Hong Kong Cantonese Data Retrieval in Multilingual Perspectives: The Case of a Cantonese-Dagaare-English e-Lexicon

Sally Y.K. MOK
Department of Linguistics, The University of Hong Kong
Pokfulam, Hong Kong
sallymok@hkusua.hku.hk

Abstract
The release of the Hong Kong Supplementary Character Set (HKSCS) by the Hong Kong Special Administrative Region, in collaboration with the Chinese Language Interface Advisory Committee has provided solutions to input and display Hong Kong characters that were previously not included in the Big5 and ISO-10646/Unicode character sets. Although the enhanced character sets have improved the support of Hong Kong Cantonese characters, there are few internet lexical resources on Hong Kong Cantonese, not mentioning the lack of trilingual translation resources. This paper discusses the issue of Hong Kong Cantonese data processing in the case of a Cantonese-Dagaare-English electronic lexicon, in particular, how the existing web and database technology retrieves multilingual lexical entries in Cantonese, Dagaare and English. It investigates the practical procedures in the course of developing the Cantonese-Dagaare-English e-lexicon. It stresses the significance of further development of these language resources for social and cultural exchange among Hong Kong and many other multilingual societies.

Keywords: Character encoding, HKSCS, Big5, ISO-10646, Unicode, e-lexicon, fonts, multilingual data processing, character input, Cantonese

1 Introduction
With the globalization of the World Wide Web, many languages other than English are emerging on the internet to serve the need of non-English web users. Multilingual transfer has become a necessity for various kinds of governmental, business and educational purposes. Yunker (2004) contends that web globalization is in action and the trend shows no signs of slowing. He believes that it is the fastest-growing segment of the global language translation industry.

Chinese web users, comprising 13.7% of the total internet population, according to Global Reach Statistics 2004, rank the second largest internet population in the world (c.f. English internet users, 35.2%). This result reveals that there has been an enormous demand in using Chinese language online. As far as the economic development of China and Hong Kong is concerned, global communication has prompted further demand for typing and displaying Chinese characters on the internet both for the people in mainland China and Hong Kong, as well as Chinese communities in some other countries. Two encodings, namely Big5 and GBK were first established in Taiwan to handle the transmission of traditional and simplified Chinese characters respectively on the internet. Before the creation of the Hong Kong Supplementary Character Set (HKSCS), some unique characters that Hong Kong people use such as ‘啲’, ‘啫’, ‘嘝’ etc. could not be displayed on the screen because these characters were not included in these coded Chinese character sets.

In the development of the Cantonese-Dagaare-English e-lexicon, this paper addresses the following research questions: How do we store and retrieve Hong Kong characters together with other language data using the HKSCS? How do the existing web and database technology retrieve Cantonese and other languages? What are the limitations of the employed web and database technology involved in developing a trilingual e-lexicon?
In the following sections, we will introduce the Cantonese-Dagaare-English e-lexicon, followed by an overview of the HKSCS and the discussion of the storage and retrieval of Hong Kong Cantonese along with Dagaare and English characters in the trilingual e-lexicon.

2 The Cantonese-Dagaare-English e-Lexicon

The Cantonese-Dagaare-English electronic lexicon is located in the Department of Linguistics server, University of Hong Kong. This is an online version of the trilingual Cantonese-Dagaare-English lexicon in paper form published in 2004 whose 'objective was to compile a concise on-line dictionary or lexical resource of the basic vocabulary items in Dagaare with Cantonese and English equivalents' (Bodomo, 2004:9). The lexicon records lexical entries of three languages: Dagaare, Cantonese and English glosses in which the best possible equivalences of these languages were established since the paper lexicon was published. ‘This lexicon comprises more than 1200 entries or headwords. Within each headword or entry are several other words in paradigmatic relation to the headword. Thus the dictionary actually contains 3000-4000 words.’ (Bodomo, 2004:7).

Table (1) below lists out the general specifications of the Cantonese-Dagaare-English trilingual e-lexicon and diagram (2) is a snapshot of the e-lexicon.

<table>
<thead>
<tr>
<th>Server</th>
<th>Apache HTTP Server 1.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Server-side</td>
<td>PHP 4.0</td>
</tr>
<tr>
<td>Browser-side</td>
<td>(x)HTML, CSS</td>
</tr>
<tr>
<td>Database System</td>
<td>MySQL server 4.0.9 (phpMyAdmin as the administration tool)</td>
</tr>
<tr>
<td>Size of the lexicon</td>
<td>Approx. 3000-4000 entries including different verb forms (continues to enlarge)</td>
</tr>
<tr>
<td>Languages</td>
<td>Dagaare and English data sets:</td>
</tr>
<tr>
<td></td>
<td>Orthography and pronunciation;</td>
</tr>
<tr>
<td></td>
<td>Cantonese: Orthography, Jyutping</td>
</tr>
<tr>
<td></td>
<td>romanization and pronunciation</td>
</tr>
<tr>
<td>Search functions</td>
<td>Fuzzy search, Diacritics</td>
</tr>
<tr>
<td></td>
<td>Insensitive/Sensitive Search</td>
</tr>
<tr>
<td>Search Directions</td>
<td>3 ways:</td>
</tr>
<tr>
<td></td>
<td>English -&gt; Cantonese and Dagaare</td>
</tr>
<tr>
<td></td>
<td>Dagaare-&gt; English and Cantonese</td>
</tr>
<tr>
<td></td>
<td>Cantonese-&gt; English and Dagaare</td>
</tr>
<tr>
<td>Multimedia</td>
<td>YES (in the form of sound clips recorded by the native speaker of the corresponding languages)</td>
</tr>
<tr>
<td>Others</td>
<td>Dagaare input installation package and user guide (with Dagaare IME file *For Windows User)</td>
</tr>
</tbody>
</table>

The e-lexicon allows lexical transfer in three directions: English to Cantonese and Dagaare equivalents, Cantonese to Dagaare and English equivalents and Dagaare to Cantonese and English equivalents. Instead of locating the entries by the alphabetical order of Dagaare words in the paper lexicon, this online lexicon is searchable by web queries in Cantonese, English or Dagaare. Additionally, it can perform diacritics-sensitive or diacritics-insensitive search. In other words, the user can search for the lexical equivalents in Cantonese and English by submitting a Dagaare query (whether with diacritics or not) and the system will output the corresponding lexical translations in Cantonese and English.

3 Multilingual Transfer on the Web: Cantonese, Dagaare and English Lexical Data

The trilingual e-lexicon deals with the manipulation of three languages with two languages which have very different writing systems. Both of these languages contain some non-ASCII (American Standard Code for Information Interchange) characters which have to be encoded differently from English character repertoires. In developing the Cantonese-Dagaare-English e-lexicon, multilingual data storage and retrieval are two essential processes for information exchange on the web. These are not two
separate procedures, but counterparts that are highly dependent on each other. The retrieval of the data relies on how the data is stored. In the following section, we will pin down our discussion on the process of inputting and displaying the lexical entries of Cantonese and Dagaare, two non-ASCII encoded languages.

3.1 Cantonese Data Processing

Cantonese is a tonal language spoken by about 66 million people mainly in the south east of China, particularly in Hong Kong, Macau, Guangdong, Guangxi and Hainan. It is also spoken in Malaysia, Indonesia, Thailand, Singapore and among Overseas Chinese communities in many other countries (Ager, 2004). Colloquial Hong Kong Cantonese, the language data collected in this research, is written with a mixture of standard Chinese characters and hundreds of extra characters invented specifically for Cantonese. Most of these extra characters have been included in the Hong Kong Supplementary Characters Set (HKSCS). Colloquial Cantonese is an important idiosyncratic feature of a Chinese dialect and there is an essential need for representing them on the internet both for communicative purpose and language learning. The Hong Kong S.A.R Government and its collaborative parties have therefore attempted to provide Unicode-based solutions to support Hong Kong-specific Chinese language requirements. This research has explored the use of HKSCS in the development of the Cantonese multilingual e-lexicon.

Chinese characters including Hong Kong Cantonese are stored as double-bytes in computers, same as other ideographic writing systems such as East Asian languages like Japanese and Korean. These languages altogether are called CJK (Chinese, Japanese and Korean) languages. Big5, a legacy character set used to encode Traditional Chinese used in Taiwan and Hong Kong, while GB(Simplified Chinese, National Standard used for Chinese in the PRC and Singapore), JIS(Japanese Industry Standard) and EUC (Extended Unix Code, other variants) are coded character set families for encoding other CJK languages. However, this research, as we will see, has shown that the use of Big5 to encode traditional Chinese characters together with Dagaare and English on the same web document is problematic.

3.1.1 An Overview of HKSCS, the Hong Kong Supplementary Character Set

The first version of the Hong Kong Character Set, Government Common Character Set (GCCS), was developed in 1995. It was the initial attempt for the Hong Kong S.A.R Government in coordinating the use of user-defined Chinese characters as a supplement to the standard character set of Big5. In September 1999, Hong Kong S.A.R Government in collaboration with the Chinese Language Interface Advisory Committee published the Hong Kong Supplementary Character Set (HKSCS). This version of HKSCS contains 4,702 Chinese characters that are specific to the Hong Kong environment and that are used by the public and the Government in electronic communication and data exchange conducted in Chinese. In December 2001, a new version of the HKSCS, namely HKSCS-2001, was published by Government. The HKSCS-2001 contains 116 more Chinese characters in addition to those contained in the HKSCS which was published in 1999. These characters are defined in the user-defined area of the coding schemes used on the computers, Big5 and the ISO 10646/Unicode standard.

Big5 is a fixed 2-octet coding scheme which consists of 13,053 traditional Chinese characters. It is an industrial standard established before the invention of Unicode and is commonly used in Taiwan and the HKSAR. The ISO Working Group and the Unicode Consortium have created one universal standard, ISO-10646/Unicode for coding all major languages in the world, including traditional and simplified Chinese characters. For Big5, the HKSCS-2001 is defined in the User-Defined Areas and Vendor-Defined Areas and is fully compatible with the Government Common Character Set (GCCS). The additional characters in the HKSCS-2001 are defined in the Private Use Area of ISO 10646/Unicode. To improve the character set, Hong Kong S.A.R Government allows public application for the inclusion of new characters in the HKSCS-2001. The significance of HKSCS to facilitate

---

electronic communication and data exchange is indicated by the ongoing research by international computing corporations such as Oracle. Tanaka’s presentation on ‘Conversion Between HKSCS-2001 and Unicode -- Going Beyond Basic Multilingual Plane’ in the 23rd Internationalization & Unicode Conference (2003) has shown the effort of Oracle to implement HKSCS-2001, and her dedication to associated future research.

### 3.1.2 The Storage and Retrieval of Cantonese Characters

#### 3.1.2.1 The Language Input

As mentioned, Hong Kong Cantonese which contains an additional set of unique characters had not been encoded in Big5 or ISO-10646/Unicode until the Hong Kong S.A.R Government has established the HKSCS-2001 (GCCS, the previous version). These new characters are included in the two code allocation schemes, Big5 and Unicode/ISO-10646. The HKSCS-2001 package consists of the new character set, reference font and input software. A separate package has to be obtained from Microsoft Hong Kong if MS Access is used for database retrieval. This package contains important system settings necessary to support HKSCS so as to encode Cantonese data retrieved from the MS Access database (Note that MySQL does not require this package to be installed). With the HKSCS, the Hong Kong characters can be inputted and displayed on locale and non-locale platforms. The new Unicode Chinese input tool can be selected from the language toolbar (See 3a) and the new characters will be added to the Unicode (See 3b) and the existing Big5 input applications (See 3c). The language input of Cantonese Characters which cannot be encoded owing to the lack of the HKSCS will appear on the screen as empty boxes, ‘□’ (See 4). As seen in the example, the Cantonese character “俤” was missing from the coded character set, hence the lexical entry was displayed as “俽”, but not “俵”。

#### 3.1.2.2 Web Scripting and Database System: PHP & MS Access or PHP & MySQL?

This research has used PHP as the scripting language. It runs on the Web server when a page is requested via HTTP. It uses UTF-8 (multi-byte) as its internal encoding. Many technical writers credit PHP as an easy-to-use web development tool which has been designed from the outset for the web environment.

---

Dan (2001) comments that PHP is built with the needs of web developers and it can solve complex problem scenarios faster and easier than comparable technologies like Java. PHP, as he explains, is naturally more compatible with all operating systems than ASP.

The choice of the database system is based on its capability of handling multilingual data exchange with PHP. This paper is restricted to compare the capabilities of MS Access and MySQL associated with PHP to handle Cantonese, Dagaare and English lexical entries. Among some other database drivers and tools options such as Oracle, PostgreSQL and SQL server, other factors may include the amount of simultaneous data access, security control, ease of user interface but they are not the scope of this research.

Although there are other existing web scripting and database systems, it is not clear which of these tools can be used jointly to produce the most effective result in storing and retrieving the lexical data of Cantonese and other languages. This research has performed a series of tests on PHP and two database systems, MS Access and MySQL on the Windows platform.

(A) MS Access
The database has been set up in MS Access on Apache server 1.3 with ODBC connection. But there exists several problems:

- Accents of Dagaare characters cannot be inputted directly into the database.
- Cantonese characters cannot be retrieved properly if the HKSCS support package for Windows is not properly installed.
- Although non-ASCII Dagaare letters such as accented open-e and open-o can be inputted into the MS Access database, the browser cannot encode these characters correctly as when they are retrieved on the web.
- The Unicode characters typed in the query which are not supported by Big5 is converted into numeric character references (NCRs). Thus, string matching will fail unless the database stores NCRs.

These problems are mainly due to the fact that MS Access uses UTF-16 which uses two 8-bit bytes for each code point. The retrieval of Cantonese characters is automatically converted into Big5 encoding (non-unicode legacy encoding) when UTF-16 encoded data in MS Access is called by the PHP program which uses UTF-8. The problem of Big5 is, however, that it does not contain all the Unicode characters. Hence, some of the Dagaare characters including diacritics and IPA symbols (such as open-e and open-o) cannot be displayed correctly as shown in (5).

(5)

<table>
<thead>
<tr>
<th>Cantonese translation:</th>
<th>Dagaare translation:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Characters: 講大話</td>
<td>agmaʔaʔ nọiʔiʔiʔiʔ</td>
</tr>
<tr>
<td>Romanization: tgamn ziri, idiom xkom</td>
<td></td>
</tr>
<tr>
<td>Pronunciation:</td>
<td>Pronunciation:</td>
</tr>
</tbody>
</table>

This observation confirms the view of Addison P. (Chair, W3C Internationalization Working Group) in the Yahoo internationalization (i18n) discussion forum that MS Access provides a special “wide char” set of functions that work with UTF-16 only, not UTF-8 or code page 65001 (UTF-8’s Windows code page number). He explains that Windows has itself support for non-Unicode legacy
multi-byte encodings (or code pages) and automatically converts from UTF-16 to these code pages when these methods are called.

Therefore, as in the case of this e-lexicon, we still can see these characters correctly displayed in MS Access. Owing to the limitation of MS Access in using double-byting (UTF-16) data storage, however, PHP cannot access MS Access data by its multibyte method. As a result, the data is converted back into its non-Unicode encoding, Big5 and hence Cantonese characters can be encoded correctly but non-ASCII accented Dagaare characters have become empty boxes ‘ ’ or question marks ‘??’, indicating a mismatch of encodings. Encoding problems on Cantonese occurs if Big5 is not specified (See samples on Internet Explorer 6.0: 6a (Big5 encoding); 6b (UTF-8 & UTF-16 encoding)). Other browsers such as Mozilla, Firefox, Netscape and Opera have the same encoding failures but the appearance of the improperly encoded characters may be different.

(6)

(a)  
2. to look for, search; to love  

<table>
<thead>
<tr>
<th>Cantonese translation:</th>
<th>Dagaare translation:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Characters: 譜；愛 / 愛</td>
<td>b? v, b? v</td>
</tr>
<tr>
<td>Traditional: wan2, ai3 / xung1</td>
<td>b? v, b? v</td>
</tr>
</tbody>
</table>

(b)  
2. to look for, search; to love  

<table>
<thead>
<tr>
<th>Cantonese translation:</th>
<th>Dagaare translation:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Characters: 粟 / ??n</td>
<td>̀v y</td>
</tr>
<tr>
<td>Traditional: wan2, ai3 / xung1</td>
<td>̀v y</td>
</tr>
</tbody>
</table>

The solution to this, as suggested by Addison P., is to find a suitable driver for the PHP installation that claims to support UTF-16 in Access. However, he does not recommend such an approach because compatibility has not been guaranteed and he doubts the portability of this PHP program with the new driver to platforms other than Windows.

B) MySQL  
This research has opted to use MySQL with PHP. The problem of encoding experienced with MS Access no longer exists with the use of MySQL. This is because both MySQL and PHP support UTF-8 encodings.

UTF-8 and UTF-16 can encode a full set of characters, hence it is not a matter of which encoding scheme is used that has caused errors. The central technique is to ensure that both the database system and the web scripting (or possibly with the help of a database connection driver) use the same encodings (same byte sequence in data storage and retrieval). A disparity of encodings often results in unknown or hardly identifiable display problems.

3.1.2.3. Methods of Encoding: ISO-10646/Unicode or Big5?

ISO-10646/Unicode and Big5 are able to handle Cantonese characters. However, when Cantonese characters are encoded with Big5 together with other languages (in this case Dagaare and English), it is likely to cause encoding problems. As we see from MS Access data retrieval, Big5 is limited in the number of characters coded. It fails to encode characters such as open-e, “é” and open-o, “ö”. Conversely, the use of UTF-8, a character encoding form (CEF) or Unicode Transformation Format (UTF) of Unicode, can encode all the three languages and well-supported by PHP and MySQL. It is ‘typically the preferred encoding form for HTML and similar protocols, particularly for the internet…it preserves ASCII transparency, making use of 8-bit code units’ (The Unicode Standard 4.0, 2004:30). The multilingual capacity of UTF-8 enables consistency and avoids ambiguity in encoding texts. Besides, the latest version of Unicode 4.0 does not require the installation of HKSCS package for displaying Cantonese characters. As far as accuracy and efficiency are concerned, Unicode is more preferred in the transmission of Cantonese data, along with other languages on the web.
3.2 **Dagaare Data Processing**

This section briefly discusses the approach used to process Dagaare data in the e-lexicon. Dagaare is a two-tone language spoken in West Africa. Accent is important in this language for the disambiguation of meanings in many cases. To preserve the meanings of words, this research explores ways to represent diacritics. Yet, some of the Dagaare characters consist of IPA Symbols and combining diacritics which cannot be easily typed and encoded by ASCII coding.

3.2.1 **The Storage and Retrieval of Dagaare with Cantonese lexical data**

In my previous work with Bodomo, Marfo and Cunningham (2004), we point out that with the ever-growing technological advancements on font representations, a keyboard for Dagaare input is long overdue. We compare two software applications towards developing a Unicode keyboard layout for Dagaare (and Twi) input. We have created a Unicode keyboard input system for typing Dagaare characters. In this e-lexicon, Microsoft Keyboard Layout Creator is used to develop a Dagaare input application. This software provides an application interface that helps with the assignment of Unicode code points (or scalar values) on keystrokes. It also assists in building a windows installer package for distribution. The Dagaare input installation package and user guide are available in the e-lexicon. This research has explored one of the effective solutions among other possibilities such as Java Multilingual Input Method Engine proposed by Leung, Liu & Wu (1998).

Regarding the display of Dagaare characters, the installation of one of the Unicode compliant fonts such as Lucida Sans Unicode, Gentium and Doulos SIL is necessary for displaying these non-Roman characters or glyphs correctly on the browsers. Although Big5 can encode Chinese including Hong Kong Cantonese characters, some of the Dagaare characters are defined only in Unicode. Therefore, as this research has shown, Unicode offers a better solution for retrieving and displaying Cantonese together with Dagaare characters, and perhaps other languages as well.

4 **Limitations and Findings**

This research has investigated the practicality and effectiveness of web technologies and database systems including PHP, MySQL, MS Access and Microsoft Keyboard Layout Creator in facilitating multilingual lexical data exchange in Cantonese, Dagaare and English on Windows Platforms. Yet, this research can be enhanced by examining some other tools such as Java, C, SQL server and Oracle, etc. and other platforms such as Mac OS, Unix, Linux so as to give a fuller scenario of the current multilingual computing environment. This research will continue to improve the search functions and query interface. Further development may aim at a larger lexicon that contains translation in more languages including minority languages, hence providing a richer lexical resource for more advanced and in-depth analysis.

Despite these limitations, this research has put forward the following preliminary observations:

i. Does the Hong Kong S.A.R. Government website contain enough information on developing a Cantonese database for multilingual web retrieval? It is not clear from the website that one has to obtain a Microsoft systems package that supports the HKSCS character set in order to use it for MS Access data retrieval on Windows Platform.

ii. This research supports the view that Unicode (UTF-8) is more preferred than Big5 in encoding Chinese or Cantonese texts with other multilingual data such as Dagaare and English.

iii. PHP works better with MySQL than MS Access in multilingual data transfer on the World Wide Web, as shown in the case of Dagaare, Cantonese and English e-lexicon.

iv. More of these Hong Kong Cantonese multilingual resources, which seem to be lacking on the World Wide Web, should be developed.

---

See Bodomo, Marfo, Cunningham & Mok (2004) for the comparison of Microsoft Keyboard Layout Creator and Tavultesoft Keyman Developer.
5 Conclusion

The new definitions of the Big5 and ISO-10646/Unicode character sets as a result of the establishment of the HKSCS have significantly contributed to the electronic transfer of Hong Kong Cantonese. As far as internationalization (i18N) and globalization of the web are concerned, more web-based Hong Kong Cantonese resource should be developed. It is hoped that this paper contributes to the existing research on computational lexicography and web internationalization, thus promoting further development on multilingual e-lexicons especially Cantonese lexical resources. This paper serves to address the need for more research on the development of web-based Hong Kong Cantonese lexical resources, in particular multilingual e-lexicons that would facilitate language transfer and enrich the lexical knowledge base. These valuable lexical resources are essential for language, social and cultural exchange among Chinese communities and the other parts of the world, hence benefiting global communications as a whole.

Acknowledgements

I would like to express my sincere gratitude to Dr. Adams Bodomo, my supervisor for his dedicated supervision and constructive advice throughout the entire process of writing this paper. As a native speaker of Dagaare, he also helped with the pronunciation recordings of Dagaare. I also wish to thank Mr. John Guest, a native English speaker and demonstrator in the Department, for recording the English pronunciations. Thanks to the Department of Linguistics, HKU for providing the server space for the trilingual lexical database.

References

Books & Journals:

Web Resources: