is a setting which determines the formats used by default to display dates, times, currency, and numbers, and the sorting order of text. A user locale is specified for each account created on a machine. If you want to change your user locale in Windows 2000, take the following path: Start | Settings | Control Panel | Regional Options and then in the General tab select your preferable locale from 'Your locale (location)' drop-down list.

While using Windows XP take the following path: Start | Settings | Control Panel | Regional and Language Options and then in the Regional Options tab select your preferable locale in the 'Standards and formats' frame.

**Input locales** are pairings of an input language with an input method (which might be a particular keyboard layout, an Input Method Editor, or speech-to-text converter, for example). Specifically, an input locale describes the language being entered and how it is being entered. It is possible to install multiple input locales for each account and user can switch between them when entering text. This is the method to write multilingual documents in languages that has characters from different character sets (like in this document both English and Simplified Chinese).

There are several paths to go to set input locales in Windows 2000, one of them being the following path: Start | Settings | Control Panel | Text Services where you can select one language to be the default language when you start your computer. Additional input languages can be set in the 'Installed Services' frame.

This is how you can set input locales in Windows XP: Take path Start | Settings | Control Panel | Regional and Language Options and in the Languages tab in the frame 'Text services and input languages' go to Details where you can select one language to be the default input language when you start your computer. Additional input languages can be set in the 'Installed Services' frame.

**The system locale** determines which ANSI, OEM and MAC codepages and associated bitmap font files are used as defaults for the system. These codepages and fonts enable non-Unicode applications to run as they would on a system localized to the language of the system locale. Windows 2000 supports system locales for any supported locale on all language versions. System locale is pre-selected by the language version but can be modified via path: Start | Settings | Control Panel | Regional settings.

In Windows operating system there are five locales for Chinese language: the PRC, Hong Kong, Macau, Taiwan and Singapore have different locales.

	<b>Locale ID (dec)</b>	<b>Region abbrev</b>	<b>Language abbrev</b>	<b>ANSI code page</b>
<b>the PRC</b>	2052	CHN	CHS	936
<b>Hong Kong</b>	3076	HKG	ZHH	950
<b>Macau</b>	5124	MCO	ZHM	950
<b>Taiwan</b>	1028	TWN	CHT	950
<b>Singapore</b>	4100	SGP	ZHI	936

**GB18030-2000 standard**

The encoding standard of Windows operating system is UTF-16 Little Endian. Microsoft code page identifier for GB18030-2000 is 54936. It does not have a System Locale, only a code page identifier to allow for conversions to and from Unicode.

There is no Command Shell support for GB18030-2000. Only applications performing windowed mode I/O can use this encoding.

Only the latest Microsoft Platforms support GB18030-2000:

Windows XP	Yes (with add-on)
Windows 2000	Yes (with add-on)
Windows 95, 98 & Millennium and NT4	No
Internet Explorer 6.0	Yes
Internet Explore 5.5 and earlier	No

Internet Explorer version 6 supports GB18030-2000. It recognizes GB18030-2000 as a character set and will display such web pages, but only when it is run on Windows XP. Older versions of Internet Explorer do not support GB18030-2000.

To get the best functional GB18030-2000 Windows platform you have to do the following:

1. Install Windows XP, English or Simplified Chinese.
2. If you have English Windows XP, install East Asian language support (Start | Settings | Control Panel | Regional and Language Options).
3. Install Microsoft GB18030-2000 Support Package. If you have Windows XP Simplified Chinese the package is on a separate CD-ROM. Otherwise you can download it from [www.microsoft.com/china/windows2000/downloads/18030.asp](http://www.microsoft.com/china/windows2000/downloads/18030.asp). This package includes
  - a. SimSun18030.ttc A font file for GB18030-2000. It does not have the complete GB18030-2000 repertoire. It is missing the entire CJK Extension B set and many other characters but it does have CJK Extension A characters.
  - b. c\_g18030.dll A system library that extends the MultiByteToWideChar and WideCharToMultiByte program interfaces to support GB18030-2000.
  - c. Gbunicnv.exe An end-user applet that converts between GB18030-2000 plain text and HTML files and Unicode.
  - d. Ms4bsp.dll A system library that provides a few program interfaces useful for applications that use GB18030-2000 as their character type.
4. Install Microsoft Office XP Simplified Chinese Version with Office Proofing Tools, OR Microsoft Office XP English version with Office Proofing Tools to get the Enhanced Unicode IME and SimSun (Founder Extended) font where you get the required input editor and the CJK Extension B glyph set.

The Microsoft 4-byte Character Set Encoding Support Package (MS4BSP) provides six functions that support multiple byte encodings that can be up to four bytes long. The API set is drawn from the set of WCHAR (Unicode) functions provided by Windows 95/98/Me. In each case, the function name is identical to the ANSI and WCHAR function except with an 'L' suffix instead of 'A' or 'W'.

- ExtTextOutL
- GetTextExtentExPointL
- GetTextExtentPoint32L

MessageBoxL  
MessageBoxExL  
TextOutL

The function parameters are identical to the 'A' version of the interface. The package is designed to expedite conversion of a code page 936 based application to GB18030-2000 or other four-byte encoding. The Windows XP and Windows 2000 implementation of MS4BSP is delivered as a single dll, ms4bsp.dll. Each function invokes the MultiByteToWideChar function to convert any multibyte string inputs to UTF-16. It then invokes the 'W' version of the function and returns the output parameters that function. The dll assumes the fonts, IMEs and registry settings for the encoding are present on the system.

### **Other conversions tools**

NJStar Communicator is a tool with code converters for CJK (Chinese, Japanese and Korean) languages and it is also available as a dll. It is a product with CJK Viewer and input method editor for Chinese, Japanese and Korean languages. It supports most CJK Coding Standards.

NJStar Universal Code Converter supports e.g. the following encodings:

GB	- Chinese GB2312-1980
GBK	- Chinese GBK (including GB2312-1980)
GB2	- Chinese GB18030-2000
HZ	- Chinese GB-HZ (GB2312-1980 in 7 bit encoding)
UCS	- Unicode UCS2/UTF16
UT8	- Unicode UTF8
UT7	- Unicode UTF7
UT5	- Unicode UTF5
IDN	- Unicode IDN (such as xn--6kry05bgnc.com)
URT	- Unicode RTF (such RTF created by Word 2000/XP)
UHT	- Unicode HTML (such as &#21335;)
UCX	- Unicode UCS4/UTF32 Hex (such as \u6975\u661f)

More information about NJStar Communicator and other NStar products at [www.njstar.com](http://www.njstar.com) and [www.njstar.com/communicator/ucc.htm](http://www.njstar.com/communicator/ucc.htm)

### **Text input**

If you have added East Asian language package or Simplified Chinese language as an input locale on Windows XP, Windows 2000 or Windows Server 2003, you can change your input language from the default language to Simplified Chinese on all Windows application if that application does not have an input editor of its own. If a data will be entered by Pinyin an English/US keyboard is sufficient. You switch between languages with LeftAlt+Shift keys by default.

Another tool for input is NJStar Communicator, a product with CJK Viewer and input method editor for Chinese, Japanese and Korean languages. NJStar CJK Viewer enables the user to input and view Chinese, Japanese and Korean Text under any windows application. It has several input methods.

More information and links about building international software on Windows platform is available at [www.microsoft.com/globaldev/default.msp](http://www.microsoft.com/globaldev/default.msp)

### 4.1.2 Sun Solaris

The following information is from Simplified Chinese Solaris User's Guide (January 2005).

#### **Chinese Locales**

In Sun Solaris system a name of a locale is composed of a language identifier, a region identifier and a code set identifier. There are several locales for the PRC: zh\_CN.EUC, zh\_CN.GBK, zh\_CN.GB18030 and zh\_CN.UTF-8. In addition there are five more locales for Taiwan and HongKong that use Traditional Chinese.

In the locale zh\_CN.GB18030 the user interface is in Simplified Chinese, region is the PRC and used character set is GB18030-2000. This enables input, display and print Simplified Chinese text.

#### **GB18030-2000 standard**

Sun Solaris 8 Operating Environment has satisfied the PRC's GB18030-2000 requirements. GB18030-2000 support in Solaris 9 release also includes backward compatibility to previous Chinese codesets, GBK, and GB2312-1980.

Sun has globalization APIs such as mbtowc(), mbstowcs(), and mblen() to help convert GB18030-2000 multibyte strings to wide character format.

The most common conversion command is *iconv* and with zh\_CN.GB18030 locale one can use the following Simplified Chinese code conversion modules:

Code	Symbol	Target Code	Symbol
UTF-8	UTF-8	GB18030-2000	zh_CN.gb18030
GB18030-2000	zh_CN.gb18030	UTF-8	UTF-8
GB18030-2000	zh_CN.gb18030	BIG5HK	zh_HK-big5hk
GB18030-2000	zh_CN.gb18030	BIG5P	zh_TW-big5p
BIG5HK	zh_HK-big5hk	GB18030-2000	zh_CN.gb18030
BIG5P	zh_TW-big5p	GB18030-2000	zh_CN.gb18030

BIG5HK and BIG5P in the previous table refer to Traditional Chinese used in Hongkong and Taiwan respectively.

There are five fonts available for the zh\_CN.GB18030 locale: four TrueType Fonts, FangSong, Song, Hei and Kai from FangZheng vendor. In addition there is Song bitmap font available.

#### **Input editors**

Sun Solaris system has an Input Method Auxiliary Window that supports the following functions:

Input method switching

Input methods properties configuration

Lookup tables for the following character sets:

GB2312-1980

GBK

GB18030-2000

Unicode

User-defined characters

Input method help

Virtual keyboard

The following input methods are supported for the zh\_CN.GB18030 locale:

1. New QuanPin
2. New ShuangPin
3. QuanPin
4. ShuangPin
5. GB18030–2000 NeiMa
6. WangMa Wubi
7. English-Chinese

The first five of them are based on pronunciation, WangMa Wubi is based on character structure. English-Chinese gives an opportunity to insert English words that are mapped to Chinese phrases.

More information on building international software with Sun Microsystems environment: Sun Microsystems information on Solaris 10 International Language Support Collection: <http://docs.sun.com/app/docs/coll/767.3>

### 4.1.3 HP-UX

#### **Chinese Locales**

In HP-UX system a name of a locale is composed of a language identifier, a region identifier and a code set identifier. In HP-UX 11i Version 1.6 release the locale zh\_CN.gb18030, that enables support of GB18030-2000, has been added to available locales. System level support is provided to allow for the input, storage, retrieval, display and printing of the set of characters defined in the GB18030-2000 character set standard.

#### **GB18030-2000 standard**

HP-UX support to GB18030-2000 contains support to Input method, CDE login, fonts, and code conversion. There are two conversion tables for *iconv* command: the first one is between GB18030-2000 and UCS2 and the second one is between GB18030-2000 and UTF8.

There are three fonts that support GB18030-2000: TrueType Font ZYCJKHei and bitmap fonts Song 18 and Song 24.

#### **Input editors**

In HP-UX 11i v2 there is a new input method for the GB18030-2000 character set. The Intelligent ABC Chinese input method is a very powerful and popular input method in the PRC. It is widely used in MS Windows, IBM AIX, Apple Mac OS, and Linux. The Intelligent ABC input

method supports the full GB18030-2000 character set and has been certified by the Chinese standards agency CITS. (<http://docs.hp.com/en/5990-6737/ch10s02.html>)

The input method is based on Pinyin. It is easy to learn and supports stroke input, word input, and user-defined words. The Intelligent ABC input method will support zh\_CN.hp15CN, zh\_CN.utf8 and zh\_CN.gb18030, the three Simplified Chinese locales.

More information about HP-UX internationalization: HP-UX 11.0 - 11i Internationalization Features White Paper: <http://docs.hp.com/en/5991-1194/index.html>

### 4.1.4 Asianux (Miracle Linux, Red Flag Linux, Haansoft)

Asianux is a unified Linux distribution across China, Japan and South Korean. Red Flag Software (China), Miracle Linux (Japan), and Haansoft INC. (South Korea) will distribute and market Asianux without any modifications in each Linux distribution package. New products such as Red Flag 5 Family, Miracle Linux V4.0 and Haansoft Linux 2005 will be based on Asianux and each will be bundled with localized features in each country.

Asianux 2.0 supports Japanese, Chinese simplified and traditional, Korean and English, and provides the optimum environment for using these languages. Also UTF8, EUC, SJIS and GB18030-2000 locales are supported. The locale for the PRC is zh\_CN.GB18030.

A free download of a complete GB 18030 font is available from the Research Grants Council [http://cerg1.ugc.edu.hk/cergprod/static/download\\_font.jsp](http://cerg1.ugc.edu.hk/cergprod/static/download_font.jsp)

## 4.2 Databases

### 4.2.1 Caché

Caché is a product of Intersystems Corporation. The following description of the product is from Intersystems' web page <http://www.intersystems.com/cache/technology/what-is-cache.html> "Caché is a post-relational database that uniquely offers three integrated data access options which can be used simultaneously on the same data: a robust object database, high performance SQL, and rich multidimensional access. No mapping is required between object, relational, and multidimensional views of data, resulting in huge savings in both development and processing time. Caché enables rapid Web application development, extraordinary transaction processing speed, massive scalability, and real-time queries against transactional data."

Caché has full Unicode support and Caché version 5 has support for Simplified Chinese but there is no support for GB18030-2000. Though, it is mentioned in this document because of the interest of one participating company of this project. Caché 5.0.20 is available for the following operating systems:

- HP Alpha OpenVMS 7.2-2, 7.3-1, 7.3-2
- HP Alpha Tru64 UNIX 5.1, 5.1a, 5.1b
- HP-UX 11 (32 bit only)
- HP-UX 11i (64 and 32 bit)
- HP-UX 11i v2 (64 bit for Itanium)
- HP-UX 11i v2 (64 bit for PA Risc)
- IBM P Series AIX 4.3.3 (32 bit only)
- IBM P Series AIX 5.1, 5.2, 5.3 (64 and 32 bit)
- Mac OS X 10.3, 10.4

Red Hat Advanced Server 3.0 (64 bit for AMD, EM64T)  
Red Hat Linux (Intel) Advanced Server 3.0, 4.0  
Red Hat Linux (Intel) Advanced Server 3.0, 4.0 (64 bit for Itanium)  
Sun Solaris (SPARC) 8, 9, 10  
SuSE Linux (Intel) 7.3, 8.0, 9.0  
SuSE Linux (Intel) Enterprise Server 9.3  
SuSE Linux (Intel) Enterprise Server 9.3 , (64 bit for Itanium, AMD, EM64T)  
Windows NT 4 (SP4, SP5, SP6), 2000 (SP3, SP4), XP (SP1, SP2), Server 2003 (SP1)  
Windows Server 2003 (SP1),(64 bit for Itanium)

#### 4.2.2 Oracle

Oracle Database is a product of Oracle Corporation. It has had support for GB18030-2000 from version 10g, Release 1. Release 1 or Release 2 Standard Edition, Standard Edition One and Enterprise Edition are available for the following operating systems which do not all support GB18030-2000 at the moment of writing this document.

AIX5L  
Asianux 2.0 (Red Flag 5)  
HP OpenVMS Alpha  
HP Tru64 UNIX  
HP-UX Itanium  
HP-UX PA-RISC  
IBM z/OS  
Linux Itanium  
Linux x86  
Linux x86-64  
Mac OS X Server  
Microsoft Windows  
Microsoft Windows (64-bit Itanium)  
Microsoft Windows (x64)  
Solaris Operating System (SPARC) (64-bit)  
Solaris x86  
z/Linux

#### 4.2.3 MS SQL Server

Microsoft SQL Server 2005 is a product of Microsoft Corporation. The following description of the product is from Microsoft's web page <http://www.microsoft.com/sql/prodinfo/overview/what-is-sql-server.msp>: "SQL Server 2005 is a comprehensive database platform providing enterprise-class data management with integrated business intelligence (BI) tools. The SQL Server 2005 database engine provides more secure, reliable storage for both relational and structured data, enabling you to build and manage highly available, performant data applications for use in your business."

SQL Server 2005 provides support of GB18030-encoded characters by recognizing them when they enter the server from a client-side application, and converting and storing them natively as Unicode characters. After they are stored in the server, they are treated as Unicode characters in any subsequent operations performed on them.

When you use GB18030 characters, remember that these characters can be used in ordering and comparison operations, but when you use a collation version older than SQL Server 90, comparisons are only based on their code points and not on other linguistically meaningful ways. Therefore, you should be careful when you use GB18030 characters in operations such as ORDER BY, GROUP BY, and DISTINCT, and especially when GB18030 and non-GB18030 characters are included in the same operation. To enable meaningful string comparisons that use GB18030 characters, use the new SQL Server 90 collation version, signified by the 90 suffix added to its name. For example, instead of the Chinese\_PRC collation, use Chinese\_PRC\_90. For more information, see Collation Settings in Setup. Collation Settings in Setup.

### 4.2.4 Sybase Adaptive Server Enterprise

Sybase Adaptive Server Enterprise is a product of Sybase Incorporation. The following description of the product is from Sybase's web page <http://www.sybase.com/products/informationmanagement/adaptiveserverenterprise>: Sybase, Inc. has introduced Sybase Adaptive Server Enterprise (ASE) 15 to meet the increasing demands of large databases and high transaction volumes, while providing a cost effective database management system. ASE 15's key features include on-disk encryption, smart partitions and new, patent-pending query processing technology that has demonstrated a significant increase in performance, as well as enhanced support for unstructured data management. ASE is a high-performance, mission-critical database management system that gives Sybase customers an operational advantage by lowering costs and risks."

Sybase Adaptive Server Enterprise has built GB18030 support to version 12.5.1. Information of operating systems that it supports are not available but the following list has operating systems that are supported by Sybase Adaptive Server Enterprise 15.0:

- HP-UX 11.11 PA-RISC (64-bit)
- HP-UX 11.23 PA-RISC (64-bit)
- IBM AIX 5.2 (64-bit)
- IBM AIX 5.3 (64-bit)
- Red Hat EL 3.0 (x86) (32-bit)
- SuSE SLES 9.0 (x86) (32-bit)
- Sun Solaris 10 (SPARC) (32-bit)
- Sun Solaris 10 (SPARC) (64-bit)
- Sun Solaris 2.8 (SPARC) (32-bit)
- Sun Solaris 2.8 (SPARC) (64-bit)
- Sun Solaris 2.9 (SPARC) (32-bit)
- Sun Solaris 2.9 (SPARC) (64-bit)
- Windows 2000 (x86) (32-bit)
- Windows 2003 (x86) (32-bit)
- Windows XP (32-bit)

## 4.3 Development environment

### 4.3.1 Java

#### Locales

A Java locale is composed of a language identifier, an optional region identifier and an optional variant code. Locale-sensitive classes like `java.text.NumberFormat` and `java.text.DateFormat` use `java.util.Locale` objects to customize how they present and format data to the user.

Language identifiers base on ISO 639 standard and they are set lowercase. Region identifiers base on ISO 3166 standard and are set uppercase. Both standards have two and three letter variants but in Java both identifiers use two-letter codes. Region identifier is an important locale component because `java.text.Format` objects for dates, time, numbers and currency are sensitive to this element. An optional variant code can be added for e.g. operating system, browser for additional functionality that is not possible with just a language and country designation.

In Java the locale for simplified Chinese and the PRC is `zh_CN`. There is also a preconstructed locale `SIMPLIFIED_CHINESE` which refers to the same locale.

Localization is supported at the most basic level by the `ResourceBundle` class, which provides access to locale specific objects, including strings. When your program needs a locale-specific resource, your program can load it from the resource bundle that is appropriate for the current user's locale. In this way, you can write program code that is largely independent of the user's locale isolating most of the locale-specific information into resource bundles.

This allows you to write programs that can be easily localized, handle multiple locales at once and be easily modified to support later more locales. The ability to use multiple `Locale` objects with various `Format` objects provides developers the opportunity to create multicultural and multilingual applications.

#### Character set

The Java programming language is based on the Unicode character set and it has been certified with an A+ rating for GB18030-2000 support.

Sun's J2SE Development Kit 5.0 for all platforms (Solaris™ operating environment, Linux, and Microsoft Windows) and the J2SE Runtime Environment 5.0 for Solaris and Linux support the GB18030-2000 standard. The canonical names for `java.nio` API, `java.io` and `java.lang` API are all GB18030 and they refer to Simplified Chinese, RPC Standard.

The classes `java.io.InputStreamReader`, `java.io.OutputStreamWriter`, `java.lang.String` and classes in the `java.nio.charset` package can convert between Unicode and a number of other character encodings.

To find out what locales your Java Runtime Environment (JRE) supports, you have to ask each locale-sensitive class. Each class that supports multiple locales will implement the method `getAvailableLocales()`. The `getAvailableLocales()` method is implemented in many of the classes in the `java.text` and `java.util` packages. For example, `NumberFormat`, `DateFormat`, `Calendar` and `BreakIterator` provide it.

User interface elements in the runtime environment have been localized for `zh_CN` locale where Simplified Chinese is used. These elements include AWT and Swing components and other messages generated by the JRE and included tools.

The support for locale-sensitive behavior in the java.util and java.text packages is almost entirely platform independent, so all locales are supported in the same way and simultaneously, independent of the host operating system and its localization. The only platform dependent functionality is the setting of the initial default locale and the initial default time zone based on the host operating system's locale and time zone. Java.util and java.text packages support zh\_CN locale.

### Writing system

For the Java Foundation Classes (AWT, Swing, 2D, input method framework, drag and drop), locales can generally be characterized by just the writing system; there are no country or language specific distinctions.

Sun's J2SE Development Kit 5.0 and the international version of the J2SE Runtime Environment 5.0 support the Chinese writing systems:

Writing System	Chinese (Simplified)
Language	Chinese
Windows Encodings	936, GB18030
Solaris Encodings	GB2312-1980, UTF-8, GBK, GB18030-2000
Linux Encodings	GBK (a,b), GB18030-2000 (a,b,c,d,e), UTF-8 (a,b)
Peered AWT Components	supported

where

- a) on Sun Java Desktop System 2003
- b) on Sun Java Desktop System Release 2.
- c) on Red Hat Linux 9.0
- d) on Red Hat Enterprise Linux AS 3.0.
- e) on Turbolinux 10 Desktop.

### Text Input

Support for text input consists of two parts: interpretation of keyboard layouts, and text composition using input methods. For interpretation of keyboard layouts, the Java 2 platform relies entirely on the host operating system. For text composition using input methods, Java 2 platform supports native input methods using the host operating system's input method manager as well as input methods developed in the Java programming language.

Locale support in input methods implemented in the Java programming language depends solely on the set of installed input methods, not on the host operating system and its localization. However, support for the use of input methods implemented in the Java programming language with peered components is implementation dependent.

The following web pages of Sun Microsystems are very informative if you are planning software internationalization:

Java Internationalization: <http://java.sun.com/j2se/corejava/intl/index.jsp>

Sun software Product Internationalization Taxonomy:

[http://developers.sun.com/dev/gadc/des\\_dev/i18ntaxonomy/i18n\\_taxonomy.pdf](http://developers.sun.com/dev/gadc/des_dev/i18ntaxonomy/i18n_taxonomy.pdf)

Internationalization tutorial: <http://java.sun.com/docs/books/tutorial/i18n/index.html>

### 4.3.2 .net

In the .net framework the System.Globalization namespace is an important class in developing global applications. The CultureInfo class holds culture-specific information, such as the associated language, sublanguage, country/region, calendar, and cultural conventions. This class also provides access to culture-specific instances of DateTimeFormatInfo, NumberFormatInfo, CompareInfo, and TextInfo. These objects contain the information required for culture-specific operations, such as casing, formatting dates and numbers, and comparing strings. StringInfo and TextInfo classes provide advanced globalization functionalities, such as surrogate support and text element processing. The culture identifier for the current CultureInfo can be retrieved using CultureInfo.LCID property.

The culture names follow the RFC 1766 standard in the format "<languagecode2>-<country/regioncode2>", where <languagecode2> is a lowercase two-letter code derived from ISO 639-1 and <country/regioncode2> is an uppercase two-letter code derived from ISO 3166. The culture name for the PRC and Simplified Chinese is "zh-CHS".

The System.Text namespace contains classes representing ASCII, Unicode, UTF-7, and UTF-8 character encodings; abstract base classes for converting blocks of characters to and from blocks of bytes. Microsoft code page identifier for GB18030 is 54936.

The .net framework has System.Resources namespace to support the creation and localization of resources. It has also supports for packaging and deploying these localized resources. The ResourceManager class provides access to culture-specific resources at run time and controls how the application retrieves resources using the resource fallback process. The ResourceManager determines which resources to retrieve based upon the current thread's CultureInfo.CurrentUICulture property.

The .net framework uses Unicode UTF-16 (Unicode Transformation Format, 16-bit encoding form) to represent characters, in some cases UTF-8 internally.

### 4.3.3 Qt (Trolltech AS)

Qt is a product of Trolltech AS that has headquarters in Oslo, Norway. It also has offices in Brisbane Australia, Palo Alto California and Beijing China.

The Qt C++ framework supports the development of crossplatform GUI applications with its "write once, compile anywhere" approach. From a single source tree, recompilation is all that is required to produce applications for Windows 98 to XP, Mac OS X, Linux, Solaris, HP-UX, and many other versions of Unix with X11. Qt applications can also be compiled to run on embedded Linux platforms.

Qt uses Unicode throughout and has considerable support for internationalization. Qt includes Qt Linguist and other tools to support translators. Applications can easily use and mix text in Arabic, Chinese, English, Hebrew, Japanese, Russian, and other languages supported by Unicode. Qt has also conversion tools for GB18030 standard.

Qt makes it possible to create platform-independent database applications using standard databases. Qt includes native drivers for Oracle, MS SQL Server, Sybase Adaptive Server, IBM DB2, PostgreSQL, MySQL, Borland Interbase, SQLite, and ODBC-compliant databases.

There is also an Open Source edition of Qt.

## 5 Summary

The most important lessons that can be learnt in this document:

1. Only a fully internationalized software can be localized from any Western language to Chinese. The writing system and other cultural aspects are so different from the West that otherwise localization is not possible.
2. A software product locale that is built for China cannot be sold in HongKong or Taiwan. The different character set and transliteration systems require separate locales for them.
3. The GB18030-2000 standard must be followed. This fact has influence on all software layers of the product and it sets requirements to fonts that must be available for displays and printers of the user environment.
4. There are several other laws and regulations to follow e.g. when designing a database 'The security protecting regulations of the computer information systems of the People's Republic of China', 'The confidential law of the People's Republic of China' and 'The standard of the computer security law in China'.

The situation among products supporting the GB18030-2000 standard is changing all the time. Several products mentioned in this document have added the GB18030-2000 standard support during the past twelve months and new products supporting GB18030-2000 will be coming.

When it is time to make a decision about software products and development tools, products should be studied again one by one and new available tools and new features of old ones should be revised.

## References

1. Adobe CJKV Character Collections and CMaps for CID-Keyed Fonts Technical Note #5094, 11 June 2004: [http://partners.adobe.com/public/developer/en/font/5094.CJK\\_CID.pdf](http://partners.adobe.com/public/developer/en/font/5094.CJK_CID.pdf)
2. Adobe-GB1-4 Character Collection for CID-Keyed Fonts Technical Note #5079, 30 November 2000: <http://partners.adobe.com/public/developer/en/font/5079.Adobe-GB1-4.pdf>
3. Simon Ager: A guide to written language: [http://www.omniglot.com/writing/chinese\\_spoken.htm](http://www.omniglot.com/writing/chinese_spoken.htm)
4. AsianA Communications: <http://www.asiana.com>
5. Asianux: <http://www.asianux.com/>
6. Bitstream Incorporation: <http://www.bitstream.com/>
7. Character set standard GB18030-2000: <http://font.founder.com.cn/english/web/standard/GB18030-2000.doc>
8. Chinese languages (Chineselanguage.org) <http://www.chinalanguage.com/Language/chinese.html>
9. A Chinese-English dictionary, July 2003, 15 printing, Foreign language teaching and research press. ISBN: 7-5600-1325-2
10. Tom Emerson: The hitchhiker's guide to Chinese Encodings (20<sup>th</sup> International Unicode Conference) <http://www.basistech.com/knowledge-center/chinese/hhgtce-icu20-te2.pdf>
11. Thomas Emerson: Simple and Complex Chinese Scripts: History and Integration to Unicode (The 24<sup>th</sup> International Unicode Conference, Atlanta GA September 2003). <http://www.basistech.com/knowledge-center/chinese/iuc24-emerson-chinese.pdf>
12. Founder Electronics: <http://www.founder.com.cn/EnglishSite/>
13. GB18030 character set standard: <http://www.18030.com/isv/gb18030.asp>
14. P.A.V. Hall, R. Hudson (ed.): Software Without Frontiers: A multi-platform, multi-cultural, multi-nation approach; John Wiley & Sons 1997; ISBN 0-471-969-745 (partially in Finnish: [http://www.vtt.fi/tte/language/publications/A\\_SOFTAL.PDF](http://www.vtt.fi/tte/language/publications/A_SOFTAL.PDF))
15. HP-UX 11.0 - 11i Internationalization Features White Paper: <http://docs.hp.com/en/5991-1194/index.html>
16. Hunan Huatian Information Industry: <http://www.htit.com/english/english.htm>
17. Tauno-Olavi Huotari ja Pertti Seppälä: Kiinan kulttuuri; Kustannusosakeyhtiö Otava 1990; ISBN 951-1-10480-2
18. IBM eServer pSeries and AIX Information Center: <http://www16.boulder.ibm.com/pseries/>
19. Intersystems Corporation: <http://www.intersystems.com/>
20. The Localization Industry Primer, 2<sup>nd</sup> edition, 2003. From web page of The Localization Industry Standards Association LISA [www.lisa.org](http://www.lisa.org)

21. Longman Chinese-English Visual Dictionary of Chinese Culture, Longman 2004, ISBN: 7-81046-662-3
22. Ken Lunde: CJKV Information Processing Chinese, Japanese, Korean & Vietnamese Computing, O'Reilly&Associates, Inc., 1999, ISBN: 1-56592-224-7
23. Benson I Margulies: Your Passport to Proper Internationalization  
<http://www.basistech.com/knowledge-center/i18n/Your-Passport-to-Proper-Internationalization.pdf>
24. Dirk Meyer: Summary, Explanations and Remarks: GB18030-2000:  
[http://examples.oreilly.com/cjkvinfo/pdf/GB18030\\_Summary.pdf](http://examples.oreilly.com/cjkvinfo/pdf/GB18030_Summary.pdf)
25. Microsoft Corporation: <http://www.microsoft.com>, [www.microsoft.com/globaldev](http://www.microsoft.com/globaldev)
26. Monotype Imaging: GB2312 and GB18030  
<http://www.monotypeimaging.com/isv/gb18030.asp>
27. Oracle Corporation: <http://www.oracle.com/index.html>
28. Yip Po-Ching and Don Rimmington: Chinese, A Comprehensive Grammar, Routledge 2004, ISBN 0415-15032-9
29. NJStar Software corporation: <http://www.njstar.com>
30. 中华人民共和国国家标准标点符号用法 The PRC National Standard on the Usage of Punctuation <http://www.cmi.hku.hk/Ref/Article/article08/>
31. Manfred Rätzmann, Clinton De Young: Galileo Computing Software Testing and Internationalization (Lemoine International and the Localization Industry Standards Association 2003)
32. Sun Java Internationalization: <http://java.sun.com/j2se/corejava/intl/index.jsp>
33. Sun Java internationalization tutorial:  
<http://java.sun.com/docs/books/tutorial/i18n/index.html>
34. Sun Microsystems documentation: <http://docs.sun.com/app/docs/>,  
<http://developers.sun.com/>,
35. Sun software Product Internationalization Taxonomy:  
[http://developers.sun.com/dev/gadc/dev\\_dev/i18ntaxonomy/i18n\\_taxonomy.pdf](http://developers.sun.com/dev/gadc/dev_dev/i18ntaxonomy/i18n_taxonomy.pdf)
36. TrollTech AS: <http://www.trolltech.com/>
37. The Unicode Consortium [www.unicode.org](http://www.unicode.org)
38. URW++ Design & Development, GmbH, Amtsgericht Hamburg  
[http://www.urwpp.de/deutsch/fonttechnologie/chinurwfonts\\_cont.html](http://www.urwpp.de/deutsch/fonttechnologie/chinurwfonts_cont.html)
39. Andrea S. Vine: "Internationalization in software architecture and design". Article in Multilingual computing&Technology, Volume 13 Issue 3.
40. ZhongYi Electronics Ltd <http://www.china-e.com.cn/new/en/profile/ZhongYiProfile.htm>