Typesetting in Hindi, Sanskrit and Persian: A Beginner's Perspective

Wagish Shukla
Maths Department
Indian Institute of Technology
New Delhi, India
wagishs@maths.iitd.ernet.in

Amitabh Trehan
Mahatma Gandhi Antarrashtriya Hindi Vishwavidyalaya (MGAHV)
16, 2nd floor, Siri Fort Road
New Delhi, India
amitabhtrehan@yahoo.co.in

Abstract

This paper describes our efforts to produce what is, to our knowledge, the first book typeset totally in an Indian language using LaTeX: *Chhand Chhand par Kumkum*, published by Prabhat Prakashan for Mahatma Gandhi Antarrashtriya Hindi Vishwavidyalaya (MGAHV).

We used the devnag package, which made it possible to encode each chapter, including verses, within a single set of `\dn` commands (much like an environment). Since then, we have also tried the sanskrit and ArabTeX packages and describe some of our experiences. Using devnag alone, typesetting a large file (a full-sized book) was a stable procedure. On the other hand, when using devnag and sanskrit together, even a small file can present problems. Using devnag/sanskrit in conjunction with ArabTeX is also problematic.

Additionally, one large part of the text was used to test conversion to HTML via latex2html (l2h) which has led to substantial upgrades of l2h by Ross Moore, its maintainer. This exemplifies the advantages of the free software community we have begun to live in. Ultimately, l2h was used to typeset MGAHV's website (http://www.hindivishwa.nic.in).

The Beginning

Our tryst with TeX began around the beginning of the year 2000 A.D. Since TeX/LaTeX is the best software for writing mathematical reports and we were in the mathematics department, we had come across mention of it here and there. Later we found that there were a few serious users, but most used GUI variants such as PCTeX (and not quite the latest ones!). The previous year the department and the institute had made rapid progress in computerisation and Internet connectivity, so every member of the faculty had a computer in his/her office and everybody (faculty and students) had round-the-clock Internet access. This prompted Wagish to think of what to do with the box in his office. He had previously stayed away from it religiously, but now he didn't want a relic in his room. So he decided to get 'computerised' and that's where Amitabh, who had recently started working with him as a student and welcomed the connectivity, came in handy. We picked up a lot of new ideas from the net, the airwaves and the brain waves and went about trying a few of them. Ultimately, we would have to say the most attractive ideas for us have been TeX, GNU/Linux and the free software philosophy.

Our first experiments using MikTeX, Ghostview, etc. to view mathematics papers were with Windows98 on a Pentium-II IBM machine (4GB HDD). Later, another computer (Pentium-III 500MHZ, 27GB HDD) and a laser printer were installed at the residence of Wagish Shukla and much of our work shifted there. We put up Redhat GNU/Linux and later Debian GNU/Linux on that machine. Meanwhile, TeXLive4.0, tugIndia, the tugIndia mailing list, CVR (C.V. Radhakrishnan) and like friends came along and we could do something useful.

The devnag Experience

Wagish writes in Hindi and needs to quote extensively from Sanskrit, Farsi and English, so it was natural that we should seek suitable solutions using LaTeX. Scanning the TeXLive4.0 package list, we came across the sanskrit, devnag and Indica packages. We couldn't find sanskrit and found no documentation for Indica. Fortunately, devnag was available, well documented and seemed friendly (important points for beginners). However, devnag on TeXLive4 was outdated (and still is, as of TeXLive6), making us suspect that we were in a less visited part of the forest. So, we downloaded devnag (v2.0, which had been upgraded to LaTeX2ε) from CTAN and set about experimenting with it. From the outset, the idea was to be able to produce large texts in Devanagari from it. As we progressed, it seemed that the developers' idea must have been to use it for short passages of Devanagari texts within English text but we are happy to state that we have been able to use it to typeset a whole book.

[tuglist] devnag + Windvi = Crash While using devnag with the TeXLive system with the Windows O.S., we came across a very strange problem. The devnag example and the test files compiled fine, so we made a small file with just some Devanagari text. This compiled and previewed well. Then we added some size-changing commands to it. It compiled. But as soon as we tried to preview it using Windvi (v. 0.66-pre6), Windows either went into a spate of blue-screen exception fault errors and rebooted or just rebooted without any warning. We copied the same file onto GNU/Linux and after removing the Microsoft newlines, we had no problem with the file. This was very intriguing. This happened to any devnag file which used size-changing commands (`\small`, `\large`, etc.)! So we posted the message on the list with the subject that takes the name of this subsection. Judging from the responses, hardly anybody on the list was using Windows (or if they did, they didn't respond). The problem indeed sounded strange to whoever heard it. Nobody could suggest what was wrong. Later, we also had some problems printing English files with Windvi. In a bit of hurry, we turned our attention to GNU/Linux and moved on.

In one of the discussions on the mailing list, C.V. Radhakrishnan had written: "Franz Velthius' simple preprocessor can seldom blow up a Win32 system". This leads us to suspect that the problems may have been caused by a virus or an anti-virus (we had Norton AntiVirus 2000 by then). Recently, when we tried to repeat the experiment with the same O.S. on the same machine, with TeXLive5, Windvi 0.67 and Norton Antivirus 2002, we had no such problems.

The Book Various experiments and Devanagari articles later, we came to do something really exciting. Wagish is a creator of many unfinished symphonies. Regarding TeX, Donald Knuth has written that it inspired him to write more and even rewrite his previous works because he could see his work beautifully written. Similarly, the transformation of his ideas typeset into a beautiful form have spurred Wagish to write more. The story of the book *Chhand Chhand par KumKum* had begun long ago, but somehow the book never materialised. Enthused by the idea of writing in Devanagari in a beautiful manner using the ethically beautiful idea of free software, Wagish thought that if it could be demonstrated that the author's creativity could be simply and beautifully expressed using the TeX system, it would inspire many people in many ways.

Chhand Chhand par KumKum is actually a commentary by Wagish of the famous poem "Ram Ki Shakti Puja" by Suryakant Tripathi Nirala, a very important poem in Hindi literature and considered rather difficult to discuss. Wagish wrote the criticism for one part of it (around a third), which was published in an issue of MGAHV's Hindi language literary magazine *Bahuvachan*. Though the rest of the issue was in a separate font using a different system, this article was printed using LaTeX. Thus, this issue has two distinct parts derived from two distinct systems. The look of the devnag font met with general appreciation and we ourselves were impressed with the intuitive commands and immense power that LaTeX and devnag offered. After this, the next logical step was to write the entire book using LaTeX and devnag.

Once this idea was concretised with support from MGAHV and its Vice Chancellor Ashok Vajpeyi and the arrangements worked out, we set to work. The whole contents of the book were then recreated and typed online by Wagish in almost exactly a month. The section previously published was also totally revised. For the general layout of the book, we used fancyheadings for the headers and footers and layout for testing the layout. Of course, our constant companions were the LaTeX book [1] and the *LaTeX Companion* book [2]. Our book was then put into final shape with help from other members of MGAHV and LILA (MGAHVś Laboratory for Informatics in the Liberal Arts), along with the publishers. Actually, in this area, publishers here still look at our LaTeX experiment more as an idle curiosity than anything really useful.

While working with devnag we came across some interesting situations, described in the next section.

Critique Working with the devnag package on GNU/ Linux has been a pleasant experience. Bedore are some of our observations:

- In one of our first long articles, we just input the source file as a single paragraph without any line breaks. This is, of course, not a good practice, as it takes away from the readability of the text. When we used the devnag preprocessor, we were greeted by a segmentation fault. This was undoubtedly due to the limit of the text read into the character array in the preprocessor.
- The most useful feature is the transliteration scheme used by Frans Velthius. The whole text is typed in English and then converted by the preprocessor to a form suitable for LaTeX to generate the final output. Since this is a phonetic-based scheme, it is easy to remember. Moreover, the ligature construction is very close to the actual phonetic construction.
- The most attractive feature in devnag, which also highlights the advantage of a Character User Interface (CUI) approach versus a Graphical User Interface (GUI) approach, is the ligature construction. devnag has a wide range of ligatures. There is also the choice of switching individual ligatures on and off, as well as a broad subdivision of Hindi and Sanskrit ligatures.
- Just after a new line (\\), if a word begins with "qa", the "qa" is not processed. Thus

  {\dn
  namaskaara\\ qaafa
  }

  yields

  नमस्कार
  ाफ़
- The preprocessor does not always handle the verbatim environment properly (although it is supposed to). Thus, the segment in the item above with verbatim would be written as:

  ```
  {\dn
  nm-kAr\\  *A'
  }
  ```

  since it has preprocessed the contents.
- We wanted to write the word जुर्अत 'jurat', which reads normally as जुरत. By trial and error we discovered the way to input this was `jua\0ta`.
- For underlining a Devanagari passage, it is better to use the ulem package rather than the usual `\underline` command.
- Additional symbols were generated by using diacritics, as in a forthcoming book on Ghalib being written by Wagish; characters have been generated by using TIPA, which works well with devnag. For example, there are five letters in the Persian/Urdu alphabet which are, in India, homophonically pronounced as 'za'/ज़, but although devnag supplies 'za'/ज़, the five different versions were reproduced as follows:

  1. `za`/ज़ for Arabic ZE.
  2. `\textsubbar{za}`/ज़̱ for Persian/ Urdu ZAAL.
  3. `\textsubdot{za}`/ज़̣ for Persian/ Urdu ZVAD.
  4. `\textsubumlaut{za}`/ज़̤ for Persian/ Urdu ZOE.
  5. `\sout{za}`/~~ज़~~ for Persian ZE.

  The first four are from TIPA, the fifth from ulem. Similarly, in Persian/Urdu ख़्वाब, the व is not pronounced but written; thus, the pronounciation is ख़ाब but one must write ख़्वाब — the devnag input for ख़ाब is `.khaaba` and that for ख़्वाब is `.khvaaba` but it was impossible to indicate the same pronounciation with two differently spelled words. Instead, this was achieved by ख़्वा̫ब (`\textsubw{.khvaa}ba`), using a command from TIPA.
- The compability of many LaTeX packages such as TIPA with devnag is heartening. However, ArabTeX does not mix well and loading sanskrit with either ArabTeX or devnag creates problems. Ideally, one would like to load all three (ArabTeX, sanskrit, devnag) at the same time.

LaTeX2HTML and devnag

MGAHV, a new university dedicated to Indian languages, literature, etc. needed to establish a website. Due to the profile of the university, it was necessary to have a bilingual website. We analysed the available options and found that there really wasn't any standard solution for setting up a website in Devanagari. One important criteria for us was that our site should be accessible uniformly across platforms and browsers: that is, setting up the site with some specific font made available for download was not an attractive option. Most sites that use this solution can only be accessed on the Windows platform after installing the proper font. Needless to say, in this age of viruses and worms, one is rather hesitant to install something to view a site. There is the option of using dynamic fonts but we were not sure about reliability, the degree of complexity of such a solution and whether there was anything in the free software domain for this. So, it seemed that we needed some image-based solution for our limited needs, but one which would not bloat up the size of the files, so that access remained reasonably fast. Given our devnag experience, we hoped to find something similar in nature. And we did — LaTeX2HTML (l2h), which also provided support for devnag.

Development via the net It was a bit of a bumpy ride getting l2h working for devnag: it turned out that nobody, to our knowledge, had used it before. Thus, like Wagish's book, MGAHV's site is also the first one created via this route. We attempted to run l2h on our devnag files and constantly mailed queries to the current maintainer, Ross Moore, who kept on advising and correcting bugs till, at last, l2h ran pretty well with devnag. This was, for us, a unique experience of software development via the Internet in the free software domain and highlighted the advantages and the cooperative spirit that this approach can generate.

l2h generates PNG/GIF images for things not directly available via HTML, such as mathematics and Indian language characters. This is where things get complicated, as l2h depends on the support of a number of other applications for image generation, including the netpbm suite of files. We installed l2h from source and then tried the package made by Manoj Srivastava for Debian on our Debian system, but the images wouldn't generate. So we joined the mailing list and realised that we needed to update netpbm. Once upgraded, the `"make test"` with l2h worked and everything seemed to be ready. But when we tested it with a small sample file it wouldn't work: it couldn't locate the devnag style files and generate images, even though it would work on Ross's system. We had also copied the l2h Indic-TEX devnagri.sty and devnagri.perl files to particular locations, as indicated in the l2h documentation. That's when Ross realised that the files for the upgraded devnag had not been uploaded for distribution. So he took care of that. By default the system had been set to use the DN2 preprocessor with devnag (DN2 is used with texts in German). Ross changed the default and left DN2 as an option.

Since we had now made some progress, we decided to give it a more thorough test. We fed l2h Wagish's article, "Ram Ki Shakti Puja", mentioned in the previous section — a file of 89Kb. l2h invokes LATEX to generate images, but it complained of memory shortage and halted. Moreover, the log indicated that l2h was trying to create just three images from the whole document. The cause of this problem turned out to be very interesting.

The article actually had a very typical structure which may not, however, have been envisioned by the developers. There were many verse environments within a single set of `\dn` braces whereas the developers had probably expected a set of `\dn` braces for each verse, so l2h was trying to generate huge images and collapsed. Ross improved the paragraph breaking, also adding an option for newlines within the title command and ultimately put up the converted document on his site. And so Lord Rama now adorns the net as a test case.

Satisfied with the results, we carried the experiment forward and created the LILA website (`www.hindivishwa.nic.in`). The images are set against a white background and the web document looks good. Overall, feedback about the quality and speed of access has been positive from people who have visited the site. The ultimate solution is probably going to come with the use of Unicode and like encodings, but we think that, with some more facilities, l2h would make a good substitute in the meanwhile.

Critique

- l2h has proven to be a good solution for sites with static Devanagari content. PNG images are of a reasonable size and don't slow down the site too much.
- At times, there are problems with clipping of the boxes around images.
- We need to have an easier update system (a sort of version control and patch system) for updating image-based sites. This is because it takes longer to process the whole text, even if one just wants to add, say, a page to the original. It would also be much easier to just upload/delete a few images instead of the whole site, which may be required for changes at the present. Thus, such a package could provide content additions, deletions and updating facilities.
- There is probably a need for closer collaboration between the developers of l2h and say, netpbm, to maintain compatibility.

devnag, sanskrit and ArabTEX

In the book on Ghalib (in Hindi) that Wagish is presently working on, we needed to use Arabic. So, we tried to use ArabTEX with sanskrit and devnag, in various permutations. The results were not very good:

- sanskrit(skt) and devnag, when taken together, make the typeset words look weird.
- ArabTEX and sanskrit or devnag output certain Greek letters at the beginning of a document and don't process the text properly.

While devnag has been useful to us, there are facilities in other packages which could be useful if incorporated into devnag:

- ArabTEX doesn't use a preprocessor
- sanskrit has support for vedic Sanskrit (but not Hindi)
- both ArabTEX and sanskrit incorporate standard transliteration facilities
- sanskrit has both bold and italic fonts

Conclusions

There is a need for more language-specific development on TEX systems, if publishers in Indian languages are to be convinced to start using TEX. Some improvements which could be made do not seem extremely difficult for the developers. There is also a need for greater variety in the form of fonts, etc. Native speakers of the language should get involved in at least the testing of suitable packages, as they could provide some unique insights.

References

[1] Leslie Lamport. *LATEX: A Document Preparation System.* Addison-Wesley, Reading, Massachusetts, second edition, 1994.

[2] Michel Goosens, Frank Mittelbach and Alexander Samari. *The LATEX Companion.* Addison-Wesley, Reading, Massachusetts,1994.

[3] Wagish Shukla. *Chhand Chhand par Kumkum.* Prabhat Prakashan, New Delhi, 2001.