IPOS copy

SYSTEM AND METHOD FOR PROCESSING CHINESE LANGUAGE TEXT

FIELD OF THE INVENTION

This invention relates to an improved system and method of using a computer to process Chinese language text, specifically phonetic Chinese text and Hanzi.

BACKGROUND OF THE INVENTION

Entering (inputting) and processing Chinese language text on a computer is a very difficult problem. The shear numbers of Chinese characters illustrate this difficulty. In the square-character (Hanzi) writing system of Chinese, there are 3000 to 6000 commonly used Chinese characters (Hanzi). Including the relatively rare ones, there are more than ten thousand Hanzi. Adding to this difficulty there are problems in the Chinese language with text standardization, multiple homonyms, and ill defined word boundaries that impede effective text processing of Hanzi with computers. In spite of intensive studies for several decades and the existence of hundreds of different methods, computer input and processing of Hanzi is a major stumbling block preventing the use computers in China, particularly for text processing.

The computer systems available today for inputting and processing Chinese language text may be divided into three categories:

The first category is based on a decomposition of the square characters into elementary graphical components. Different keys on the keyboard are assigned to represent different elementary graphical components of a Hanzi. Then, each character can be keyed in with a few key strokes -- as a combination of these elementary graphical components. Examples of this approach include Changji in Taiwan and the Five-Stroke method in mainland China. The major drawback of such methods is that the assignment of the keys to Hanzi components is artificial. In both the Changji and Five-Stroke methods, the assignment of the codes has to be mechanically memorized. Memorizing the keys representing the components is difficult and time consuming. Besides, the decomposition of a Hanzi into its elementary components in many cases is not unique. Although these methods are used by professional input operators, and high speed is demonstrated by well-trained typists, they are not used much by computer experts and other professionals, let alone ordinary people. Therefore, these methods tend to restrict the use of computers by the general Chinese speaking population.

The second and third category encounter a "homonym problem" in Chinese language processing.

The second category is phonetic input, (e.g. Pinyin for mainland China and "phonetic symbols" or BPMF for Taiwan) which is the most commonly used method for everyone except professional typists. The Hanzi writing system of Chinese language is a conceptual and practical barrier to this method.

Since there are only about 1300 different phonetic syllables, in contrast to tens of thousands of characters, one phonetic syllable may correspond to many different Hanzi. For example, the pronunciation of "yi" in Mandarin can correspond to over 100 Hanzi. This creates ambiguities when translating the phonetic syllables into Hanzi.

To address this "homonym problem," most of the phonetic input systems use a multiple-choice method. See for example, German patent 3,142,138, issued 5/5/83, by J. Heinzl et al.; US Patent 5,047,932, issued 9/10/91, by K. C. Hsieh; and Chinese patent 1,064,957 issued 3/8/91, by Tan Shanguang. After a phonetic syllable is keyed in, the computer displays all possible Hanzi with the same pronunciation. In some cases, there is not enough space on the screen to display all possible characters with the same pronunciation. These cases require scrolling up and down. Therefore, these phonetic methods, based on individual syllables, are very slow.

An improvement to the multiple-choice methods based on deriving probability of the adjacent Hanzi is disclosed in the prior art. See for example, British patent 2,248,328, issued on 4/1/92 to R. W. Sproat. The probability approach can further be combined with grammatical constraints. See for example, K. T. Lua et al., Computer Processing of Chinese and Oriental Languages, Vol. 6, Num 1. page 85, June 1992. However, the conversion accuracy (phonetic to Hanzi) of these methods is typically limited to around 80%.

The third category combines a phonetic-character input method with the addition of non-phonetic letters.

Non-phonetic letters are added to the phonetic letters to artificially discriminate characters with the same pronunciation. Examples include phonetic spelling with radical marks (British patent 2,158,776, issued 11/20/85, by C. C. Chen) and phonetic spelling with number of strokes (Chinese patent 1,066,518, issued 11/25/92, by G. Xie). These methods require memorizing artificial rules or counting number of strokes which slows down the speed of input substantially.

In addition to the "homonym problem," a "word boundary problem" exists when processing the Chinese language.

Although more than 80% of words in modern Chinese have multiple syllables (thus two or more Hanzi), there is no word separation in its writing system (in contrast with all European languages, and even Korean). Further, input of phonetic Chinese is usually performed syllable by syllable without accounting for word boundaries.

In spite of the wide recognition of multisyllable words and the lack of a standard way to delimit words at a word boundary, the definition and even the existence of words in Chinese is controversial. Furthermore, because Chinese is traditionally written a continuous string of Hanzi without word spacing, an ordinary Chinese person does not have a clear concept of what a "word" means. In many cases, it is unclear where a word boundary or delimiter, e.g. a space, should be placed. The controversy is exemplified by the following cases:

Compound nouns. In English, two independently valid words

can be combined to form a compound noun, for example, blackboard or rattlesnake. As in English, controversy exists about whether these compound strings should be treated as one word or two words. Because there is no generally accepted precedence in China, controversy about compound nouns is much more severe. For example, the word "nanguangboyuan" (male announcer), as listed in Chinese Pinyin Vocabulary, may be considered as two words (nan guangboyuan), or even three words (nan guangbo yuan) by different people. Affixes. All Chinese verbs can be appended with "syntax units" -le, -guo, or -zhe, which make them past, present prefect, or progressive tense. All adjectives can be appended with -de. However, these syntax units also appear as individual words called particles. Different schools of linguists treat these syntax units differently. Some schools treat these syntax units as "proper" affixes, i.e. part of the word to which they are attached. Other schools treat them as individual particles, i.e., separate words.

An affix is part of a word while a particle is an indiyidual word. For example, while the noun endings -hua, -jia, -yuan, -xing, and -zhuyi are considered by most linguists as affixes in single words, some linguists consider them individual particles (separate words). On the other hand, endings such as -z, -r, and -tou are always treated as suffixes for nouns and not as individual particles. Compound verbs. There is a class of verbs in Chinese which is very similar to the divisible verbs in German (die zerbrechbar Zeitwort), such as aufziehen, heraufziehen, etc. Those "divisible" verbs can use infixes, -zu- and -ge-, to become infinitives or past participles.

In Chinese, similar compound verbs can have infixes, -de- or -bu-, to mean "capable" or negative. Examples are, taiqilai (raise), which has versions of taideqilai (can raise) taibuqilai (cannot raise), very similar to the above German verbs. Moreover, the phrases "taiqi tou lai", taideqi tou lai", and "taibuqi tou lai" are similar sentence structures using compound verbs (such as "ziehen dein Kapf auf".) From this point of view, "taiqilai" should be one word. However, many linguists consider those syllables as separate words (tai, qi, lai), and write them separately.

As illustrated above, in the Chinese language it is often unclear where word boundaries should be placed.

In spite of the controversy, many multiple-syllable words are universally recognized as minimal linguistic units, or morphemes, such as: (1) nouns "gada", "putao", "feiji", etc.; (2) verbs "zhuanyou", "xingwu", etc.; and (3) adjectives "heised", "pangdad", etc. Also, many phrases are universally accepted that consist of multiple words. For example, although sometimes "dianzigongye" can be considered as one word, no one would consider the phrase "fazhan dianzi dongye" as a single word. There are popular four-syllable idioms that are universally considered as words, although in different writing styles of Pinyin, hyphens may or may not be used. For these classes of words, unique word boundaries are universally recognized.

As described above, the lack of universally accepted orthographic rules and the lack a word-separation habit for Chinese, make it very difficult to develop a easily used standard for computer input and processing of Chinese

language text -- no particular linguistic school is universally followed. Even by following a narrow definition of words (i.e., treating many compound words as phrases, and treating many affixes as particles), some ambiguities will remain. By making a broad definition of words, (i.e., treating many compound words as single units, and accepting many affixes as part of words), the accuracy of identification will improve, but the volume of vocabulary required to be stored in the computer memory would be too large to account for every single unit word and words with all affix combinations.

Textbooks of spoken Chinese for foreigners are written in a spelling form called Pinyin, where multiple syllable words are considered as basic units. Pinyin uses Roman characters and has its vocabulary listed in the form of multiple syllable words. A Chinese Pinyin Vocabulary was published in 1964. A revised edition was published in 1989 by Language Press, Beijing, China, which contains some 60000 word entries. Rules of orthography for Chinese written in Pinyin form, that define the word boundaries, were published in 1984.

OBJECTS OF THE INVENTION

It is an object of the present invention to ameliorate one or more disadvantages of the prior art.

SUMMARY OF THE INVENTION

The invention is a system and method for accurately and efficiently entering phonetic Chinese (Pinyin and BPMF) into a computer system and for accurately converting the phonetic input into the Hanzi form. The system has a novel keyboard with diacritic keys (and corresponding ASCII coding) that

permit the user to annotate each entered phonetic text syllable with a diacritic that indicates the tone of the syllable. A process executing on the system determines that a syllable has been entered when a diacritic (or delimiter) key is struck. An entered phonetic syllable is then compared to a list of acceptable phonetic syllables and abbreviations. If the entered syllable is on the list, the correctly spelled and accented syllable is stored in memory and displayed on a phonetic portion of a graphical display. The process continues for succeeding syllables until a delimiter is entered.

Upon encountering a delimiter, the word string (defined as the string of characters between two delimiters) is processed to determine the proper Hanzi characters that represent the word(s) in the word string. The word string (or each word in the word string) is compared to a dictionary of phonetic words. If the word string has a unique Hanzi translation, the Hanzi translation is stored in memory and displayed on a Hanzi portion of the graphical interface. If the word string has no matches in the dictionary, a morphological analysis is done, i.e., any standard affixes (as identified on an affix list) are removed from the word string. The remaining word string (root) is re-analyzed. If the root still has no matches in the dictionary, the root (remaining word string) is decomposed into constituent components/words. If the constituent components have unique matches in the dictionary, the Hanzi translation of these unique constituent components/words is stored in memory and displayed on the Hanzi portion of the graphical interface.

If the word input or any constituent components/words have ambiguous translations (non unique matches in the dictionary), a syntactical analysis is done. Words in the word string are classified as either function words (those on a function word/affix list and having no toned syllables) or notional words (those having at least one toned syllable.) (Note that for the purposes of this invention, function words, particles, and affixes are used and treated the same way.) The function words are translated unambiguously and also determine the structure of the remaining word string (root) input. Ambiguity of any notional words is then reduced or removed based on use of the notional words with the translated function words. Any remaining ambiguity of notional words is removed by a statistical model.

The phonetic text can be entered into the system by apparatus other than the keyboard, e.g. through a network, magnetic media, or acoustic system. Mix text, both Chinese and non Chinese words, can be processed by delimiting the non Chinese words with a special character, e.g. a space.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 is a block diagram a preferred embodiment of the present system.

Figure 2A shows one preferred embodiment of a keyboard used in the present system.

Figure 2B shows an alternative embodiment of part of a keyboard with tone keys.

Figures 2C and 2D show alternative preferred embodiments of keyboards used with the present system where each keyboard has split sections of the space bar used as tone keys.

Figure 3 is a preferred embodiment of an ASCII code data structure which includes the tone diacritics in the lower 7 bits and all usable vowels with tone diacritics in the higher 7 bits.

Figure 4 is a flow chart showing a preferred process for keyboard input that produces a Pinyin word registration in the memory.

Figure 5 is a flow chart showing a morphological analysis and decomposition process for Chinese words written in Pinyin form.

Figure 6 is a flow chart showing a preferred syntactical method for processing Pinyin input with ambiguous word strings.

Figure 7 is a table with a preferred list of Chinese syllables having one code form for each syllable.

Figure 8 is a table with a preferred list of Chinese function words, particles, and affixes treated as function words.

DETAILED DESCRIPTION OF THE INVENTION

Figure 1 shows a block diagram of one preferred embodiment of the present system 1000. The system 1000 includes a

graphical interface 1020 capable of displaying Chinese characters and English text, a CPU 1010, a main memory 1100, and memory storage 1200. A keyboard profile 1070 is stored in the main memory 1100 and/or memory storage 1200. The system 1000 has a pointing device 1031, such as a mouse, and can be connected to a network 1090. An IBM RS/6000 is one embodiment of the CPU 1010, memory (1100, 1200), and graphical interface 1020.

Also shown is the novel keyboard 1030. The keyboard includes tone diacritics keys 1050 and the syllable input keys 1040 that are standard Roman characters. Alternative tone keys 1060A,B are also possible. Tones keys 1060A can be placed on a split space bar. Tone keys 1060B can be added above the standard arrow keys. Alternative preferred embodiments of the novel keyboard are discussed in the description of Figures 2A-2D.

Included in memory storage 1200 and/or in main memory 1100 is a novel data structure 300 used for converting key strokes of the keyboard 1030 to encoded (e.g. ASCII) characters and diacritics that are stored, i.e. registered, in the memory (1100, 1200). Data structures 700, 800, and 950 are also stored in memory 1200. They include a Chinese syllable list 700, a function word/affix list 800, and a Chinese dictionary or vocabulary 950.

Novel processes 400, 500, and 600 are included in memory (1100, 1200.) Process 400 is a method for entering (inputting) phonetic Chinese into the system 1000 and optionally defining word boundaries. (Note that in this disclosure phonetic Chinese will be referred to as Pinyin

without loss of generality.) Process 500 is a method for morphologically analyzing strings of Pinyin text/words to determine unique Hanzi translations. Process 600 is a method of syntactical analysis used by process 500 for resolving ambiguous strings/words, i.e., those that have more than one Hanzi translation. The data structures 300, 700, and 800 are described in Figures 3, 7, and 8 respectively. Processes 400, 500, and 600 are described in Figures 4, 5, and 6, respectively.

Typically, the user uses the keyboard 1030 to enter Pinyin text input with appropriate diacritics (see below) into the system 1000. Using the data structure 300 and process 400, the system 1000 displays the Pinyin characters 1022 (Roman letters, diacritics, punctuation, etc.) on a first section 1021 of a split screen on the graphical interface 1020. Using processes 500 and 600, the system 1000 analyses the Pinyin input, converts the Pinyin into Hanzi, and displays the Hanzi characters 1025 on a second section 1024 of the graphical interface 1020. Note that using processes 500 and 600 and additional language translation apparatus 1034, the system 1000 can convert the entered Pinyin, coded or marked with diacritics, 1022 into other languages, like English.

The system 1000 can also include components other than the keyboard 1030 that permit the entry of text, e.g., a string can be transmitted over the network 1090. Other inputs include components 1033 for audio text input. These components would have the appropriate drivers to connect to system 1000. See U.S. Patent Application Number 08/404,786 entitled STATISTICAL ACOUSTIC PROCESSING METHOD AND APPARATUS FOR SPEECH RECOGNITION USING A TONED PHONEME

SYSTEM to Chen et al. filed on March 15, 1995 which is herein incorporated by reference in its entirety. Strings of Pinyin coded or marked with the appropriate diacritics can also be entered into the system 1000 using other well known methods including magnetic media 1032, e.g. diskettes.

When entering text, diacritics are used to demarcate stressed syllables (and in some embodiments unstressed syllables) and tone values in a phonetic representation of the Chinese language. (See process 400.) A diacritic is a symbol that represents a tone of a syllable. In the present invention, syllables of words that have a tone are annotated with a diacritic that indicates the syllable has a tone. Also, the diacritic indicates the type of tone. There are four tone types in Mandarin (standard Chinese). They are: high, rising, low, and falling. Some linguists include a fifth tone (diacritic) called the "untone or no tone."

During the entry (input) of text 400 into the system 1000, tone diacritics are added on accented syllables. With this coding or demarkation, the difference between notional words and function words, particles, and affixes immediately becomes clear in further processing of the text (500, 600). (Words with accented syllables are notional words and function words, particles, and affixes are all listed 800.)

In an alternative embodiment, a "no tone" diacritic is added after each unaccented (untoned) Chinese syllable (including the function words, particles, and affixes). This embodiment is useful when "mixed language/text" is entered, i.e. English or other non Chinese text is included with the Chinese text input. Since the non Chinese text is

uncoded (unmarked) with a diacritic, the Chinese words/syllables (marked with diacritics including the untone diacritic) can be distinguished from the unmarked non Chinese text. Non Chinese text is also delimited with special delimiters, e.g. with spaces between the words.

During further language processing, i.e., translation, analysis, and/or printing, diacritics demarking the accented syllables in a syllable string and indicating the type of the tone of each accented syllable, are used by the invention to parse the phonetic representation (Pinyin) of Chinese into words and sentences. (See processes 500 and 600.) Rules are applied to parse the phonetic representation by dividing the phonetic representation of the Chinese sentence at the points where there are function words, particles, and/or affixes. The system 1000 efficiently does this because there are a relatively small number, i.e., under 100, of known and identified function words (and particles/affixes) in each Chinese dialect. Therefore, a relatively small amount of computer memory is needed to store all useful function words and particles/affixes 800.

The system 1000 accurately determines word meanings and sentence structure. (See processes 500 and 600). First the function words (and particle/affixes), that have a unique and unambiguous translation in terms of meaning, use, and tone (none), are translated. The function words (and particles/affixes) are also used to define the sentence structure. Using the translation of the function words (and particles/affixes) and the defined sentence structure, meanings of the other (notional) words in the sentence can

be determined.

Again, for the purposes of this invention, function words, particles, and affixes are used and treated in the same manner. This is because often in Chinese writing no distinction is made because there is no clear word boundary. Function words, particles, and affixes also have a similar tone structure. A preferred list of function words, particles, and affixes is stored in the system memory (1100, 1200) and is presented below in Figure 8.

One novel feature of the present invention is a new keyboard designed specifically for Pinyin Chinese input and/or for mixed text processing (Pinyin Chinese together with non Chinese languages, e.g. English.) The keyboard provides a way of inputting Pinyin syllables with diacritics. Figures 2A-2D are examples of alternative preferred embodiments of keyboards 1030 specifically designed for Pinyin input. These keyboards 1030 enable easy and rapid entry of Pinyin into the system 1000. The standard United States English keyboard is not convenient for Pinyin input and also does not provide the ability to input tones/diacritics.

Figure 2A is a drawing of a novel keyboard (1030, 2000) that is a novel modification of a US English keyboard. In this particular preferred embodiment, some of the keys are used as tone (diacritic) keys 1050 while other keys 1040 are used in the traditional way as syllable keys 1040 for entry of Pinyin and other language (English) characters. In one preferred embodiment, some of the rarely used punctuation keys are assigned as the tone keys 1050. As a non limiting

example,{" key is assigned as the high tone (diacritic) key 2040; the "&rbt., }" key is assigned as the rising tone (diacritic) key 2050; the "▪, |" key is assigned as the low tone (diacritic) key 2060; and the " ",´ " key is assigned as the falling tone (diacritic) key 2070. For processing of purely Pinyin text, these four diacritic keys (2040-2070) are adequate because untoned Pinyin syllables and function words (and particle affixes) would be coded or marked with no diacritic. However, in embodiments that are used to process Pinyin text in combination with non Chinese (English) text (and other preferred embodiments), another diacritic key, the untone (diacritic) key 2010, is added. For example, one of the "alt keys" 2010 can be used for this function. Since in processing mixed text (both Chinese and non Chinese together) the untone key will be used frequently, a second untone key 2020 can be added at the second "alt key" location on the standard English keyboard to provide additional finger access. In this embodiment, the "alt key" would be reassigned to one of the duplicate "control keys" 2030. The English keyboard can be easily modified to include equivalent alternative assignments of keys to provide diacritic inputs 1050.

Figure 2B is an alternative preferred embodiment 2100 of a modified English keyboard 1030 that has diacritic keys 1060B for the entry of diacritics in Pinyin text processing. In this embodiment 2100, diacritic keys 1060 are added to the English keyboard 1050 above the standard "arrow keys." Again the diacritic keys that would be added include the high tone (diacritic) key 2120, rising tone (diacritic) key 2130, low tone (diacritic) key 2140, and the falling tone (diacritic) 2150 key. The optional untone (diacritic) key

2110 would be added for mixed text (or other) processing.

Figure 2C is a drawing of an another alternative preferred embodiment 2200 of the keyboard 1030. This embodiment 2200 has a split space bar 2270 that allows all keys (or all but one or two) of the US English keyboard layouts 1040 to be kept intact. In one preferred embodiment, the space bar 2270 is split into four sections (2230, 2240, 2250, 2260). The leftmost section 2230 of the space bar is assigned to the "rising" tone (diacritic) 2230. The large section 2260 next to it is used as the space bar 2260. The two smaller sections (2240,2250) on the right hand side are assigned to the "falling" 2250 and "low" 2240 tones (diacritics), respectively. The high tone (diacritic) 2220 is assigned to one of the "alt keys."

Such an arrangement is based on ergonomics. The thumbs are the most important fingers in piano playing, but much under utilized in typing. This design of the Chinese keyboard will utilize the thumbs much more efficiently in tone input. The new space bar 2260 is located at the center of the keyboard, namely, between "F" and "J" keys. Both thumbs can be used to press it comfortably. Thus, the size reduction of the space bar will not affect typing of English at all in mixed text processing. By curving a thumb to the palm, the four tone keys (2220, 2230, 2240, 2250) can be accessed easily.

For pure Pinyin input, no untoned key 2210 is needed. In this embodiment, the "space bar" key 2260 acts as the "untoned" key for the Pinyin only input. In Figure 2C, the position of the falling tone 2250 and the low tone 2240 are reversed from their common linguistic order (the third and

the fourth tones) because falling tone is used about a factor of two more frequently than the low tone. This "reversing" will reduce the fatigue and improve efficiency.

For mixed text processing (Chinese together with non Chinese), and some preferred embodiments of pure Chinese text processing, the untone key 2210 is added. In this preferred embodiment, the untone key (diacritic) 2210 is assigned to the other "alt key." One of the control keys 2280 is reassigned as the "alt key."

Figure 2D shows another preferred embodiment 2300 of a keyboard 1030. This embodiment uses a split space bar 2370 with rising 2330, falling 2350, and low 2340 tones (diacritics) positioned as in keyboard 2200. However, in this keyboard 2300, the center, largest portion 2310 of the space bar 2370 is assigned to the untone tone (diacritic) 2310. The space function 2360 is provided in the position of one of the "alt keys." This is especially useful when the keyboard 2300 is used to enter mixed text. The other "alt key" is assigned the high tone (diacritic) 2320. Again one of the "control keys" is reassigned as an "alt key" 2380. This embodiment positions the untone key 2310 in a very convenient location for access of both thumbs. This is helpful when the system 1000 is processing mixed text because the untone key 2310 will be frequently used.

Keyboard 2300 can be enhanced by adding additional marks on the keys. For example, in Chinese, the left quotation mark 2338 and right quotation mark 2339 are different and are each given a key. Left 2341 and right 2342 title marks are added along with a bullet 2344, apostrophe 2345, and decimal

point 2337. These marks are also used when entering mixed text. A phase mark 2336, used as a delimiter in Chinese is also provided a key location.

Key assignments and reassignments are performed using well known methods. For example, the profile of the keyboard 1030 is a file 1070 in the memory (1100, 1200) of the system 1000 that defines the functions of the keys. In the RS/6000 this is called the "keyboard translate table 1070." This table 1070 assigns a character, typically in ASCII code, or a control to each key on the keyboard 1030. By changing the assigned key codes in this table 1070, the functions of the keys on the keyboard 1030 change.

The user enters (inputs) Pinyin or mixed text into the system by using the keyboard 1030. After the user enters a Chinese (Pinyin) syllable, the user enters the appropriate diacritic associated with the syllable. When pure Pinyin is entered, no diacritic (or alternatively a space) is entered for unaccented syllables (and function words). When mixed text is entered (or in alternative preferred embodiments used to enter Pinyin), the user enters an untone diacritic after each unaccented syllable (and function word). Spaces (or equivalent word boundary delimiters) are entered after non Chinese words. Spaces delimiting Pinyin (Chinese) words with generally accepted word boundaries improves the accuracy of the system 1000. However, it is not necessary to delimit Chinese words for the invention to operate. See process 400.

The Pinyin (or mixed text) input is stored in the system 1000 memory (1100, 1200). In a preferred embodiment, the

input is encoded in an ASCII format. A standard ASCII code table that has codes for English characters and/or codes for GB2312-80 Hanzi is modified to include diacritics. In a preferred embodiment, the high, rising, low, and falling diacritics are coded for the upper and lower cases of each commonly accented vowel so that proper diacritic can be displayed/printed over the vowel in a displayed/printed Pinyin syllable.

Figure 3 is one preferred embodiment of a data structure 300 for the ASCII coding for Pinyin or mixed input. The block of memory 300 has 256 memory locations. These locations are numbered from 0 to 255. The number is also the ASCII code that is assigned to each of the locations/ASCII codes.

The ASCII code is made of a single bytes of information, a lower 7-bit section (representing locations 0-127) and an upper 7-bit section (representing locations 128-255). A single character, punctuation, or control function is assigned to a memory location. For example, the upper case Roman letter "B" is assigned location 66, the "escape" function is assigned location 27, and the "!" is assigned to location 33.

Although the coding of the Roman letters, punctuations, and numerals are well-defined, there are many optional spaces in the ASCII code. Individual countries can assign special symbols at these optional spaces.

Each of the diacritics is represented by a tone code or mark. In a preferred embodiment, each of the five (or four)

tone codes (marks) are assigned to memory location with numbers (ASCII code) in the lower half of all the ASCII code data structure 300 (the seven bits from 0 to 127.) These memory locations are used for communication. Therefore, by assigning the diacritics ASCII codes with numbers between 0 and 127, the tone codes can be communicated over a network 1090. In this non limiting example shown in Figure 3, the tone code (mark) no tone 301, high 302, rising 303, low 304, and falling 305 are assigned the locations with the ASCII code 2, 3, 4, 5, and 6, respectively. These are the locations of the Pokers in US ASCII coding.

In this preferred embodiment, a location and an ASCII code is also assigned to each vowel (both upper and lower case) for each diacritic (high, rising, low, and falling) that the vowel can have. The no tone case can use the English vowels that are located in the lower 7 bit section of the ASCII code. In addition, upper case "I" and upper case "U" with diacritics are omitted since no phonetic syllable begins with these vowels. Accordingly there are 32 vowel designations with diacritics to be assigned ASCII codes. These vowels with diacritics are assigned ASCII codes (128-159) in the upper half of the data structure 300, i.e. between numbers 128-255.

Giving vowels with diacritics a unique ASCII code is useful when displaying the Pinyin syllables on the graphical interface 1020 and in printing the Pinyin syllables. The system 1000 uses the Pinyin syllable entered with a following diacritic (301-305) to display/print a correctly spelled syllable with the diacritic properly over the vowel of the syllable. Therefore, by using the vowels assigned

with diacritics in data structure 300 locations 128-159, the system 1000 is able to print/display Pinyin syllables in the traditional way -- with the diacritic over the vowel. For example, the Pinyin syllable entered with key strokes Pin´ will be displayed/printed as Pi´n.

The advantage of this arranging the vowels with diacritics in the upper half of the data structure 300, with ASCII codes 128 to 159, is that the GB code only uses the upper half of ASCII code data structure 300 from 160 to 255. Therefore, the Pinyin vowels with diacritics can be stored and transmitted together with GB code (and the lower 7 bit of the ACSII code).

Figure 4 shows a novel process 400 of entering a Pinyin syllable into a tangible media like a computer memory (1100,1200) or a display (1020). In a preferred embodiment, process 400 provides a spell checker and an automatic error detection and correction process that is used to check the syllables entered into the system 1000 by key strokes (or other methods.) Process 400 also allows a mixture of Chinese and non Chinese (e.g., English) text to be entered, treated, and stored with minimal effort. An input unit is the input process 400 executed by the system 1000.

In step 401, the user enters a syllable (here either Chinese or non Chinese) through the keyboard 1030 with key strokes. If the syllable is Chinese, a diacritic representing the tone of the syllable is entered following the syllable. (Here the untone diacritic is used in text entry.) No diacritics are entered after non Chinese words, e.g. English. Further non Chinese words can be identified

402 because they are delimited by special delimiters, e.g. a space.

In the case of pure Chinese text, it is unnecessary in some embodiments to enter diacritic after no tone syllables or function words. Although not required in pure Chinese text entry, system 1000 accuracy improves by entering delimiters for word boundaries between Chinese words where a well established convention for word separation is known. This can be done by entering a delimiter like a space or other code (e.g. double diacritic, "esc") at the word boundary. In this case, a list of Chinese words with well established word separation can be referenced in 403 to distinguish them from non Chinese syllables/words.

When the end of a syllable (or non Chinese word) is reached, the system 1000 reads a diacritic, e.g. a depressed tone key 1060 (or space.) In some alternative embodiments where pure Chinese text is being processed, the no tone diacritic can be eliminated. In these cases, some untoned (and un marked) syllables will be appended to one another. To parse these syllables 404, the string is compared to a dictionary of syllables 700. The longest syllables in the dictionary 700 are compared to the entered string first. If part of the entered string matches the longest string, that matched string is used as a syllable input and deleted from the entered string. The process is repeated on the remainder of the entered string. If no match is found on the longest strings in the dictionary of syllables 700, the next longest syllable strings in the dictionary of syllables 700 is compared. The process is repeated until all the syllables of the entered string have been matched. This

process is call the "longest string match" algorithm and examples of this are well known. See "Chinese text segmentation for text retrieval: achievements and problems" by Z. Wu and G Tseng, Journal of the American Society for Information Science, Vol. 44, No. 9, October, 1993, pages 532-542 which is herein incorporated by reference in its entirety.

In some preferred embodiments of entering pure Chinese text and in mixed text processing, the no tone key is used to mark unaccented syllables in Chinese. In these embodiments, all Chinese syllables are coded (marked) and unambiguously delimited by a diacritic. Therefore, the Chinese syllable boundaries are unambiguously marked. This embodiment permits Chinese syllable abbreviations to be used since the abbreviations are also unambiguously delimited and are uniquely identified by being added to the syllable list 700. Use of syllable abbreviations greatly reduces the number of key strokes required to enter phonetic Chinese text.

In step 402, used in mixed text processing, the system 1000 determines whether or not the syllable is a Chinese syllable. If there is a diacritic (including the no tone diacritic) after the syllable, the syllable is Chinese 404. If there is no diacritic, the syllable (word) is non Chinese, e.g. English 403. Non Chinese words can also be identified 403 by their delimiters. Non Chinese words can be further processed, e.g. spell checked, by known systems that process these languages.

After a tone key (either one of the four tones keys or the "untoned" key) is pressed, the computer recognizes that the

preceding letters represent a Pinyin syllable 404 and the syllable is spell checked 405, checked as an abbreviation or alternate spelling 407, or error checked 408. These steps (405, 407, and 408) use a Chinese syllable list data structure 700 shown in Figure 7.

The Chinese syllable list 700 in Figure 7 is one preferred syllable list used for Mandarin Chinese. Because there are altogether just over 1300 syllables in Mandarin (including different tones), it is convenient to code those syllables with a small portion of a two-byte code space (65536 spaces). Each of the syllables is assigned a unique two-byte code or "short integer". Specifically, the ASCII code string representing each syllable is assigned a two-byte code in the Chinese syllable list 700. For example, the Mandarin Pinyin syllable Pi´n is assigned the two-byte code "814." Each syllable corresponds to a standard spelling form, including the tone diacritics on top of the proper vowel. To add abbreviations to the list 700, the abbreviations are assigned the same two-byte code as the syllable that is abbreviated. For example, the abbreviation "d" would be assigned 212, the code on the syllable list 700 that represents "de´", the syllable that "d" abbreviates. Syllable list 700 takes a relatively small amount of memory (1100, 1200) space.

After a Chinese syllable 404 is encountered, the system 1000 searches the Chinese syllable list 700 to find a match. If a match is found, the syllable 404 was entered with a standard spelling 405, i.e. the spelling on the Chinese syllable list 700, and the entered key stroke string is replaced with the two-syllable code of the matched word on

the Chinese syllable list 700. For example, if the Chinese syllable Pin´ is entered, the system 1000 represents the key strokes with the ASCII code string {80, 105, 110, 6}. See Figure 3. When checking the Chinese syllable list 700 for a match 404, the system finds that this string corresponds to Pi´n that is represented by the two-byte code 814. See Figure 7. Thus the two bit code 814 replaces the ASCII code string for this syllable in memory. Therefore, if a standard spelling is detected 405, the entire syllable is stored 412 in memory 1010/1020 as one of the 1300 Chinese syllables (by a two-byte code) and displayed on the monitor 411, with the tone diacritics correctly on top of the vowel. This Pinyin display 411, is shown in the region 1021 of the graphical interface 1020 that displays Pinyin.

In like manner, if the user enters an abbreviation or shorthand spelling for a Chinese syllable, the ASCII codes representing the shorthand spelling are compared 406 to the entries in the Chinese syllable list 700. If there is a match, the shorthand (abbreviated) syllable is converted 407 to the proper, unabbreviated Pinyin syllable form by replacing the entered ASCII codes of the shorthand spelling with the two-byte code corresponding to the proper, unabbreviated Pinyin syllable. Again, abbreviations are added to the Chinese syllable list 700 with codes that will print the full syllable that is abbreviated.

Some example example abbreviations or shorthand spellings of Chinese syllables follow. For example, "zi", "si", "ci", "ji", "qi", and "xi" can be spelled as "z", "c", "s", "j", "q", and "x" without any confusion. Similarly, "yi", "wu", "yu" can be spelled as "i", "u", and "v" without confusion.

Other shorthands such as "shg" for "sheng", "xg" for "xing", "qag" for "qiang", "jo" for "jiao", etc, can be recognized without ambiguity.

In syllables with syllable separation marks, the syllable separation mark, such as that in "Xi'an", can be avoided altogether in typing, although it should appear during display. Using the present system 1000 and process 400, the above examples of shorthand typing are so convenient that no learning is required. With a little learning the number of shorthand syllables can be increased (added to the Chinese syllable list 700.)

If an erroneous spelling is detected 408, i.e., there is not match in table 700, the most probable syllable is displayed 409. This is done by presenting a menu of probable choices, i.e. best matches 1023, selected from the Chinese syllable list 700. The user selects 409 the proper syllable from the menu by using a selection apparatus, e.g., a mouse 1031 or a key. As before, the computer codes this selected syllable as a two-byte value and stores it in the memory 412.

If a delimiter, e.g., a space bar, carriage return, or punctuation key, is not encountered 413, the process 400 repeats 415 for the next syllable. If a delimiter is encountered the entire string is stored as a word string 414. On the monitor, the correct Pinyin spelling of this word string is displayed in the Pinyin portion 1021 of the graphical interface 1020. Note that the word string 414 between delimiters is treated as a single word when the delimiters are entered (according to a generally accepted

convention) and the word will have unique word boundaries coded or marked by the delimiters. However, the word string 414 can include two or more words (constituents) and/or affixes if the delimiters were not placed at each of the constituent word boundaries, e.g., when there are no generally accepted word boundaries to delimit the constituent words.

In step 414, the text, in the coded Pinyin form, also can be printed as Pinyin syllables using a printer 1080 with a special read only memory 1081, ROM, which converts the two-byte codes of table 700 into a string of Roman letters with tone symbols on the proper vowel of the required syllables. The text then becomes a continuous and readable printout in the Pinyin form, same as in all textbooks on spoken Chinese for foreigners.

Figures 5 and 6 are flow charts of methods for processing the Pinyin text that is entered into the system 1000, e.g. by process 400. In one preferred embodiment, Pinyin text 1022 is entered by key strokes on the keyboard 1030 of the system and displayed, as discussed above, on the Pinyin portion 1021 of the graphical interface 1020. Concurrently, the Pinyin input in word form, stored above in step 414 is converted by process 500 into Hanzi form and displayed 1025 in the Hanzi portion 1024 of the graphical interface 1020. Therefore, the user entering Pinyin text and diacritics through the keyboard 1030 sees the Pinyin text 1022 (with the diacritic placed over the proper vowel by process 400) on the Pinyin portion 1021 of the graphical interface 1020 and the correct Hanzi translation 1025 of the entered Pinyin displayed on the Hanzi portion 1024 of the graphical

interface by process 500. In alternative embodiments, Pinyin text with diacritics can be entered into the system 1000 by means other than the keyboard 1030, e.g., over a network 1090, audio system 1033, or with a magnetic medium 1032 like diskette.

Process 500 translates the Pinyin input into Hanzi. The system 1000 can process Pinyin and Hanzi in other ways. For example, the Pinyin 1022 text and/or the Hanzi text 1025 can be printed. Alternatively, the Pinyin 1022 and/or Hanzi 1025 text can be entered into known language translators 1034 and translated into other languages like English.

Processes 500 and 600 takes full advantage of the word boundary and the diacritics coding (marks) on the accented syllables. The two processes together handle a large variety of different styles of Pinyin writing theories and habits, and achieve high conversion accuracy. For example, if the Chinese words are delimited at well established boundaries, the word input 501 will be recognized as a unique word 511. If the word, is not and/or can not be delimited from other "words, " a morphology unit being the process 500 executing on system 1000, morphologically analyzes and/or decomposes the word string. The invention is therefore able to determine the constituent "words" in the word string. This is done using the vocabulary 950 and the "longest string matching" algorithm described above.

The invention is also able to accurately processes Pinyin into Hanzi for either of the following conventions affixes included as part of the word or 2. affixes as a particles separate from the word. In one preferred embodiment, the

second type of affix (particle) is treated as a function word. For convention 1, process 500 performs a morphological analysis, i.e., standard affixes included in unrecognized words (convention 1) are removed and the remaining word string (root) is analyzed without the affix. In later processing the removed affix is treated as a function word.

Process 500 uses morphology to process word string input. Morphology is the study of word formation in a language. Here the formation of words specifically involves affixes and how the meaning of words is affected by affixes. In the orthography of most European languages, such as English, German, Spanish, Finnish, and Hungarian, word meanings change depending on how affixes are used. For example, in the English expression "Calvin and Dorothy's Department Store", the "'s" belongs to "Calvin and Dorothy", but is appended only after "Dorothy".

Process 500 first analyzes the morphology of entered words or word strings, e.g., from step 414. The algorithm starts with comparing 510 the entered word (string) 501 to a known vocabulary 950. The vocabulary 950 is a table of phonetic Chinese words and/or abbreviations (Pinyin and/or BPMF) with their associated Hanzi translation. A dictionary 950 of Chinese words can also be used. Such dictionaries 950 are well known. For example, see "Hanyu Pinyin Cihui" by the language publishing house "Yuwen Chubanshe," 1989. Dictionaries such as this can be improved by combining other known dictionaries and/or adding Chinese words with their Hanzi translations from other sources.

If the word input 501 (word or word string) matches one or more words in the vocabulary 510, its uniqueness is checked, 511. If there is no ambiguity, i.e., there is only a single match in the vocabulary 510, the word is converted to Hanzi form and displayed 550 in the Hanzi portion 1024. If it is ambiguous (that is, there are more than one Hanzi forms in the vocabulary (510, 950) corresponding to the phonetics of the word input 501), the different words are stored for further consideration (step 540).

If the word input 501 is not in the vocabulary, a morphological analysis is performed on the word input (string) 501. Part of the word input is compared with the list of function morphines, or affixes, step 522. A preferred list of affixes 800 is provided in Figure 8. If a standard affix is found, i.e., it is on the affix list 800, the found affix is cut off 524 and the remaining part of the word input (root) is compared 510 with the dictionary 950 again. If the root of the root is in the vocabulary, step 511 is applied. If the root is not in the vocabulary 510 but there are more affixes in the root 522, step 524 is repeated for any other standard affixes remaining in the root.

If no standard affixes are recognized 522, i.e. no string in the word input 501 is in the affix list 800, the word string/root is decomposed 530. In one preferred embodiment, the root is decomposed using the "longest match method" above by comparing the root to the vocabulary 950.

If the word string is composed of more than one constituent, the uniqueness 531 of each constituent is searched. If a constituent is unique 531, the word is

considered as a compound word and the corresponding string of Hanzi is displayed and recorded 550. If the word is ambiguous 531, e.g. one or more of the constituents of the word input 501 have more than one Hanzi translation and/or there are more than one ways to decompose the word string, these different alternatives are stored for further consideration, step 540.

In step 540, the whole phrase (or clause, or sentence), i.e., word string input 501, is determined as that string of characters that is delimited by phrase delimiters such as punctuations, carriage return, or any other phrase delimiters assigned by the user. The whole phrase is analyzed to determine the meaning of the ambiguous word(s). See process 600 of Figure 6 below. After step 540, the entire sentence (phrase, clause) in Hanzi form is displayed, step 550. If there is any error, step 560, in the text displayed 560 on the graphical interface 1024, the user corrects the phrase, sentence or clause manually, 561.

To support the explanation of the syntactical analysis used by present invention given below (process 600), a discussion of a discovery of phonetic characteristics of acoustics of speech (tone and accents), function words, particles, affixes, and notional words is now presented.

There is an important phonetic phenomenon in spoken Chinese often overlooked. In Chinese, as in all spoken languages on the world, there is a great deal of information provided by the acoustics of speech in addition to the information provided by the mere string of individual syllables. In Chinese, each word has an accent structure that results in a

speech acoustic for the word that is often crucial in understanding the word meaning. How syllables are grouped to form a word can affect the accent structure, speech acoustic, and understanding of the word.

Again, the separation of words (syllable grouping) and the commensurate accent structure of words (speech acoustic) are part of the information conveyed in the spoken language. For the Chinese language, tone, a part of the speech acoustic, is important as well. In Chinese, tone of a syllable is prominent only in accented syllables. For syllables that have no accent, the tone is either substantially weakened or completely lost.

Further note that in most languages, including Chinese, there are function words and notional words. Notional words include all nouns, all verbs, all adjectives, and most adverbs in a language. For example, "man, run, fast, and quickly" are notional words. Function words include articles, prepositions, and some conjunctions. For example, "a, the, to, from, and, but etc.," are function words. Affixes denoted as separate particles have characteristics of function words.

Function words in Chinese are unstressed words that express primarily a grammatical relationship and/or functions of other words rather than expressing a lexical meaning. Function words include prepositions, auxiliary verbs, conjunctions, adverbs, interjections, articles, and certain pronouns. In Chinese, the function words (articles, prepositions, postpositions, etc.), affixes in the form of a particle, and any affix in a multiple syllable word are

always NOT accented. Function words (and affixes in particle form and affixes in a multiple syllable word) have no tone or a substantially weakened tone.

Notional words are words that have a full lexical meaning. In Chinese, notional words include all nouns, all main verbs, all adjectives, and most adverbs. Notional words include all words in Chinese that are not function words. The accent structure of notional words depends on the dialect.

Until now, it has not been recognized how a person listening to a language intuitively identifies notional words and function words in order to understand the words of the language. This is particularly true in the Chinese language. Since the tone of function words and affixes is very weak or completely missing in Chinese and because there are a limited number of known and identified function words in the Chinese language, function words are intuitively used by people listening to spoken Chinese to understand what is being said.

To further amplify the importance of function words, notional words, and tone in understanding the Chinese language, the accent structure of notional words and the use of function words (and affixes) is now examined.

The Shanghai dialect represents the clearest example of the accent structure of notional words and function word use. Every notional word starts with an accented syllable, which has a prominent tone. All other syllables in the word have no accents at all, therefore the tones of the syllables that

follow the first syllable are muted. For example, in Shanghainese, the nouns "wenti" (question) and "dianhua" (telephone) are always accented at the first syllable.

The function words in the Shanghai dialect have no accent. When a function word follows a notional word, the function word sometimes can be interpreted as a suffix (i.e., an affix) of that notional word rather than a particle. However, part of the discovery is that from the point of view of human listener, these two interpretations (a function word being a particle or as a suffix) have no difference in substance.

In the Beijing and Mandarin dialects the accent structure is slightly different.

In most cases of the Beijing dialect, the accent of a notional word is at the first syllable. However, there are many cases where the accent is located at the last syllable instead of the first one.

Mandarin is the standard Chinese language. In Mandarin, the accent of a word is usually at the first syllable. In some words with more than 3 syllables, a "framed" accent structure is observed. That is, the first and the last syllables are accented, whereas all other syllables in between have no accent and have very weak (or no) tone. A multiple syllable notional word in Mandarin has at least one accent, usually at the first syllable. For example, in the syllable string "taiqilai", only the first syllable "tai" (lift) has an accent. For the remaining two syllables, not only is the acoustic energy much reduced, but also the tones

are substantially weakened or completely lost. In the phrase "taibuqi tou lai", only the two syllables "tai" and "tou" have accents. The remaining syllables "ba", "qi", and "lai" are weak, and are atonic or almost atonic. (Atonic means without tone.) Note that "lai" is an afix that has no tone and, in this convention, is written as a separate particle.

In both the Beijing and Mandarin dialects, the function words, as a rule, have no accent and the tones of the function words are weak or missing (degenerate).

The discovery is that listeners to the spoken Chinese language intuitively use the accent structure of notional words (syllables with tone and syllables without tone) and untoned function words (and affixes) to demark word boundaries, classify words, and determine the meaning of words. This natural phonetic phenomenon is independent of the differing views of linguistic philosophy of the Chinese language and/or how one writes words or phrases in Chinese. The present invention develops and applies rules based on this phenomenon, and implements these rules with appropriate processes, to improve the accuracy of the processing (entering, analyzing, translating, and writing) phonetic Chinese.

Having completed the discussion of the discovery of phonetic relationships that were until now unrecognized in the prior art, the following discussion explains how these relationships are used in the syntactical analysis done by process 600.

A novel syntactical analysis method is shown in Figure 6 as

process 600. The process analyzes an entire phrase (or clause, or sentence) using principles discovered and disclosed above.

A word string (clause or a sentence) 501 is delimited with phrase delimiters like punctuations, a carriage return, or any input assigned by the user. In a preferred embodiment, these delimiters, or phrase delimiters, include the space (key 2360), phrase mark (key 2336), Chinese period 2335, left quotation mark (key 2338), right quotation mark (key 2339), left title mark (key 2341), right title mark (key 2342), exclamation mark (key 2311), comma (key 2312), slash (key 2313), left parenthesis (key 2314), right parenthesis (on key 2315), colon (on key 2316), semicolon (key 2316), bullet (key 2344) and question mark (key 2317). A double stroke on any of the tone keys (2310, 2320, 2330, 2340, and 2350) can also be a phase delimiter. (The keys indicated in parenthesis here designate the key in this embodiment of the keyboard 2300 that enters the phrase delimiter.) These phrase delimiters unambiguously define a word boundary, i.e. the end of one phrase and the beginning of another phrase 501.

Phrases with one or more ambiguous words, as defined in steps 511 and 531, are input 601 to the process 600. A syntactical unit is process 600 being executed on the system 1000.

First, the word components of a phrase 601 are classified by the existence of any tone diacritics. The words with at least one diacritic 610 are notional words 620.

Words without diacritics 611, however, could be: notional words with the diacritic inadvertently omitted, non-Chinese words, function words, affixes or affixes written as independent particles. In mixed text entry, non-Chinese words are delimited by spaces (or another equivalent character, e.g. the "Esc") and therefore are recognized.

To determine if the words without diacritics 611 are function words, the words 611 are compared to a function word list 800. The number of function words is limited, under 100 according to most linguists. If the word 611 is on the function word list 800, it is a function word. The meaning of each function word is uniquely defined. A preferred function word list 800 is provided in Figure 8.

In similar manner, the words without diacritics 611 can be compared to an affix list (also included in list 800) to determine whether they are affixes and/or affixes written as independent particles. The meaning of these independent particles is uniquely defined. One preferred list of particle/affixes is also relatively short, having under 20 entries. Particle affixes with their Hanzi translations are included on the list of function words in Figure 8.

Data structure 800 that lists function words, particles, and affixes take relatively little space in the computer memory (1100, 1200.)

Once all the function (and/or particle affixes) words are identified, and therefore uniquely defined, the phrase/sentence structure is analyzed in step 630. Since the meanings and the function of the function words

(particle/affixes) are known, the types of words preceding and following any particular function word (affix) are limited. Therefore, the ambiguity of any notional word can be further reduced or eliminated in step 640 by using the relationship of the ambiguous notional word with one or more function words (particle/affixes.) For instance, if there are five possible words identified in step 540 of process 500 for an ambiguous notional word, some or all but one of these possible words might be eliminated by knowing the function word (affix) preceding or following the ambiguous word.

For example, prepositions 810, considered function words, are listed in Figure 8. These prepositions must be followed by a noun or a noun phrase. If any notional word following these prepositions has a translation that is not a noun or a noun phrase, this ambiguous translation is removed.

Tense particles 820 (le, zhe, guo) usually follow verbs. Therefore, any ambiguous notional word with these particles following is translated as a verb. If any of the alternative translations for this notional word are not a verb, they are deleted.

Relation particles 830 (de, zhi) usually follow nouns or adjectives. Therefore, any ambiguous notional word with these particles following is translated as a noun or adjective. Alternative ambiguous translations are deleted.

Conjunctions (he, ji, and gen) 840 are function words and usually go between nouns. Therefore, any ambiguous notional word on either side of these conjunctions 840 is translated

as a noun. Alternative ambiguous translations are deleted.

Conjunctions 850 are function words that usually go between phrases. Notional words are either side of these conjunctions 850 are considered as part of a phrase.

Noun suffixes 860 are affixes that are usually are part of a noun. Notional words preceding these suffixes 860 are translated as nouns. Alternative ambiguous translations are deleted.

There are adjective and adverb suffixes 870 that must be preceded by an adjective or an adverb. Any notional words preceding these suffixes (particles) having a translation that is not an adjective or an adverb are removed.

There are also adjective and adverb prefixes that include hen-, zui-, bu-, and geng-. These particles/affixes must be followed by adjectives or adverbs. If any notional word following these prefixes (particles) has a translation that is not an adjective or an adverb, these ambiguous translations are removed.

There are rare exceptions to these rules which might result in an erroneous translation that would be corrected by hand.

Note that the list 800 does not include abbreviations of the function words, particles, and affixes. Abbreviations could be included in alternative embodiments.

If there is a remaining ambiguity in step 640, i.e. more

than one possible notional word still exists, statistics of the relative word occurrence in the phrase are applied 641 to predict the most probable word of the remaining notional word candidates. Statistical models 641 like this are well known in the speech recognition arts. See for example, .usFundamentals of Speech Recognition by Rabiner and Juang, 1993, pages 447-450 which is herein incorporated by reference.

In one preferred embodiment, the results are displayed 650 on the Hanzi 1024 portion of the graphical interface 1020. If there are words that can not be resolved unambiguously, the remaining candidates are presented to the user on the graphical interface 1020 for selection by a pointing device 1031. In addition, the displayed result 650, 1025 is subject to manual correction by the user 650. In alternative embodiments, the final display could be printed on a printer or transmitted over a network connection 1090. The displayed result 650 can also be Hanzi input to other computer systems, for example, a translator 1034 that translates the Hanzi text into another language.

By using processes 500 and 600, the system 1000 can determine the meanings and Hanzi translation of constituent words in the word input 501, even though the words were not separated according to any generally accepted convention or delimited at a boundary.

Given this disclosure one skilled in the art could develop alternative equivalent embodiments that are within the contemplation of the inventor. For example, a keyboard like the one in Figure 2D can be designed to include the 37

characters in the BPMF phonetic alphabet. The key assignment for this phonetic alphabet depends on the choice of the user/manufacturer. ASCII code assignments for the 37 characters can be made in the upper 7 bit region of Figure 3. Diacritics for BPMF can also be provided on the keyboard and in the ASCII code table. However, some modification may be necessary. For example, a diacritic for the high tone might be provided. The syllable table 700 and function word, particle, and affix table 800 for BPMF has a one to one correspondence to the Pinyin case. The processes shown is Figure 4, 5, and 7 still apply.

CLAIMS

1. A computer system for processing Chinese language text comprising:

a computer memory;

an input apparatus for entering a plurality of Chinese phonetic language syllables into the system, each syllable having one or more characters, the input apparatus marking one or more of the accented syllables with diacritic that indicates a tone of the accented syllable;

an input unit that provides a character code for each character and a tone code for each diacritic entered by the input apparatus, the input unit recognizing a syllable as a string of character codes marked by tone code;

a syllable list having a plurality of syllable strings each associated with one or more syllable string codes, the input unit matching the string of character codes marked by the diacritic code to one of the syllable string codes and storing in the computer memory a syllable representation for the syllable string associated with the syllable string code that matches the string of character codes marked by the tone code.

2. A system as in claim 1, where the Chinese language text is Pinyin.

3. A system as in claim 1, where the Chinese language text is BPMF.

4. A system as in claim 1, where the input apparatus is a keyboard with one or more keys for entering diacritics.

5. A system as in claim 1, where the input apparatus is an acoustic device for entering spoken Chinese.

6. A system as in claim 1, where the system further comprises a printer and the syllable representations stored in the computer memory are converted into printed syllables that are printed on the printer.

7. A system as in claim 1, where the system further comprises a graphical user interface and the syllable representations are converted to a syllable text that is displayed on the graphical user interface.

8. A system as in claim 1, where one or more of the syllable string codes represents an abbreviation of the associated syllable string.

9. A system as in claim 1, where the input unit provides a list of best matches on a graphical interface to a user if no syllable string code matches the string of character codes marked by the tone code.

10. A system as in claim 1, where the input unit recognizes the string of character codes as a non Chinese syllable where the string of character codes is unmarked by a tone code.

11. A computer system for processing Chinese language text comprising:

an input apparatus for entering a phonetic Chinese language phrase, the phrase having one or more words, each word having one or more syllables, each syllable having one or more characters, the phrase being a string of the characters between a first and second phrase delimiter;

an affix list having a plurality of entries being phonetic Chinese affixes;

a vocabulary of Chinese words, the vocabulary being a list of a plurality of phonetic Chinese words with a Hanzi translation; and

a morphology unit that removes one or more affixes from the phrase to create a root, the removed affix being one of the affixes on the affix list, the morphology unit comparing the root to one or more of the words in the vocabulary to find a match, and storing in a computer memory the Hanzi translation of the word in the vocabulary that matches the root.

12. A system as in claim 11, further comprising a syntactical unit, the syntactical unit parsing the root into accented words and unaccented words, the accented words being marked with diacritics and unaccented words being not marked with diacritics.

13. A system as in claim 12, where the entries include affixes, function words, and particles, each entry having a Hanzi translation, the syntactical unit matching the unaccented words with one or more of the entries, the entry matching the unaccented word being a matched entry, and the syntactical unit using the Hanzi translation of the matched entry as a translation of unaccented word in the phrase.

14. A system as in claim 13, where the syntactical unit further translates one or more of the accented words as an ambiguous notional word, the ambiguous notional word having an ambiguity because of there is more than one Hanzi translation of the ambiguous notional word in the vocabulary, and the syntactical unit removing the ambiguity by using a relationship of the ambiguous notional word with one or more of the unaccented words in the phrase.

15. A system as in claim 12, where the first and second phrase delimiters are any one of the following: a carriage return, a space, a phrase mark, a Chinese period, a left quotation mark, a right quotation mark, a left title mark, a right title mark, an exclamation mark, a comma, a slash, a left parenthesis, a right parenthesis, a colon, a semicolon, a bullet, a question mark, and two or more strokes of a tone key.

16. A system as in claim 12, where the morphology unit decomposes the word into constituent words if the word does not match one of the words in the vocabulary and the syntactical unit parses the constituent words.

17. A computer system for processing Chinese language text comprising:

a computer memory means;

an input means for entering a plurality of Chinese phonetic language syllables into the system, each syllable having one or more characters, the input means marking one or more of the accented syllables with a diacritic that indicates a tone type of the accented syllable;

an input process means that provides a character code for each character and a tone code for each diacritic entered by the input apparatus means, the input process means recognizing a syllable as a string or character codes marked by the tone code;

a syllable list means having a plurality of syllable strings each associated with one or more syllable string codes, the input process means matching the string of character codes marked by the tone code to one of the syllable string codes and storing in the computer memory means a syllable representation for the syllable string associated with the syllable string code that matches the string of character codes marked by the tone code.

18. A method of storing syllables and words of phonetic Chinese in a computer memory of a computer system comprising the steps of:

a. entering one or more characters into the computer system using an input apparatus, a syllable having one or more characters, a word having one or more syllables, and a phrase having one or more words, the words of the phrase being between two phrase delimiters, the syllables being marked with a diacritic representing a tone of the syllable and the syllables being between two diacritics, the entering being the storage of a character code for each character entered and the storage of a tone code for each diacritic marking one of the syllables;

b. recognizing one or more of the syllables as a Chinese syllable, a Chinese syllable being a string of one or more character codes and a tone code;

c. matching the Chinese the string to a syllable list, the syllable list having a plurality of syllable strings each associated with one or more syllable string codes; and

d. storing in the computer memory a matched syllable string, the matched syllable string being the syllable string that is associated with a syllable string code that matches the string.

19. A method, as in claim 18, where steps a-d are repeated until a phrase delimiter is entered, the one or more matched syllable strings stored in the computer memory being a word string.

20. A method of inputting a phonetic Chinese language string through an alphanumerical keyboard, comprising the steps of:

inputting a phonetic representation of the Chinese language into an input device of a computer system, the phonetic representation having one or more words each made of one or more phonetic syllables, one or more of the syllables being accented syllables and one or more of the syllables being unaccented syllables; and

marking the accented syllables with a diacritic that indicates a tone type of the accented syllable.

21. An program storage device readable by a computer, tangibly embodying a program of instructions executable by the computer to perform method steps for storing syllables and words of phonetic Chinese in a computer memory of the computer, the method comprising the steps of:

a. entering one or more characters into the computer system using an input apparatus, a syllable having one or more characters, a word having one or more syllables, and a phrase having one or more words, the words of the phrase being between two phrase delimiters, the syllables being marked with a diacritic representing a tone of the syllable and the syllables being between two diacritics, the entering being the storage of a character code for each character entered and the storage of a tone code for each diacritic

* C L M S *

marking one of the syllables;

b. recognizing one or more of the syllables as a Chinese syllable, a Chinese syllable being a string of one or more character codes and a tone code;

c. matching the Chinese the string to a syllable list, the syllable list have a plurality of syllable strings each associated with one or more syllable string codes; and

d. storing in the computer memory, a matched syllable string, the matched syllable string being the syllable string that is associated with a syllable string code that matches the string.