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

Contents

CJK Dictionary Publishing Society 日中韓字典刊行会
1-3-502 3-chome Niiza, Niiza-shi, Saitama 352-0006 JAPAN
Phone: +81-48-481-3103 Fax: +81-48-479-1323
E-mail: jack@cjk.org WWW: http://www.cjk.org

The Pitfalls and Complexities of Chinese to Chinese Conversion

Jack Halpern, Editor-in-Chief, CJK Dictionary Publishing Society
Jouni Kerman, Chief of Software Development, CJK Dictionary Publishing Society

0. Abstract

Standard Chinese is written in two forms: **Simplified Chinese** (SC), used in the PRC and Singapore, and **Traditional Chinese** (TC), used in Taiwan, Hong Kong, Macao, and among most overseas Chinese. 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, such as from GB 2312-80 to Big Five.

Although many code conversion tools are mere implementations of this fallacy, nothing can be further from the truth. There are major differences between the systems on various levels: character sets, encoding methods, orthography (choice of characters), vocabulary (choice of words), and even semantics (word meanings).

With the growing importance of East Asia in the world economy, localization and translation companies face an urgent need to convert between SC and TC, but must contend with such obstacles as: (1) current conversion tools produce unacceptable results, (2) the lack of knowledge to develop good conversion tools, (3) no access to high quality dictionary data, and (4) the high cost of manual conversion.

In 1996, the **CJK Dictionary Publishing Society** (CDPS) launched a project to investigate these issues in-depth, and to build a comprehensive SC<>TC database (now at 700,000 items and growing) whose goal is to enable conversion software to achieve near 100% accuracy. The CDPS is collaborating with **Basis Technology** in developing the sophisticated segmentation technology required to achieve this.

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

1. Introduction

1.1 Historical Background

The forms of Chinese characters (汉字 hànzì) 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. Some of the full, complex forms were elevated to the rank of "correct characters" (正字 zhèngzì), while the bewildering plethora of variants were often relegated to the status of "vulgar forms" (俗字 súzì).

Soon after the establishment of the People's Republic of China in 1949, the new regime launched a vigorous campaign to implement large-scale written language reforms. In the 1950s, Mao Zedong and Zhou Enlai led the way by announcing that character simplification was a high priority task. In 1952, the Committee on Language Reform was established to study the problem in-depth, and to undertake the task of compiling lists of simplified characters.

As a result of these activities, various written language reforms were undertaken, the most important of which include: the development of a standardized romanization system known as *pinyin*, limiting the number of characters in daily use, and the drastic simplification of thousands of character forms. Although at one point the ultimate goal was to abolish the use of Chinese characters altogether and to replace them with a romanized script, this policy was abandoned in favor of character form simplification.

Various simplified character lists were published in the subsequent years, the most well-known of which is the "definitive" **Comprehensive List of Simplified Characters** (简化字总表 jiănhuàzì zŏngbiăo) published in 1964, which was reissued several times with minor revisions. The latest edition, published in 1986, lists 2244 simplified characters [Zongbiao 1986].

Taiwan and Hong Kong, and most overseas Chinese, did not follow the path of simplification. Taiwan, in particular, has adhered fairly strictly to the traditional forms. The Taiwanese Ministry of Education has published various character lists, such as the 常用國字標準字體表 (chángyòng guózì biāozhŭn zìtĭbiăo), which enumerates 4808 characters, as guidelines for correct character forms.

1.2 Simplified and Traditional Chinese

Although the most important difference between Simplified Chinese and Traditional Chinese lies in character form, there are, as we shall see, also differences in character sets, encoding methods, and choice of vocabulary.

From a practical point of view, the term **Simplified Chinese** 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** 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. Character sets and encoding methods are less restricted, as described in section 1.4 below.

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.

1.3 The Nature of the Problem

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, etymologically two distinct characters, which were merged into the single simplified form 发. The table below shows these and other examples of SC forms that map to multiple TC forms.

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

SC Source	TC Target	Meaning	TC Example
发 fā	發	emit	出發 start off
发 fà	髮	hair	頭髮 hair
干 gān	乾	dry	乾燥 dry
干 gàn	幹	trunk	精幹 able, strong
干 gān	干	intervene	干涉 interfere with
干 gàn	榦	tree trunk	楨榦 central figure
面 miàn	麵	noodles	湯麵 noodle soup
面 miàn	面	face	面具 mask
后 hòu	後	after	後天 day after tomorrow
后 hòu	后	queen	王后 queen

As can be seen, 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.

1.4 Character Sets and Encodings

This paper does not aim to present a detailed treatment of Chinese character sets and encoding methods. This can be found in Ken Lunde's outstanding book *CJKV Information Processing* [Lunde 1999]. This section gives only a brief overview of some of the important issues, since our main goal is to deal with the higher level linguistic issues.

SC typically uses the GB 2312-80 (GB0) character set, or its expanded version called GBK, and is typically encoded in EUC-CN. For Internet data transmission, it is often encoded in HZ, or in the older zW. TC is typically encoded in Big Five, and less frequently in EUC-TW based on the Taiwanese CNS 11643-1992 (Chinese National Standard) character set.

In Japan, some wordprocessors handle Chinese characters via the JIS X 0208:1997 character set plus extensions. Similarly, it is possible to encode Chinese in the Korean character set KS X 1001:1992. However, in neither case are sufficient numbers of TC or SC characters available to adequately serve for general Chinese usage. This by no means exhausts the list of character sets for encoding Chinese (CCCII is an older Taiwanese standard still in use), and shows how complicated the situation is.

From the point of view of SC<>TC code conversion, one major issue is that GB 2312-80 is incompatible with Big Five. The former contains 6763 characters, as opposed to 13,053 characters in the latter. Approximately one-third of the GB 2312-80 characters are simplified forms not present in Big Five. This leads to many missing characters on both sides, as shown in the table below.

Table 2: GB and BIG Five Incompatibilities

Hanzi	GB0 (EUC)	Big Five	Unicode
頭	*	C059	982D
發	*	B56F	767C
計	*	AD70	8A08
头	CDB7	*	5934
发	B7A2	*	53D1
计	BCC6	*	8BA1
干	B8C9	A47A	5E72
里	C0EF	A8BD	91CC

The difficulties in SC<>TC conversion are not limited to the GB 2312-80 and Big Five character sets. In fact, Big Five contains only a subset of traditional forms, while GB 2312-80, surprisingly, does not contain some simplified forms, as shown in the table below.

Table 3: SC-to-TC Mappings not in GB and Big Five

SC Unicode	SC Source	TC Target	TC Unicode
7EBB	纻	紵	7D35
8BEA	诪	譸	8B78
8D51	赑	贔	8D14
94D4	铔	錏	930F
9613	阓	闠	95E0
98CF	飏	颺	98BA
9978	饸	餄	9904
9A89	骉	驫	9A6B
9C97	鲗	鰂	9C02
9E40	鹀	鵐	9D50

The international standard ISO-2022:1994 [ISO 1994] attempted to address these incompatibility issues by establishing a portmanteau encoding system in which escape sequence mechanisms signal a switch between character sets, but this fell short of a complete solution.

The advent of the international character set Unicode/ISO 10646 [Unicode 1996] has solved many of the problems associated with SC<>TC code conversion. With a Unicode-enabled system, it is possible to represent all Big Five and GB 2312-80 codepoints, and to display them in the same document, since Unicode is a superset of both these standards. This greatly simplifies SC<>TC conversion at the codepoint level. Although there are some issues that still need to be addressed (e.g. numerous characters have been excluded from the current version [Meyer 1998]), Unicode has effectively solved the problems caused by incompatibility between the Big Five and GB 2312-80 character sets.

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; that is, to do a simple substitution by replacing a source codepoint of one character set (such as GB 2312-80 (EUC) `0xB9FA` for SC 国) with a target codepoint of another character set (such as Big Five `0xB0EA` for TC 國) 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. That is, the text stream is not parsed into higher level linguistic units, but is treated merely as a sequence of code values of discrete multiple-byte characters.

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
门	C3C5	門	AAF9	
汤	CCC0	湯	B4F6	

Since such tables map each source character to only one target character, the other possible candidates (shown in the “Omitted Candidates” column) are ignored, which frequently results in incorrect conversion.

For example, an SC string such as 头发 ‘hair’ is not treated as a single unit, but is converted character by character. Since SC 头 maps only to TC 頭, the conversion succeeds. On the other hand, 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.’ On the other hand, if the table maps 发 to 髮, 头发 will be correctly converted to 頭髮, but other common words, such as SC 出发 ‘depart’, will be converted to the nonsensical 出髮 ‘go out’ + ‘hair.’

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 chose from, there is a high probability that a one-to-one code converter will output the incorrect combination. This demonstrates that code conversion cannot be relied upon to give accurate results without (often significant) human intervention.

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. Its only advantage is that it is easy and inexpensive to implement.
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 (in section 2.1.1) 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, as in the case of 发, which maps to both TC 發 and 髮.

   We have investigated several systems based on the frequency approach, and found numerous errors and omissions. The greatest difficulty in building a frequency-based code converter is that accurate and comprehensive mapping tables, based on reliable statistics, did not hitherto exist, and require extensive research to develop. Appendix C shows an example of incorrect mappings found in a well-known converter, compared with the mapping tables developed by the CDPS.
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, either interactively in the user interface (UI), or as a list in brackets.

Several major Chinese electronic dictionaries and wordprocessors, which claim to support TC, seem to be based on the simplistic approach. Some Chinese input systems take an approach that combines both (1) and (2). Approach (3), which is implemented in one of our in-house code converters, is rarely found.

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. (Actually, this example is complicated by the fact that 梳頭 ‘comb one's hair’ is also a legitimate word-unit.)

The following is an example of an orthographic (word-unit) mapping table. Appendix B gives a more detailed table.

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	qiúqiān	a swing

It is important to note that in both code conversion and orthographic conversion, the results must be in **orthographic correspondence** with the source. That is, the source and target are merely orthographic variants of the same underlying *lexeme* (see section 2.3.1 below). This means that each source character must be either identical to, or in exact one-to-one correspondence with, the target character.

For example, in converting SC 计算机 (jìsuànjī) to TC 計算機 'computer', 计 corresponds to 計, 算 corresponds to 算 (identical glyph), and 机 corresponds to 機 on a one-to-one basis. No attempt is made to "translate" SC 计算机 to TC 電腦 (diànnǎo), as is done in lexemic (Level 3) conversion.

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. Probably the best known of these is *computer,* which is normally 计算机 (jìsuànjī) in SC and 電脳 (diànnǎo) in TC.

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.

In a sense, converting one lexeme to another is like translating between two languages, but we call it lexemic conversion, not "translation," since it is limited to words and phrases of closely-related varieties of a single standard language, and no change is made in the word order (as is done in normal bilingual translation).

2.3.2 The Conversion Process

Let us take the SC string 信息处理 (xīnxì 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', which are orthographic variants of the same lexeme. This analogy to English must not be taken too literally, since the English and Chinese writing systems are fundamentally different.

Lexemic conversion is similar to orthographic conversion, but differs from it 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, *not* to its orthographic equivalent 計算機.
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 (see section 2.3.3 below).

The following is an example of a lexemic mapping table.

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	xīnxì	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	lǐbaì

As can be seen, the above table maps the semantic content of the lexemes of one variety of Chinese to the other, and in that respect is identical in structure to a bilingual glossary.

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 (kěnnídí), 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.

Following is an example of a mapping table for non-Chinese names that are not in orthographic correspondence.

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
Georgia	佐治亚	喬治亞	佐治亞
Kennedy	肯尼迪	甘迺迪	肯尼迪
Wisconsin	威士康星	威士康辛	威士康星

There are numerous other examples of this kind. These differences are not only extremely interesting in themselves, but have practical consequences. That is, since code and orthographic converters ignore them, they produce the unacceptable results shown in the "Incorrect TC" column above

Below is an example of two-dimensional mappings, as explained in item (3) above:

Table 10: Two-Dimensional Mappings

SC Source	Pinyin	TC as Name	TC as Word
周	zhōu	周	周 週 賙
发	fā	發	發 髮
才	cái	才	才 纔

This means that SC 发, when used as a name, must always be converted to TC 發, never to TC 髮. This is quite difficult, since the segmenter must be intelligent enough to distinguish between a character used as a word as opposed to a proper noun. This is a complex issue that deserves a paper in its own right.

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.

The following is an SC sentence that will no doubt confuse even the most sophisticated conversion engine:

> 发！请发这封传真可以吗？ 发点了点头发了传真。
>
> Hey, Fa! Could you please send this fax?
> Fa nodded his head and sent the fax.

The most advanced converters today could not possibly do better than:

> 發！請發這封傳真可以嗎？ 發點了點頭髮了傳真。

A Chinese speaker will find it humorous that the converter confused the independent SC words 头 (tóu) 'head' and 发 (fā) 'send' with the compound word 头发 (tóufa) 'hair'. The ideal contextual converter should be able to identify these as independent words that happen to be contiguous, and, hopefully, should be able to generate the correct:

> 發！請發這封傳真可以嗎? 發點了點頭發了傳真。

Ironically, a simplistic code converter, precisely *because* it does not recognize word-units, will probably give the correct results in this case, but for the wrong reasons! Admittedly, this is a contrived example. But it is a perfectly natural Chinese sentence, and clearly demonstrates the pitfalls and complexities of Chinese to Chinese conversion.

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 (普通话简体字)

根据「计算机周报」的报道、「佐治亚软件研究所」所长的威廉肯尼迪氏、和广东大学的「信息处理研究所」所长的周东丰教授、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 (臺灣的國語繁體字)

根據「電腦週報」的報導、「喬治亞軟體研究所」所長的威廉甘迺迪氏、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 Georgia 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, which is an example of SC-to-TC lexemic conversion, 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]、「[佐治亞][軟件]研究所」所長的威廉[肯尼迪]氏、和廣東大學的「[信息]處理研究所」所長的周{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. For now, we will ignore lexemic errors (such as 计算机 being converted to 計算機), all of which were converted incorrectly. The table below shows the orthographic errors ("TC Result"), the correct TC equivalents, and other potential candidates.

Table 12: SC-to-TC Conversion Results

SC Source	TC Result	Correct TC	Correct	Other Candidates
所长	所長	所長	yes	
大学	大學	大學	yes	
香港	香港	香港	yes	
未来	未來	未來	yes	
发表	發表	發表	yes	發表 髮表 發錶 髮錶
东丰	東丰	東豐	no	東丰
周报	周報	週報	no	周報 賙報
并且	并且	並且	no	併且 并且
合并	合并	合併	no	合并 合並
关于	關于	關於	no	關于
对于	對于	對於	no	對于

SC compound words consisting of characters that map to only one TC character have only one TC candidate, and were therefore converted with 100% accuracy. Some compounds containing polygraphic characters, such as SC 发 (which maps to TC 發 and 髮), were sometimes converted correctly, as in the case of 发表 to 發表. But in other cases, as in SC 周 (which maps to TC 周, 週 and 賙), they were often converted incorrectly, as happened with 周报 being converted to 周報, as well as in five other cases.

The above analysis demonstrates how unreliable code conversion can be.

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. Such compounds are recognized as word-units by the segmenter, are looked up in the orthographic mapping tables, and are then unambiguously converted to their correct TC equivalents.

The following is an example of a table that maps SC word-units to TC word-units on the orthographic level.

Table 13: Orthographic Equivalents

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

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

3.4 Lexemic Conversion Issues

As we have seen, code and orthographic converters are incapable of dealing with lexemic differences, such as between SC 计算机 and TC 電腦, since these are distinct lexemes for the same concept. There are also many non-Chinese proper nouns that are not transliterated with the same characters. For example, SC 佐治亚 (zuŏzhìyà), a phonetic transliteration of ‘Georgia’, should map to TC 喬治亞 (qiáozhìyà), *not* to its orthographically equivalent 佐治亞.

As the “Correct” column in the table below shows, all the SC lexemes and proper nouns which are not in orthographic correspondence with their TC equivalents were converted incorrectly.

Table 14: Lexemic Equivalents

English	SC Lexeme	SC Pinyin	TC Lexeme	TC Pinyin	Correct
computer	计算机	jìsuànjī	電腦	diànnăo	no
database	数据库	shùjùkù	資料庫	zīliàokù	no
Georgia	佐治亚	zuŏzhìyà	喬治亞	qiáozhìyà	no
information	信息	xīnxì	資訊	zīxùn	no
Internet	因特网	yīntèwăng	網際網路	wăngjì-wănglù	no
Kennedy	肯尼迪	kĕnnídí	甘迺迪	gānnăidí	no
report	报道	bàodào	報導	bàodăo	no
software	软件	ruănjiàn	軟體	ruăntĭ	no

The above analysis demonstrates that the use of lexemic mapping tables is essential to the attainment of a high level of conversion accuracy.

3.5 TC-to-SC Conversion

The one-to-many mapping problem is not limited to the SC-to-TC direction. In fact, most of the difficulties encountered in SC-to-TC conversion are present in TC-to-SC conversion as well. However, the one-to-many mappings on the orthographic level are far less numerous in the TC-to-SC direction.

Nevertheless, we have found a few dozen polygraphic traditional characters that map to two simplified forms, as shown in the table below.

Table 15: TC-to-SC One-to-Many Mappings

TC Source	SC Target	Meaning	SC Example
著 zhe	着	particle	沿着
著 zhù	著	writings	著作
乾 gān	干	dry	干燥
乾 qián	乾	male	乾坤
徵 zhēng	征	go on journey	长征
徵 zhǐ	徵	ancient note	宫商角徵羽
於 yú	于	at, in	关于
於 yú	於	Yu (a surname)	於先生

Some of these characters, such as TC 著 which maps to SC 著 and 着, are of high frequency and are found in hundreds of compound words, so that TC-to-SC conversion is not as trivial as may appear at first sight.

It is worthwhile noting that TC-to-SC mappings are not always reversible. For example, SC 后 (hòu) 'after; queen' maps to both TC 後 (hòu) ‘after’ and to TC 后 (hòu) ‘queen’, whereas the SC surname 後 maps only to TC 後. This means that SC-to-TC mapping tables must be maintained separately from TC-to-SC mapping tables.

3.6 How Severe is the Problem?

What is the extent of this problem? Let us look at some statistics. A number of surveys, such as [Xiandai 1986], 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.

Going in the other direction, from TC to SC, the scope of the problem is much more limited, but we have found that about 20 of the 2000 most frequent Big Five characters, based on a corpus of more than 170 million TC characters [Huang 1994], map to multiple SC characters.

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 700,000 items [Halpern 1994, 1998], 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, especially in the SC-to-TC direction, will lead 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

4.1 Project Overview

In 1996, the Tokyo-based **CJK Dictionary Publishing Society** (CDPS), which specializes in CJK computational lexicography [Halpern 1994, 1998], 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 this end, we have engaged in the following research and development activities:

1. In-depth investigation of all the technical and linguistic issues related to Chinese to Chinese conversion.
2. Construction of SC<>TC mapping tables for the first three conversion levels.
3. Development of Chinese word segmentation technology in collaboration with Basis Technology (Cambridge, Massachusetts) .

To achieve a high level of conversion accuracy, our mapping tables are comprehensive, and include approximately 700,000 general vocabulary lexemes, technical terms, and proper nouns. They also include various other attributes, such as pinyin readings, grammatical information, part of speech, and semantic classification codes.

4.2 System Components

Below is a brief description of the principal components of the conversion system, especially of our mapping tables:

1. **Code mapping tables:** Our SC<>TC code mapping tables are comprehensive and complete. They are not restricted to the GB 2312-80 and Big Five character sets, but 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. See Appendix A for an example.
2. **Orthographic mapping tables:** Constructing accurate orthographic mapping tables for tens of thousands of polygraphic compounds requires extensive manual labor. Our team

of TC specialists has been compiling such tables by examining and double-checking each word individually. See Appendix B for an example.

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 (see section 2.4.2). See section 2.3.2 for an example.
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 270,000 items. Unlike lexemic tables, these tables present a special complication because of the need for two-dimensional mappings. See section 2.3.3 for details and an example.
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.

Acknowledgements

We would like to extend our heartfelt gratitude to the various individuals who have contributed to this paper by reviewing it and offering their comments and constructive criticisms. This includes, in alphabetical order, Glenn Adams, James Breen, Carl Hoffman, Timothy Huang, Ken Lunde, Dirk Meyer, Frank Qian, Tsuguya Sasaki, David Westbrook, and Christian Wittern. Several members of the review team are recognized authorities in the field of CJK information processing.

Special recognition is due to Glenn Adams and James Breen, who have reviewed the paper carefully and made many invaluable suggestions.

References

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

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

[Halpern 1998] **Halpern, Jack** (1998): “Building A Comprehensive Database for the Compilation of Integrated Kanji Dictionaries and Tools,” 43rd International Conference of Orientalists in Tokyo.

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

[Huang 1994] **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 1994]: *ISO 2022:1994 Information Technology -- Character Code Structure and Techniques.*

[Lunde 1999] **Lunde, Ken** 1999: *CJKV Information Processing.* Sebastopol: O'Reilly & Associates.

[Meyer 1998] **Meyer, Dirk** (1998): “Dealing With Hong Kong Specific Characters,” *Multilingual Computing & Technology,* Vol. 9 No. 3. Multilingual Computing, Inc.

[Unicode 1996]: *The Unicode Standard, Version 2.0.* Reading: Addison-Wesley.

[Xiandai 1986] 现代汉语频率词典 xiàndaì hànyǔ pínlǜ cídiǎn (1986). Beijing: Beijing Language Institute.

[Zongbiao 1986]: 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.

Appendixes

Appendix A: Code Conversion Mapping Tables

Table A-1: SC-to-TC Code Mapping Table

GB Code	Source SC	Target TC	Big Five Codes
B0B5	暗	暗 闇	B774 EEEE
B2C5	才	才 纔	A47E C5D7
B3D4	吃	吃 喫	A659 B3F0
B5D6	抵	抵 牴 觝	A9E8 ACBB DBD3
B6AC	冬	冬 鼕	A556 C35D
B7E1	丰	豐 丰 風	C2D7 A4A5 ADB7
B8F6	个	個 箇	ADD3 BAE7
C0DB	累	累 纍	B2D6 F5EC
C3B9	霉	霉 黴	BE60 C5F0
CAAC	尸	屍 尸	ABCD A472
D5F7	征	徵 征	BC78 A9BA
DAD6	谥	諡 謚	EBAC EEB0
F3BD	蠼	蠼 蠷	F96E F8BE

Table A-2: TC-to-SC Code Mapping Table

Big Five	Source TC	Target SC	GB0 (EUC)
AB5D	侷	局	BED6
ADB7	風	风 丰	B7E7 B7E1
B054	訊	讯	D1B6
B0A2	陝	陕	C9C2
B0AE	乾	干 乾	B8C9 C7AC
B16A	強	强	C7BF
B3CA	傘	伞	C9A1
B3F2	圍	围	CEA7
B6C4	傭	佣	D3B6
BAE0	箋	笺	BCE3
BBB1	跼	局	BED6
BC78	徵	征 徵	D5F7 E1E7
BECA	彊	强	C7BF
BFFD	錄	录	C2BC

Appendix B

Table B: Orthographic Conversion Mapping Table

SC Source	TC Target
暗杀	暗殺
暗码	暗碼
暗里	暗裡
暗昧	闇昧
幽暗	幽闇
霉菌	黴菌
霉雨	霉雨
霉菌	黴菌
特征	特徵
象征	象徵
秋征	鞦徵
长征	長征
出征	出征
累进	累進
系累	繫纍
丰姿	丰姿
丰韵	風韻

Appendix C

Table C: Some Incorrect Mappings in a Popular Converter

GB (EUC)	SC Source	Incorrect TC	Correct TC
B7E1	丰	丰 豐	豐 丰 風
C3B4	么	么	麼 么
D4C6	云	云 雲	雲 云
CAB2	什	什	甚 什
B6AC	冬	冬	冬 鼕
BCB8	几	几 幾	幾 几
BACF	合	合	合 閤
BAF3	后	后	後 后
B8B4	复	復 複	復 複 覆 复
CAAC	尸	尸	屍 尸
B8C9	干	干 幹	幹 乾 干 榦
D5F7	征	征	徵 征
B5D6	抵	抵	抵 牴 觝
BDDC	杰	杰 傑	傑 杰
B4D6	粗	粗	粗 麤
B7B6	范	范	范 範
B8B2	覆	覆	覆
C3B9	霉	霉	霉 黴

About the Authors

JACK HALPERN 春遍雀來 (ハルペン・ジャック)

Editor in Chief, **CJK Dictionary Publishing Society**
Editor in Chief, **Kanji Dictionary Publishing Society**
Research Fellow, **Showa Women's University**

Born in Germany in 1946, Jack Halpern lived in six countries and knows twelve languages. Fascinated by kanji while living in an Israeli kibbutz, he came to Japan in 1973, where he compiled the **New Japanese-English Character Dictionary** [Halpern 1990] for sixteen years. He is a professional lexicographer/writer and lectures widely on Japanese culture, is winner of first prize in the International Speech Contest in Japanese, and is founder of the International Unicycling Federation.

Jack Halpern is currently the editor-in-chief of the **Kanji Dictionary Publishing Society** (KDPS), a non-profit organization that specializes in compiling kanji dictionaries, and the head of the **CJK Dictionary Publishing Society** (CDPS), which specializes in CJK lexicography and the development of a comprehensive CJK database (DESK). He has also compiled the world's first Unicode dictionary of CJK characters.

Following is a list of Jack Halpern's principal publications in the field of CJK lexicography.

Halpern, Jack (1982): “Linguistic Analysis of the Function of Kanji in Modern Japanese,” 27th International Conference of Orientalists in Tokyo.

Halpern, Jack (1985): “Function of Kanji in Modern Japanese,” *Transactions of the International Conference of Orientalists in Japan.* The Tōhō Gakkai (The Institute of Eastern Culture). 27th International Conference of Orientalists in Japan in Tokyo.

Halpern, Jack (1985): “Kenkyusha's New Japanese-English Character Dictionary,” *Calico Journal*, December 1985.

Halpern, Jack (1987): 漢字の再発見 *Kanji no Saihakken* ‘Rediscovering Chinese Characters’. Tokyo: Shodensha.

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

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 (1993): *NTC's New Japanese-English Character Dictionary.* Chicago: National Textbook Company.

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 (1995): *New Japanese-English Character Dictionary, Electronic Book Edition.* Tokyo: Nichigai Associates.

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.

Halpern, Jack and **Kerman, Jouni** (1999): “The Pitfalls and Complexities of Chinese to Chinese Conversion,” Fourteenth International Unicode Conference in Boston.

Halpern, Jack: *Dictionary of Unified CJK Characters -- for the Unicode Standard*. Forthcoming.

JOUNI KERMAN 華留萬陽貳 (ケルマン？ヨウニ)

Chief of Software Development, **CJK Dictionary Publishing Society**
Research Fellow, **Showa Women's University**

Born in 1967 in Finland, Jouni Kerman took a broad interest in languages, computer programming, and economics since his early teens. Besides his native tongue Finnish, he has studied English, Swedish, French, German, Italian, Japanese, Mandarin and Cantonese. He has received a Monbusho scholarship for advanced studies of Japanese from the Japanese Ministry of Education in 1992, and in 1996 he graduated from Helsinki School of Economics and Business Administration with a Master's Degree.

In 1996, Jouni Kerman joined the Kanji Dictionary Publishing Society in Tokyo on a Research Fellow grant from Showa Women's University to develop a page composition system for *The Kodansha Kanji Learner's Dictionary* [Halpern 1999]. Since the completion of the project, he has been involved in developing CJK data processing systems and database design for the CJK Dictionary Publishing Society.

CJK Dictionary Publishing Society
日中韓字典刊行会

The **CJK Dictionary Publishing Society** (CDPS) consists of a small group of researchers that specialize in CJK lexicography. The society is headed by **Jack Halpern,** editor-in-chief of the *New Japanese-English Character Dictionary* (http://www.kanji.org), which has become a standard reference work for studying Japanese.

The principal activity of the CDPS is the **development and continuous expansion of a comprehensive database** that covers every aspect of how Chinese characters are used in CJK languages, including Cantonese. Advanced computational lexicography methodology has been used to compile and maintain a Unicode-based database that is serving as a source of data for:

1. Dozens of lexicographic works, including electronic dictionaries.
2. Pedagogical, linguistic and computational lexicography research.
3. CJK input method editors (IME), online translation tools, and search engines.

The database has currently about **1.7 million Japanese and about 700,000 Chinese items**, including detailed grammatical, phonological and semantic attributes for general vocabulary, technical terms, and hundreds of thousands of proper nouns. The single-character database covers every aspect of CJK characters, including frequency, phonology, radicals, character codes, and other attributes. See http://www.cjk.org/datasrc.htm for a list of data resources.

The CDPS has become one of the world's prime resources for CJK dictionary data, and is contributing to CJK information processing technology by providing software developers with **high-quality lexical resources,** as well as through its ongoing **research activities and consulting services.**

Visit the CDPS website at

http://www.cjk.org

CJK Dictionary Publishing Society 日中韓字典刊行会

1-3-502 3-chome Niiza, Niiza-shi, Saitama 352-0006 JAPAN

Phone:	+81-48-481-3103	Fax:	+81-48-479-1323
E-mail:	jack@cjk.org	WWW:	http://www.cjk.org

For more information, contact jack@cjk.org.