



Bringing home the global message

## The Final Frontier of the Non-standard Meeting the technical challenges of localizing Indic language Web sites

John O'Shea, VP Technical Strategy, Wordbank and Lindsay Johnson, Assistant Director, Technical Services. They can be reached at [john\\_oshea@wordbank.com](mailto:john_oshea@wordbank.com) and [lindsay\\_johnson@wordbank.com](mailto:lindsay_johnson@wordbank.com)

---

Having been in the localization industry for nearly two decades, the technology team at Wordbank has tackled just about every kind of font, character set and encoding issue that can manifest itself during the localization process. Ten years ago, working with Eastern European character sets represented a host of challenges. Five years ago, Asia-Pacific languages and Hebrew and Arabic stretched our technical ability to its limit. Or so we thought.

This year, safe in the assumption that we'd seen it all and could rely on our expertise of dealing with the peculiarities of character set encoding, we found ourselves wrestling with the localization of a Web site into the Indic languages of Hindi, Urdu and Gujarati. This was for an independent government department in the United Kingdom, which had already localized its site into French and Spanish but recognised both the need and the demand to communicate to both the UK's major ethnic minority communities and the global marketplace.

We embarked upon this task with full confidence in our ability to deliver using our standard working practice of delivering Web sites with UTF-8 encoding and the specification of standard fonts. Researching the issue, from a variety of sources, revealed no evidence to the contrary and so we began the project in the anticipation that we could rely on our usual, proven process.

In terms of Web site localization, "our usual, proven process" includes running the source files through our WordWeb tool (an in-house automatic text extraction tool that manipulates tagged file formats). Without going into the particulars of WordWeb and all its workings, the resultant file format distributed to suppliers is Microsoft Word. Word is generally known to support multilingual content (albeit in a mildly problematic manner) and considered by many of our freelance language suppliers to be the preferred environment in which to carry out the translation process, not least because of the vast support provided by linguistic tools, such as spell checkers, glossary and translation memory software.



**WORDBANK LIMITED**  
33 CHARLOTTE STREET  
LONDON W1T 1RR  
TEL: +44 (0) 20 7903 8800  
FAX: +44 (0) 20 7903 8888  
EMAIL: [word@wordbank.com](mailto:word@wordbank.com)  
**www.wordbank.com**

WORDBANK LIMITED REGISTERED AT THE OFFICE ADDRESS  
REG. NO. 2299752 VAT NO. GB 510 4082 02





Bringing home the global message

## These are fonts, but not as we know them

Having distributed the source content for localization in Word format to our in-country suppliers (leaving the HTML-specific content safe and sound in a separate file), we waited expectantly for the translated result. When the translations arrived, we faced the first of many hurdles to producing a localized deliverable - our old favourite, the non-standard font.

We found that several of our Gujarati and Hindi suppliers were using fonts such as Shusha, Vakil and Kruti that had non-UTF-8 encodings. Not altogether surprisingly, we found that these fonts were unable to display any characters apart from those for the language in question. This rendered them useless for the Web site we were localizing, which combined English with one Indic language in each localized version. Using an invaluable freeware tool, Unicode Font Inspector (for OS X), we were able to determine that the collation order of these non-standard fonts was, as expected, completely different from the Unicode collation order, rendering any programmatic alteration of the translated files a particularly ugly task.

Our natural inclination is always to automate the kind of tasks that would send any normal human being round the bend. The obvious solution seemed to be to create lookup tables based on the character map for the non-standard fonts used and their equivalents in the Unicode character set. We were unable, however, to find any electronic representations of the collation order for the non-standard fonts used. This meant that the only way to create a character map was to painstakingly identify each character in the non-standard font and to find the Unicode equivalent.

Given that the Indic languages we were working with each had around 100 characters, this wasn't a viable solution given the time constraints on the project. We were forced to have the work re-keyed in standard fonts (Mangal and Shruti) using either in-house resources or alternative third-party suppliers.

The whole process was very reminiscent of the situation we encountered on a regular basis six or seven years ago when dealing with Eastern European and Cyrillic typesetting projects, where our suppliers would regularly be working with fonts in a wide variety of encodings.



### WORDBANK LIMITED

33 CHARLOTTE STREET  
LONDON W1T 1RR  
TEL: +44 (0) 20 7903 8800  
FAX: +44 (0) 20 7903 8888  
EMAIL: [word@wordbank.com](mailto:word@wordbank.com)

[www.wordbank.com](http://www.wordbank.com)

WORDBANK LIMITED REGISTERED AT THE OFFICE ADDRESS  
REG. NO. 2299752 VAT NO. GB 510 4082 02

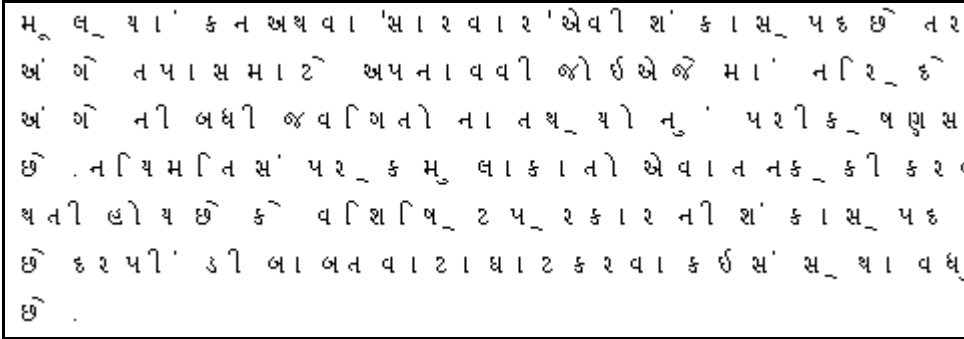


## Display difficulties

Our next surprise was the display of the translated documents in Word 2000 (running on Windows 2000). While working with in-country linguistic suppliers undoubtedly provides a better standard of translation, one of the challenges this represents is the arbitrary nature of their technical setups. Freelance suppliers run a variety of versions of Windows and MS Office and we found that this threw up difficulties and important considerations in the display of Indic languages.

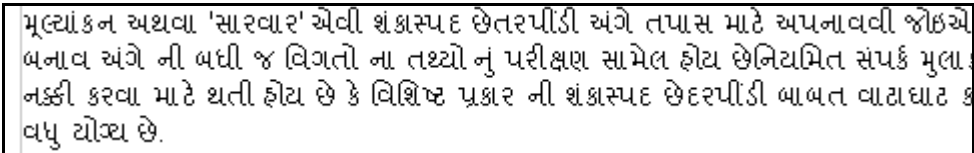
Urdu was not a major problem and did not display any particular issues in Word having been completed in Tahoma throughout the localization process. However, Hindi and Gujarati displayed what appeared to be spaces between each character. This seems to be a 'feature' of Word 2000, as the spaces were not actually present when saving and viewing the content in RTF. Curiously, both languages displayed correctly in Word 2002 and the open-source equivalent, OpenOffice.

### Screengrab of Gujarati text displayed in Word 2000:



મૂલ્યાંકન અથવા 'સારવાર' એવી શંકાસ્પદ છેતરપીડી અંગે તપાસ માટે અપનાવવી જોઈએ તપાસ માટે અપનાવવી જોઈએ અંગે ની બધી જ વિગતો ના તથ્યો નું પરીક્ષણ સામેલ હોય છેનિયમિત સંપર્ક મુલાકાત નક્કી કરવા માટે થતી હોય છે કે વિશિષ્ટ પ્રકાર ની શંકાસ્પદ છેતરપીડી બાબત વારાઘાટ કરવા વધુ યોગ્ય છે.

### Screengrab of the same document displayed in OpenOffice:



મૂલ્યાંકન અથવા 'સારવાર' એવી શંકાસ્પદ છેતરપીડી અંગે તપાસ માટે અપનાવવી જોઈએ તપાસ માટે અપનાવવી જોઈએ અંગે ની બધી જ વિગતો ના તથ્યો નું પરીક્ષણ સામેલ હોય છેનિયમિત સંપર્ક મુલાકાત નક્કી કરવા માટે થતી હોય છે કે વિશિષ્ટ પ્રકાર ની શંકાસ્પદ છેતરપીડી બાબત વારાઘાટ કરવા વધુ યોગ્ય છે.



#### WORDBANK LIMITED

33 CHARLOTTE STREET  
LONDON W1T 1RR  
TEL: +44 (0) 20 7903 8800  
FAX: +44 (0) 20 7903 8888  
EMAIL: word@wordbank.com

[www.wordbank.com](http://www.wordbank.com)

WORDBANK LIMITED REGISTERED AT THE OFFICE ADDRESS  
REG. NO. 2299752 VAT NO. GB 510 4082 02



**Screengrab of Hindi text displayed in Word 2000:**

यह मू ल्यां कन या जां च करने में  
सं भा वि त मा मले को जां च-पड ता ल  
या न हीं , भे जे गए मा मले के सा रे  
जां च की जा ती है । यह तय करने के  
के कि सी वि शे ष मा मले से नि ब ट ने  
अधि क उप यू क्त हैं नि य मि त सम्प र्क  
हैं ।

**Screengrab of the same document displayed in OpenOffice:**

यह मूल्यांकन या जांच करने में कि क्या धोखा-धड़ी के संभा  
पड़ताल के लिए स्वीकार किया जाए या नहीं, भेजे गए मामले  
जांच की जाती है। यह तय करने के लिए कि संभावित धोखाधड़

**WORDBANK LIMITED**33 CHARLOTTE STREET  
LONDON W1T 1RR

TEL: +44 (0) 20 7903 8800

FAX: +44 (0) 20 7903 8888

EMAIL: word@wordbank.com

**www.wordbank.com**

WORDBANK LIMITED REGISTERED AT THE OFFICE ADDRESS

REG. NO. 2299752 VAT NO. GB 510 4082 02



## Finding the best font solutions

Based on the technical challenges we encountered through working with our in-country suppliers, we recommended to our client that using the default, or readily available, Windows font choices (and their Macintosh equivalents) would be the sensible approach. Looking at the OpenType layout table support for various fonts, Arial Unicode MS (available on the MS Office and Publisher CDs) seemed to be a good fallback choice, as it supported Devanagari, Gujarati and Arabic. Our font choices, pre-browser testing, were as follows (in order of likelihood of occurrence):

### Wordbank Font Choices for Indic Languages

1. **Gujarati:** Shruti (Windows XP standard Gujarati font), Arial Unicode MS, Gujarati MT (OS X)
2. **Hindi (Devanagari):** Mangal (Windows XP standard Devanagari font), Arial Unicode MS, Devanagari MT (OS X)
3. **Urdu:** Tahoma (Windows 2000/XP standard Arabic font), Arial Unicode MS, Geeza Pro (OS X)

On re-engineering the files (by this we mean automatically merging the localized text with the source HTML content), there were display issues with the Urdu, which previously had caused no difficulty in the production process. The 'Farsi Yeh' character (Unicode character 0x6CC) refused to display correctly, regardless of the browser in question. Some further research revealed a not-especially-well-highlighted feature in that the character in question only appeared in particular versions of Tahoma, and our freelance proofreader was using an earlier Windows 98 version.

We determined that we required version 2.80 or 3.0 of Tahoma (which come as standard with Windows 2000 and XP, respectively), and that the only other standard Windows font capable of displaying all Arabic characters correctly is 'Microsoft Sans Serif'. Discussion with the client led to the decision to stay with Tahoma as the main font, as the target browser audience could be assumed to have relatively new machines, and thus have the later versions of the font.

The Hindi files proved less problematic in terms of display, though we noticed one display anomaly with Internet Explorer, which we are at a loss to explain. The following screenshots show the same file loaded into both Internet Explorer 6 and Firefox 0.93. The font settings are the same on both browsers and the screengrabs were taken on the same machine.



WORDBANK LIMITED  
33 CHARLOTTE STREET  
LONDON W1T 1RR  
TEL: +44 (0) 20 7903 8800  
FAX: +44 (0) 20 7903 8888  
EMAIL: [word@wordbank.com](mailto:word@wordbank.com)  
[www.wordbank.com](http://www.wordbank.com)

WORDBANK LIMITED REGISTERED AT THE OFFICE ADDRESS  
REG. NO. 2299752 VAT NO. GB 510 4082 02



Hindi document displayed in Internet Explorer 6 with Arial Unicode MS as default font:

मामला जांच-पड़ताल के लिए स्वीकार किया जाए या नहीं, यह मामले की मूल धोखा-धड़ी के संभावित और मेजे गए मामले के सारे ज्ञात तथ्यों की वस्तुतः है। यह तय करने के लिए कि संभावित धोखाधड़ी के किसी विशेष मामले से न सिर्फ सबसे अधिक उपयुक्त है नियमित सम्पर्क बैठकें आयोजित की जाती हैं। यह संभावित धोखाधड़ी के किसी विशेष मामले से निबटने के लिए कौन-सा संगठन नियमित सम्पर्क बैठकें आयोजित की जाती हैं।

Hindi document displayed in Firefox 0.9.3 with Arial Unicode MS as default font:

मामला जांच-पड़ताल के लिए स्वीकार किया जाए या नहीं, यह मामले की मूल्यांकन संभावित और मेजे गए मामले के सारे ज्ञात तथ्यों की विस्तृत जांच को बाद किया कि संभावित धोखाधड़ी के किसी विशेष मामले से निबटने के लिए कौन-सा संगठन नियमित सम्पर्क बैठकें आयोजित की जाती हैं। यह तय करने के लिए कि संभावित मामले से निबटने के लिए कौन-सा संगठन सबसे अधिक उपयुक्त है नियमित सम्पर्क बैठकें आयोजित की जाती हैं।

The circled element shows that Internet Explorer combines the characters in a different order to that of Firefox, which gets it right. Given that Firefox uses the same underlying font-rendering engine (the Windows Uniscribe technology), we can only presume this is a 'rendering feature' or bug in Internet Explorer.

## Grappling with Gujarati

Gujarati proved to be the most challenging of the three languages as it exposed some fairly fundamental display problems in all the browsers we tested, including Internet Explorer 5, IE6, Mozilla and Firefox.

We found that, using the 'Arial Unicode MS' font for Gujarati, Internet Explorer 6 and Firefox both display 'half characters' correctly but display consonants and their dependent vowel signs as two distinct characters, rather than combining them into one onscreen character. Internet Explorer 5.x does the reverse - it combines the characters correctly but does not display half-characters at all.



WORDBANK LIMITED  
33 CHARLOTTE STREET  
LONDON W1T 1RR  
TEL: +44 (0) 20 7903 8800  
FAX: +44 (0) 20 7903 8888  
EMAIL: [word@wordbank.com](mailto:word@wordbank.com)  
[www.wordbank.com](http://www.wordbank.com)

WORDBANK LIMITED REGISTERED AT THE OFFICE ADDRESS  
REG. NO. 2299752 VAT NO. GB 510 4082 02



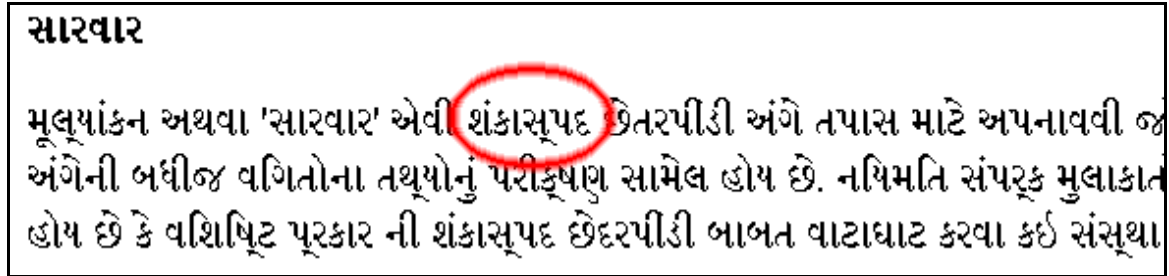
Most full / base consonants of the Gujarati alphabet also have half-forms which appear to the left of the full / base consonant, thus forming conjunct characters or consonant clusters. Their rendered forms often resemble the full form but are missing the vertical stem, which marks a syllabic core. Without the combination of the full and half consonants in a word, not only would the pronunciation of the word change, but it could also alter the meaning of a word.

In the worst-case scenario, the lack of a half character in a word could make it totally incomprehensible as it may not exist in Gujarati at all. It is similar to the occurrence of single and double consonants in the English language. For example, "bellow" with only one "l" would be "below", thus changing both pronunciation and meaning.

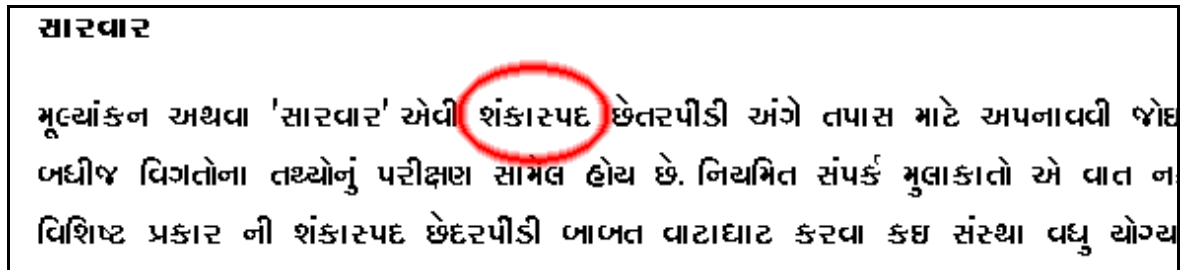
We thus had a situation where the same software and operating system will display the same text differently depending on the font selected, even though the fonts both appear to conform to the OpenType font standard.

Through exhaustive (and exhausting) experimentation, we discovered a freely available Gujarati font called Padmaa that actually displayed correctly in both Internet Explorer (all versions) and Firefox. Quite why it displays correctly is another matter as viewing the font in several font-inspection tools (on both Windows and Macintosh machines) revealed no significant differences at all. Both Padmaa and Arial Unicode MS encode the Gujarati character set in exactly the same way.

#### Gujarati document displayed in Internet Explorer 6 with Arial Unicode MS as default font:



#### Same document with Padmaa-Medium as default font:



WORDBANK LIMITED  
33 CHARLOTTE STREET  
LONDON W1T 1RR  
TEL: +44 (0) 20 7903 8800  
FAX: +44 (0) 20 7903 8888  
EMAIL: word@wordbank.com  
[www.wordbank.com](http://www.wordbank.com)

WORDBANK LIMITED REGISTERED AT THE OFFICE ADDRESS  
REG. NO. 2299752 VAT NO. GB 510 4082 02



However, that left us with the thorny issue of how our client's users would get hold of the font in the first place. Research into the relatively few non-English-language-only Gujarati sites revealed that many of them required their users to download and install the font (specified in English instructions), and that a significant proportion used fonts with non-Unicode encodings (cf. our supplier issues outlined earlier).

An option we considered was the use of the Internet Explorer-only ability to embed fonts in the pages in question. This would require use of Microsoft's WEFT tool to convert the TrueType font to EOT format and would imply a significant extra initial download upon first visit to the site (we estimated the converted font would be 30-90K in size).

However, a combination of the client's understandable dislike of Internet Explorer-only solutions and the discovery that the WEFT tool would not convert Padmaa anyway, forced us to consider a different solution.

### Reaching a compromise solution

In the end, we came to a compromise. This was to specify Padmaa first in the style sheet, followed by Shruti and Arial Unicode MS. The client also created a 'Display problems?' link on every page (in English), and created a page detailing (again, in English) the steps necessary to satisfactory font viewing. This approach, while not ideal, would mean that a visitor with standard fonts installed would be able to read the page to a reasonable extent, and, hopefully, have sufficient English knowledge to follow the 'Display problems?' link and the instructions on the resultant page.

### The dangers of editing UTF-8 encoded files

Having covered the anomalies in standard support for Unicode fonts for Indic languages and display issues in varying versions of Microsoft Office applications and browsers, we offer some advice on the dangers of editing UTF-8 encoded files in non-compatible text editors. This certainly isn't an issue limited to Indic languages but the mountain of problems in this project was exacerbated further by our client editing some of the delivered (UTF-8) files and then inadvertently saving them with Windows ANSI encoding. This, as expected, resulted in character corruption, which was detected and resolved during the final QA of the pages on the staging server. This required a small amount of re-education on the client's part and serves to emphasise the point that manipulating files with complex charactersets should always be approached with caution.

### The lessons learnt

This challenging project stretched the ingenuity, resourcefulness and technical skills of our technical team to the final frontier but we consider it to be an 'experience investment' because, although we have not undertaken many projects in these languages to date, we regard them as a major growth area for localization. As well as being significant



WORDBANK LIMITED  
33 CHARLOTTE STREET  
LONDON W1T 1RR  
TEL: +44 (0) 20 7903 8800  
FAX: +44 (0) 20 7903 8888  
EMAIL: [word@wordbank.com](mailto:word@wordbank.com)  
[www.wordbank.com](http://www.wordbank.com)

WORDBANK LIMITED REGISTERED AT THE OFFICE ADDRESS  
REG. NO. 2299752 VAT NO. GB 510 4082 02





Bringing home the global message

minority languages in the UK market, there are more than 180 million Hindi, 58 million Urdu and 44 million Gujarati speakers throughout the world. If the “global market” is to become truly global, working with these languages may well become the norm.

In the absence of standards, which will no doubt be forthcoming as demand grows, we can make the following recommendations for both localization specialists and clients alike:

- 1) Don't assume that freelance language suppliers are using UTF-8-compliant fonts. Whilst the general assumption is that this is the case, it is worth confirming this before attempting projects in new linguistic areas.

For the three Indic languages of Gujarati, Hindi and Urdu our final font choices (based on browser testing) were as follows:

**Gujarati:** Padmaa (Windows 2000/XP standard Gujarati font), Arial Unicode MS, Gujarati MT (OS X)

**Hindi:** Mangal (Windows XP standard Devanagari font), Arial Unicode MS, Devanagari MT (OS X)

**Urdu:** Tahoma (Windows 2000/XP standard Arabic font), Arial Unicode MS, Geeza Pro (OS X).

- 2) Consider the fact that character corruption issues may vary in the way they display from one application to another. What you think you see is not necessarily what is happening. When working with Word processing applications, it is always wise to test “corruption” in different versions of Microsoft Word. The open-source equivalent, OpenOffice, can also prove invaluable in this respect as OpenSource applications seem to provide more consistent and rapidly improving multilingual support. The same rule applies for browser applications, as different versions of Internet Explorer can produce different results as can the Mozilla-derived FireFox.
- 3) Even with rigorous testing, don't assume that users will have the font setup required to view your pages. Backing up your localization with an English description of “what to do if you can't view this” on your Web site, appears to be a standard approach for the few Indic sites we found and, with the current lack of support for world-wide standards, is an entirely advisable countermove.

Words: 2,489



WORDBANK LIMITED  
33 CHARLOTTE STREET  
LONDON W1T 1RR  
TEL: +44 (0) 20 7903 8800  
FAX: +44 (0) 20 7903 8888  
EMAIL: word@wordbank.com  
[www.wordbank.com](http://www.wordbank.com)

WORDBANK LIMITED REGISTERED AT THE OFFICE ADDRESS  
REG. NO. 2299752 VAT NO. GB 510 4082 02

