The Pitfalls and Complexities of Chinese to Chinese Conversion

汉字简繁转换的复杂性和陷阱
漢字簡繁轉換的複雜性和陷阱

Jack Halpern
Editor-in-Chief
CJK Dictionary Publishing Society

Jouni Kerman
Chief of Software Development
CJK Dictionary Publishing Society

Abstract

Chinese is written in two forms: **Simplified Chinese** (SC), used in the PRC and Singapore, and **Traditional Chinese** (TC), used in Taiwan, Hong Kong, and elsewhere. A common fallacy is that there is a straightforward correspondence between the two systems, and that conversion between them merely requires mapping from one character set to another. In fact, there are major differences between the systems on various levels: character sets, encoding methods, orthography, vocabulary, and even semantics.

With the growing importance of East Asia, localization and translation companies face an urgent need to convert between SC and TC, but must contend with such obstacles as the lack of knowledge to develop good conversion tools, no access to high quality dictionary data, and the high cost of manual conversion.

In 1996, the **CJK Dictionary Publishing Society** launched a project to investigate these issues in-depth, and to build a comprehensive database whose goal is to enable conversion software to achieve near 100% accuracy. The CDPS has collaborated with **Basis Technology** in developing the sophisticated segmentation technology required to achieve this.

This paper explains the complex issues involved, and shows how this technology can significantly reduce the time and costs of Chinese localization and translation projects.

1 Introduction

1.1 Simplified and Traditional Chinese

The term **Simplified Chinese** (SC) typically refers to a Chinese text that meets the following conditions:

1. **Character forms:** SC must be written with the simplified character forms (unless no simplified form exists).
2. **Character Sets:** SC normally uses the GB 2312-80 character set, or its expanded version called GBK.
3. **Encoding:** SC normally consists of GB 2312-80 text encoded in EUC-CN, or in HZ used for Internet data transmission.
4. **Vocabulary:** Choice of vocabulary follows the usage in mainland China.

Similarly, the term **Traditional Chinese** (TC) typically refers to a Chinese text that meets the following conditions:

1. **Character forms:** TC must be written with the traditional character forms.
2. **Character Sets:** TC normally uses the Big Five character set.
3. **Encoding:** TC is normally encoded in Big Five.
4. **Vocabulary:** Choice of vocabulary follows the usage in Taiwan or Hong Kong.

Only the first of these is a necessary condition. "Simplified" Chinese, by definition, cannot be written with the traditional character forms, except in those cases where a traditional form has no corresponding simplified form. Similarly, "Traditional" Chinese must not be written in the simplified forms, with some minor exceptions, such as in certain proper nouns.

There is also some variation in vocabulary usage.

Taiwanese texts, for example, may include some PRC-style vocabulary, while Singaporean texts may follow Taiwanese-style, rather than PRC-style, computer terminology. Nevertheless, on the whole, the terms Simplified Chinese and Traditional Chinese are used as defined above.

1.2 The Nature of the Problem

The forms of Chinese characters underwent a great deal of change over the several thousand years of their history. Many calligraphic styles, variant forms, and typeface designs have evolved over the years. The language reforms in the PRC have had a major impact on the Chinese written language. From the point of view of processing Chinese data, the most relevant issues are:

1. Many character forms underwent major simplifications, to the point where they are no longer recognizable from their traditional forms, e.g. TC 徵 → SC 征.
2. In numerous cases, one simplified form corresponds to two or more traditional forms (less frequently the reverse is also true), e.g. SC 征 maps to TC 徵 and 征. Normally only one of these is the correct one, depending on the context.
3. Sometimes, one simplified form maps to multiple traditional forms, *any* of which may be correct, depending on the context.
4. The GB 2312-80 standard used for SC is incompatible with the Big Five standard used for TC, resulting in numerous missing characters on both sides.

Item (2) above is the central issue in SC-to-TC conversion, and is what this paper focuses on. The "classical" example given in such discussions are the traditional characters 發 and 髮, which were merged into the single simplified form 发.

Table 1: SC-to-TC One-to-Many Mappings

SC Source	TC Target	Meaning	TC Example
发 fā	發	emit	出發
发 fà	髮	hair	頭髮
干 gān	乾	dry	乾燥
干 gàn	幹	trunk	精幹
干 gān	干	intervene	干涉
干 gàn	榦	tree trunk	樹榦
后 hòu	後	after	後天
后 hòu	后	queen	王后

Successfully converting such SC forms to their corresponding TC forms depends on the context, usually the word, in which they occur. Often, the conversion cannot be done by merely mapping one codepoint to another, but must be based on larger linguistic units, such as words.

There are hundreds of other simplified forms that correspond to two or more traditional ones, leading to ambiguous, one-to-many mappings that depend on the context. In this paper, such mappings may be referred to as **polygraphic**, since one simplified character, or *graph*, may correspond to more than one traditional (graphic) character, or vice versa.

2 The Four Conversion Levels

The process of automatically converting SC to TC (and, to a lesser extent, TC to SC) is full of complexities and pitfalls. The conversion can be implemented on four levels, in increasing order of sophistication, from a simplistic code conversion that generates numerous errors, to a sophisticated approach that takes the semantic and syntactic context into account and aims to achieve near-perfect results. Each of these levels is described below.

Table 4: The Four Conversion Levels

Level 1	**Code**	Character-to-character, *code*-based substitution
Level 2	**Orthographic**	Word-to-word, *character*-based conversion
Level 3	**Lexemic**	Word-to-word, *lexicon*-based conversion
Level 4	**Contextual**	Word-to-word, *context*-based translation

2.1 Level 1: Code Conversion

2.1.1 Basic Concepts

The easiest, but most unreliable, way to convert SC to TC, or vice versa, is to do so on a codepoint-to-codepoint basis by looking the source up in a hard-coded, one-to-one mapping table. This kind of conversion can be described as character-to-character, *code*-based substitution, and is referred to as **code conversion**, because the units participating in the conversion process are limited to single codepoints. The following is an example of a one-to-one code mapping table.

Table 5: Code Mapping Table

SC Source	GB0 (EUC)	TC Target	BIG Five	Omitted Candidates
出	B3F6	出	A558	齣
发	B7A2	發	B56F	髮
干	B8C9	幹	A47A	乾 干 榦
暗	B0B5	暗	B774	闇
里	C0EF	裡	B8CC	里 裏
征	D5F7	徵	BC78	征
汤	CCC0	湯	B4F6	

Since such tables map each source character to only one target character, the other possible candidates are ignored, which frequently results in incorrect conversion. For example, since SC 发 'hair' maps to both TC 髮 'hair' and TC 發 'emit', the conversion may fail. That is, if the table maps 发 to 發, which is often the case, the result will be the nonsensical 頭發. 'head' + 'emit.'

These problems are compounded if each element of a compound word maps to more than one character (polygraphic compounds), since the number of permutations grows geometrically, as shown in the table below.

Table 6: SC-to-TC Polygraphic Compounds

SC Source	Meaning	Correct TC	Other TC Candidates
特征	characteristic	特徵	特征
出发	start off	出發	出髮 齣髮 齣發
干燥	dry	乾燥	干燥 幹燥 榦燥
暗里	secretly	暗裡	暗里 闇里 闇裡 暗裏 闇裏
千里	long distance	千里	韆里 千裡 韆裡 千裏 韆裏
秋千	a swing	鞦韆	秋千 秋韆 鞦千

It is self-evident that, when there are several candidates to choose from, there is a high probability that a one-to-one code converter will output the incorrect combination.

2.1.2 The Conversion Process

Code conversion can be implemented in three different ways, in increasing order of sophistication:

1. **Simplistic conversion:** This refers to system based on one-to-one mapping tables in which the target codepoint is one of several alternatives selected without sufficiently considering its frequency of occurrence. Simplistic conversion frequently leads to unacceptable results, and requires considerable effort in human post-editing. Unfortunately, many conversion utilities take this approach.
2. **Frequency-based conversion:** This refers to a system based on one-to-one mapping tables in which the target codepoint is the *first* of several alternatives, selected from a list ordered by frequency of occurrence. Table 5 is an example of a frequency-based mapping table.

   Although this approach frequently leads to correct results, it is likely to fail in the many cases where the second (or third) alternative of multiple target mappings is itself of high frequency.
3. **Candidate-based conversion:** This refers to a system based on one-to-many mapping tables, with the alternative candidates listed in order of frequency of occurrence. In the case of one-to-many mappings, the user is presented with a list of candidates.

To sum up, code conversion has the following disadvantages:

1. If implemented as simplistic conversion, it will normally produce unacceptable results.
2. Even if implemented intelligently (approaches (2) and (3) above), it may require considerable human intervention in the form of candidate selection and/or post-editing.
3. It totally ignores differences in vocabulary (discussed below).

2.2 Level 2: Orthographic Conversion

2.2.1 Basic Concepts

The next level of sophistication in SC↔TC conversion can be described as word-to-word, *character*-based conversion. We call this **orthographic conversion,** because the units participating in the conversion process consist of

orthographic units: that is, characters or meaningful combinations of characters that are treated as single entries in dictionaries and mapping tables.

In this paper, we refer to these as **word-units.** Word-units represent meaningful linguistic units such as single-character words (free forms), word elements such as affixes (bound morphemes), multi-character compound words (free and bound), and even larger units such as idiomatic phrases. For brevity, we will sometimes use *word* as a synonym for *word-unit* if no confusion is likely to arise.

2.2.2 The Conversion Process

Orthographic conversion is carried out on a word-unit basis in four steps:

1. Segmenting the source sentence or phrase into word-units.
2. Looking up the word-units in orthographic (word-unit) mapping tables.
3. Generating the target word-unit.
4. Outputting the target word-unit in the desired encoding.

For example, the SC phrase 梳头发 (shū tóufa) 'comb one's hair,' is first segmented into the word-units 梳 'comb' (single-character free morpheme) and 头发 'hair' (two-character compound), each is looked up in the mapping table, and they are converted to the target string 梳頭髮. The important point is that 头发 is *not* decomposed, but is treated as a single word-unit.

Table 7: Orthographic Mapping Table

SC Word-Unit	TC Word-Unit	Pinyin	Meaning
头发	頭髮	tóufa	hair
特征	特徵	tèzhēng	characteristic
出发	出發	chūfā	start off
干燥	乾燥	gānzào	dry
暗里	暗裡	ànlǐ	secretly
千里	千里	qiānlǐ	long distance
秋千	鞦韆	qiūqiān	a swing

2.3 Level 3: Lexemic Conversion

2.3.1 Basic Concepts

Orthographic conversion works well as long the source and target words are in **orthographic correspondence**, as in the case of SC 头发 and TC 頭髮. Unfortunately, Taiwan, Hong Kong, and the PRC have sometimes taken different paths in coining technical terminology. As a result, there are numerous cases where SC and TC have entirely different words for the same concept.

The next level of sophistication in SC↔TC conversion is to take these differences into account by "translating" from one to the other, which can be described as word-to-word, *lexicon*-based conversion. We call this **lexemic conversion,** because the units participating in the conversion process consist of semantic units, or *lexemes*.

A **lexeme** is a basic unit of vocabulary, such as a single-character word, affix, or compound word. In this paper, it also denotes larger units, such as idiomatic phrases. For practical purposes, it is similar to the word-units used in orthographic conversion, but the term *lexeme* is used here to emphasize the semantic nature of the conversion process.

2.3.2 The Conversion Process

Let us take the SC string 信息处理 (xìnxī chǔlǐ) 'information processing', as an example. It is first segmented into the lexemes 信息 and 处理, each is looked up in a lexemic mapping table, and they are then converted to the target string 資訊處理 (zīxùn chǔlǐ).

It is important to note that 信息 and 資訊 are *not* in orthographic correspondence; that is, they are distinct lexemes in their own right, not just orthographic variants of the same lexeme. This is not unlike the difference between American English 'gasoline' and British English 'petrol'. The difference between 处理 and 處理, on the other hand, is analogous to the difference between American English 'color' and the British English 'colour'. This analogy to English must not be taken too literally, since the English and Chinese writing systems are fundamentally different.

Lexemic conversion differs from orthographic conversion in two important ways:

1. The mapping tables must map one lexeme to another on a semantic level, if appropriate. For example, SC 计算机 must map to its TC lexemic equivalent 電腦.
2. The segmentation algorithm must be sophisticated enough to identify proper nouns, since the choice of target character could depend on whether the lexeme is a proper noun or not.

Table 8: Lexemic Mapping Table

English	SC Lexeme	SC Pinyin	TC Lexeme	TC Pinyin
Bit	位	wèi	位元	wèiyuán
Byte	字节	zìjié	位元組	wèiyuánzǔ
CD-ROM	光盘	guāngpán	光碟	guāngdié
Computer	计算机	jìsuànjī	電腦	diànnǎo
Database	数据库	shùjùkù	資料庫	zīliàokù
File	文件	wénjiàn	檔案	dàng'àn
Information	信息	xìnxī	資訊	zīxùn
Internet	因特网	yīntèwǎng	網際網路	wǎngjì-wǎnglù
Software	软件	ruǎnjiàn	軟體	ruǎntǐ
Week	星期	xīngqī	禮拜	lǐbai

2.3.3 Proper Nouns

Another aspect of lexemic conversion is the treatment of proper nouns. The conversion of proper nouns from SC to TC, and vice versa, poses special problems, both in the segmentation process, and in the compilation of mapping tables. A major difficulty is that many non-Chinese (and even some Chinese) proper nouns are not in orthographic correspondence. In such cases, both code converters and orthographic converters will invariably produce incorrect results.

The principal issues in converting proper nouns are:

1. **Segmentation:** The segmentation algorithm must be sophisticated enough to identify proper nouns, since the choice of target character(s) could depend on whether the lexeme is a proper noun or not.

2. **Non-Chinese names:** For some non-Chinese proper nouns, TC and SC use different characters. For example, SC 肯尼迪 (kěnídí), a transliteration of 'Kennedy', maps to TC 甘迺迪 (gānnǎidí). Note how 肯 and 尼 do *not* orthographically correspond to 甘 and 迺.

3. **Two-dimensional mappings:** Sometimes, a source must map to a target along two dimensions: ordinary vocabulary and proper nouns. For example, SC 周 maps to either TC 周 or 週 (or even 賙) in ordinary words, but only to 周 in personal names.

Table 9: Lexemic Mapping Table for Non-Chinese Names

English	SC Source	Correct TC	Incorrect TC
Berlin Wall	柏林墙	柏林圍牆	柏林牆
Chad	乍得	查德	乍得
Oahu	瓦胡岛	歐胡島	瓦胡島
Kennedy	肯尼迪	甘迺迪	肯尼迪
Wisconsin	威士康星	威士康辛	威士康星

2.4 Contextual Conversion

2.4.1 Basic Concepts

The highest level of sophistication in SC↔TC conversion can be described as word-to-word, *context*-based translation. We call this **contextual conversion,** because the semantic and syntactic context must be analyzed to correctly convert certain ambiguous polysemous lexemes that map to multiple target lexemes.

As we have seen, orthographic converters have a major advantage over code converters in that they process

word-units, rather than single codepoints. Thus SC 特征 (tèzhēng) 'characteristic', for example, is correctly converted to TC 特徵 (not to the incorrect 特征). Similarly, lexemic converters process lexemes. For example, SC 光盘 (guāngpán) 'CD-ROM' is converted to the lexemically equivalent TC 光碟 (guāngdié), *not* to its orthographically equivalent but incorrect 光盤.

This works well most of the time, but there are special cases in which a polysemous SC lexeme maps to multiple TC lexemes, *any* of which may be correct, depending on the semantic context. We will refer to these as **ambiguous polygraphic compounds.**

One-to-many mappings of polysemous SC compounds occur both on the orthographic level and the lexemic level. SC 文件 (wénjiàn) is a case in point. In the sense of 'document', it maps to itself, that is, to TC 文件; but in the sense of 'data file', it maps to TC 檔案 (dàng'àn). This could occur in the TC-to-SC direction too. For example, TC 資料 (zīliào) maps to SC 资料 in the sense of 'material(s); means', but to SC 数据 (shùjù) in the sense of 'data'.

2.4.2 The Conversion Process

To our knowledge, converters that can automatically convert ambiguous polygraphic compounds do not exist. This requires sophisticated technology that is similar to that used in bilingual machine translation. Such a system would typically be capable of parsing the text stream into phrases, identifying their syntactic functions, segmenting the phrases into lexemes and identifying their parts of speech, and performing semantic analysis to determine the specific sense in which an ambiguous polygraphic compound is used.

The CDPS is currently developing a "pseudo-contextual" conversion system that offers a partial solution to this difficult task. It does not do syntactic and semantic analysis, but aims to achieve a high level of accuracy by a semi-automatic process that requires user interaction. To this end we are:

1. Building a database of one-to-many mappings for ambiguous polygraphic compounds.
2. Developing a user interface that allows the user to manually select from a list of candidates.

The following is an example of a mapping table for ambiguous polygraphic compounds, both on the orthographic and the lexemic levels.

Table 11: Ambiguous Polygraphic Compounds

SC Source	TC Alternative 1	TC Alternative 2
编制	編制 organize; establish	編製 make by knitting
制作	制作 creation (music etc.)	製作 manufacture
白干	白幹 do in vain	白干 strong liquor
阴干	陰乾 let pickles dry	陰干 even numbers
文件	檔案 (data) file	文件 document

2.4.3 The Ultimate Converter

Our ultimate goal is to develop a contextual converter that will achieve near-perfect conversion accuracy. Such a converter should be capable of, among other things, to:

1. Perform sophisticated parsing based on syntactic and semantic analysis.
2. Identify proper nouns and other parts of speech.
3. Include comprehensive, frequency-based, one-to-many code mapping tables.
4. Include comprehensive orthographic and lexemic one-to-many mapping tables.
5. include comprehensive two-dimensional, one-to-many mapping tables for proper nouns.
6. Automatically convert polygraphic lexemes, including ambiguous polygraphic compounds.
7. Operate in batch mode or through user interaction.

3 Discussion and Analysis

3.1 SC-to-TC Conversion Sample

Following is an example of SC-to-TC lexemic (Level 3) conversion.

Simplified Chinese

根据《计算机周报》的报道，瓦胡岛软件研究所所长威廉肯尼迪氏和广东大学的信息处理研究所所长周东丰教授在香港举办了关于"因特网的现状"及"信息高速公路的未来"的发表会，并且对于明年两研究所将合并开发的因特网信息数据库进行了讨论。

Traditional Chinese

根據《電腦週報》的報導，歐胡島軟體研究所所長威廉甘迺迪氏和廣東大學的資訊處理研究所所長周東豐教授在香港舉辦了關於“網際網路的現狀”及“資訊高速公路的未來”的發表會，並且對於明年兩研究所將合併開發的網際網路資訊資料庫進行了討論。

English Translation

According to the *Computer Weekly*, the director of the Oahu Software Research Institute William Kennedy, and the director of Canton University's Information Processing Institute Professor Dongfeng Zhou, held a press conference in Hong Kong on the topics "The Internet Today" and "The Future of the Information Superhighway." They also discussed the plans of both institutes to build a "Database of Internet Information."

The above passage has several interesting features that demonstrate the principal challenges that must be overcome to achieve near-perfect conversion. Below we will examine the various issues related to the conversion process for each of the first three levels.

3.2 Code Conversion Issues

Let us first consider what would happen if the above passage were converted with a plain code converter. We did this with a popular wordprocessor developed by a Chinese university, and got the following (highly unacceptable) results:

根據《[計算機]{周報}》的[報道]，[瓦胡島][軟件]研究所所長威廉[肯尼迪]氏和廣東大學的[信息]處理研究所所長周{東丰}教授在香港舉辦了{關于}“[因特網]的現狀”及“[信息]高速公路的未來”的發表會，{并且}{對于}明年兩研究所將{合并}開發的[因特網][信息][數據庫]進行了討論。

The above brief passage contains six orthographic errors, enclosed in braces, and 11 lexemic errors, enclosed in square brackets. 29 out of 105 characters, or about 28%, were converted incorrectly.

Some compounds containing polygraphic characters, such as SC 发, were sometimes converted correctly, as in the case of 发表 to 發表. But in other cases, as in SC 周, they were often converted incorrectly, as happened with 周报 being converted to 周報, as well as in five other cases.

3.3 Orthographic Conversion Issues

The failure to convert SC 周报, 并且 and other words correctly could be resolved by using Level 2 orthographic conversion.

Using mapping tables ensures correct conversion on a word-unit level, and avoids the problems inherent in one-to-one code converters.

Table 12: Orthographic Equivalents

SC Source	TC Target	Pinyin	English
大学	大學	dàxué	university
举办	舉　辦	jǔbàn	conduct, hold
所长	所長	suǒzhǎng	chief
处理	處理	chǔlǐ	processing
东丰	東豐	dōngfēng	Donfgeng (a name)
周报	週報	zhōubào	weekly publication
并且	並且	bìngqiě	moreover
合并	合併	hébìng	merge
关于	關於	guānyú	about, concerning
对于	對於	duìyú	regarding

3.4 Lexemic Conversion Issues

There are also many non-Chinese proper nouns that are not transliterated with the same characters, (e.g. 瓦胡岛 for "Oahu"). As the "Correct" column in the table below shows, all the SC lexemes and proper nouns that are not in orthographic correspondence with their TC equivalents were converted incorrectly.

Table 13: Lexemic Equivalents

English	SC Lexeme	SC Pinyin	TC Lexeme	TC Pinyin	Correct
Computer	计算机	Jìsuànjī	電腦	diànnǎo	no
Database	数据库	Shùjùkù	資料庫	zīliàokù	no
Oahu	瓦胡岛	Wǎhúdǎo	歐胡島	ōuhúdǎo	no
Information	信息	xìnxī	資訊	zīxùn	no
Internet	因特网	yīntèwǎng	網際網路	wǎngjì-wǎnglù	no
Kennedy	肯尼迪	kěnnídí	甘迺迪	gānnǎidí	no
Report	报道	bàodào	報導	bàodǎo	no
Software	软件	ruǎnjiàn	軟體	ruǎntǐ	no

3.5 How Severe is the Problem?

What is the extent of this problem? Let us look at some statistics. A number of surveys have demonstrated that the 2000 most frequent SC characters account for approximately 97% of all characters occurring in contemporary SC corpora. Of these, 238 simplified forms, or almost 12%, are polygraphic: that is, they map to two or more traditional forms. This is a significant percentage, and is one of the principal difficulties in converting SC to TC accurately.

But these figures tell only part of the story, because they are based on single characters. To properly grasp the full magnitude of this problem, we must examine the occurrence of all word-units that contain polygraphic characters.

Some preliminary calculations based on our comprehensive Chinese lexical database, which currently contains more than 900,000 items, show that more than 20,000 of the approximately 97,000 most common SC word-units contain at least one polygraphic character, which leads to one-to-many SC-to-TC mappings. This represents an astounding 21%. A similar calculation for TC-to-SC mappings resulted in 3025, or about 3.5%, out of the approximately 87,000 most common TC word-units. These figures demonstrate that merely converting one codepoint to another leads to unacceptable results.

Since many high-frequency polygraphic characters are components of hundreds, or even thousands, of compound words, incorrect conversion will be a common occurrence unless the one-to-many mappings are disambiguated by (1) segmenting the byte stream into semantically meaningful units (word-units or lexemes) and, (2) analyzing the context to determine the correct choice out of the multiple candidates.

4 A New Conversion Technology

In 1996, the Tokyo-based **CJK Dictionary Publishing Society** (CDPS), which specializes in CJK computational lexicography, launched a project whose ultimate goal is to develop a Chinese-to-Chinese conversion system that gives near-perfect results. This has been a major undertaking that required considerable investment of funds and human resources.

To achieve a high level of conversion accuracy, our mapping tables are comprehensive, and include approximately 900,000 general vocabulary lexemes, technical terms, and proper nouns. They also include various other attributes, such as pinyin readings and parts of speech.

Below is a brief description of the principal components of the conversion system:

1. **Code mapping tables:** Our SC↔TC code mapping tables are comprehensive and cover all Unicode codepoints. In the case of one-to-many SC-to-TC mappings, the candidates are arranged in order of frequency based on statistics derived from a massive corpus of 170 million characters, as well as on several years of research by our team of TC specialists.
2. **Orthographic mapping tables:** Constructing accurate orthographic mapping tables for tens of thousands of polygraphic compounds requires extensive manual labor. Our team of specialists has been compiling such tables by examining and double-checking each word individually.
3. **Lexemic mapping tables:** Constructing accurate lexemic mapping tables is even more laborious, since there is no orthographic correspondence between the SC and TC characters, and since dictionaries showing SC/TC differences do not (seem to) exist. Each word must be examined individually, while taking into account the extra complications resulting from ambiguous polygraphic compounds
4. **Proper noun mapping tables:** Special treatment has been given to proper nouns, especially personal and

place names. Our mapping tables for Chinese and non-Chinese names currently contain about 500,000 items. Unlike lexemic tables, these tables present a special complication because of the need for two-dimensional mappings.

5. **Conversion Engine:** The conversion engine was developed by Basis Technology in collaboration with the CDPS. Its major components are: (1) a sophisticated **Chinese word segmenter,** which segments the text stream into word-units and identifies their grammatical functions, and (2) the **conversion module,** which looks up the word-units in the mapping tables and generates the output in the target encoding.

4.3 Conclusions

Chinese to Chinese conversion has become increasingly important to the localization, translation, and publishing industries, as well as to software developers aspiring to penetrate the East Asian market. But, as we have seen, the issues are complex and require a major effort to build mapping tables and to develop segmentation technology.

The CJK Dictionary Publishing Society finds itself in a unique position to provide software developers with high quality Chinese lexical resources and reliable conversion technology, thereby eliminating expensive manual labor and significantly reducing costs. We are convinced that our ongoing research and development efforts in this area are inexorably leading us toward achieving the elusive goal of building the perfect converter.

Bibliography

Halpern, Jack (1990). "New Japanese-English Character Dictionary: A Semantic Approach to Kanji Lexicography" Euralex '90 Proceedings. Actas del IV Congreso Internacional. 157-166. Benalmádena (Málaga): Bibliograf.

Halpern, Jack (1990). New Japanese-English Character Dictionary (Sixth Printing). Tokyo: Kenkyusha.

Halpern, Jack, Nomura Masaaki, and Fukada Atsushi (1994). "Building a Comprehensive Chinese Character Database," Euralex '94 Proceedings. International Congress on Lexicography in Amsterdam.

Halpern, Jack (1998). "Building A Comprehensive Database for the Compilation of Integrated Kanji Dictionaries and Tools," 43rd International Conference of Orientalists in Tokyo.

Halpern, Jack (1999). The Kodansha Kanji Learner's Dictionary. Tokyo: Kodansha International.

Huang, Shih Kun (1994). Chinese Usenet Postings. Department of Computer Science and Information Engineering, National Chiao-Tung University, Taiwan (http://www.csie.nctu.edu.tw/).

ISO 2022:1994 Information Technology -- Character Code Structure and Techniques.

Lunde, Ken (1999). CJKV Information Processing. Sebastopol: O'Reilly & Associates.

Meyer, Dirk (1998). "Dealing With Hong Kong Specific Characters," Multilingual Computing & Technology, Vol. 9 No. 3. Multilingual Computing, Inc.

The Unicode Standard, Version 2.0. Reading: Addison-Wesley.

现代汉语频率词典 xiàndài hànyǔ pínlǜ cídiǎn (1986). Beijing: Beijing Language Institute.

国家语言文字工作委员会 (1986). 简化字总表 jiǎnhuàzì zǒngbiǎo (Second Edition): 语文出版社.