

BABEL SPEAKS

हिंदी

Zdeněk Wagner

<http://icebearsoft.euweb.cz>

Basic facts

- Hindi, together with English, is an official language of the Republic of India.
- Hindi is the first language in 6 Indian states (Uttar Pradesh, Bihar, Madhya Pradesh, Rajasthan, Haryana, Himachal Pradesh)
- 180 million people use Hindi or one of its dialects as their mother tongue
- 418 million people speak Hindi all over the world
- Economic boom expected in China and India

Current situation

- Typesetting in most Indic scripts supported
- Some scripts requires preprocessing (preprocessors are available)
- Language switching not supported
- Package `devanagari.sty` defines “mini-babel”

Note

Babel support for Kannada is being developed by the Kannada \TeX group — see <http://www.sarovar.org/projects/kannadatex>

Enabling Hindi in Babel

- No hyphenation patterns are available (words are not hyphenated), therefore `zerohyph.tex` is used.
- A few macros must be defined, see samples below.

<code>\partname</code>	खण्ड
<code>\chaptername</code>	अध्याय
<code>\contentsname</code>	विषय - सूची
<code>\indexname</code>	सूची
<code>\tablename</code>	तालिका
<code>\figurename</code>	चित्र
<code>\pagename</code>	पृष्ठ

`\today: ५ नवम्बर २००६`

Why not in hindi.ldf

- The definitions must be preprocessed independently (once forever).
- They already exist in dev(anagari).sty since 2.13 and are documented.
- It would be error-prone to manage the definitions at two places.
- For typesetting Hindi the preprocessor, fonts and macros from the Velthuis Devanāgarī for T_EX are necessary.

Solution

- Some potentially conflicting macros defined via `\providecommand` during `\AtBeginDocument`.
- Package `devanagari.sty` loaded by `\RequirePackage` with version check.
- Options declared as language attributes.

Advantage of Babel

- Unified language environment
- More flexible redefinition of captioning macros

Put the following into the preamble:

```
\addto\captionmodernhindi{\def\indexname{{\dn anukrama.nikaa}}}
```

Index will be titled as अनुक्रमणिका instead of the default सूची.

Velthuis encoding versus Unicode

- Velthuis: Both traditional शक्ति and modern शक्ति encoded as "sakti"
 - i-matra (ॆ) moved in front of the consonant by the preprocessor
 - conjuncts (क + त) denoted by missing a between consonants
 - form of conjuncts switched by preprocessor commands.
 - independent and dependent vowels, e. g. long a in आसान (aasaan) handled by the preprocessor
- Unicode: encoded as letters sha + ka + virama + ta + i-matra
 - i-matra (ॆ) moved in front of the consonant by the rendering engine
 - a-matra not encoded
 - conjuncts denoted by virama between consonants
 - form of conjuncts depends on availability in the font
 - half form (क्त) may be forced by zero width joiner
 - independent and dependent vowels (matras) have different codes

Searchable PDF

- Font encoding for Devanāgarī named as X0900 (for cmap.sty)
- ToUnicode map defined
- Each half-form consonant encoded as the respective consonant followed by *vīrama*
- Subscript repha encoded as *ra* preceded by *vīrama*
- Superscript repha encoded as *ra* followed by *vīrama*
- Ligatures encoded as corresponding sequences of Unicode characters

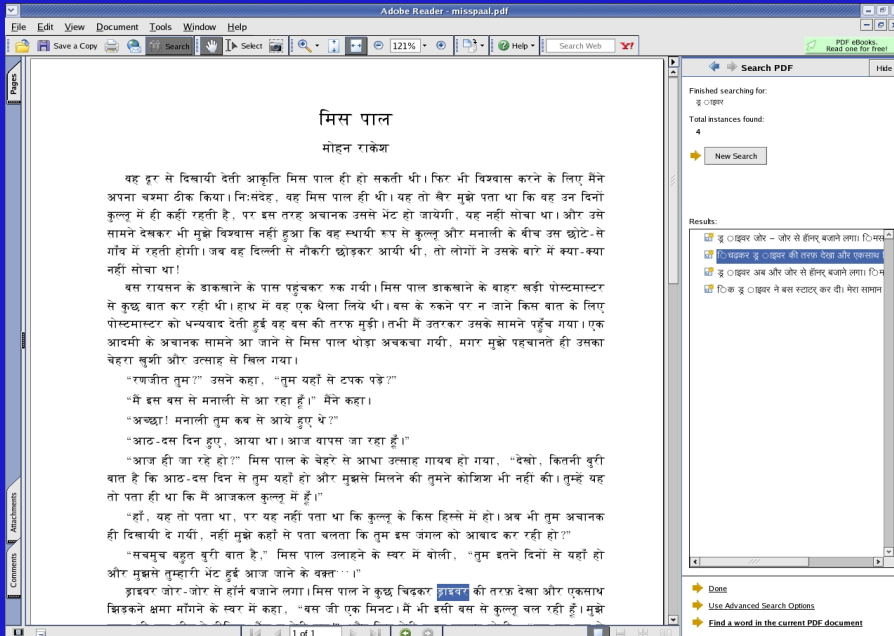
Not handled

- Order of i-matras and superscript rephas is wrong
- Extra spaces generated by Acrobat Reader after subscript matras and subscript repha
- Long independent *a* (आ) split into two Unicode characters: अ + ा

Mappings of type **many-to-many** would be needed to solve this problem.

This causes problems not only when searching but also when cutting & pasting the text from Acrobat Reader to other applications.

Searching **इंटरनेट** in Linux



The screenshot shows the Adobe Reader interface with a PDF document open. The document text is in Hindi and discusses a character named 'मिस पाल' (Miss Pal) and her search for a husband. A search overlay is visible on the right side of the document, showing search results for the term 'इंटरनेट' (Internet).

मिस पाल
मोहन राकेश

यह दूर से दिखायी देनी आकृति मिस पाल ही हो सकती थी। फिर भी विश्वास करने के लिए मैंने अपना चश्मा ठीक किया। निःसंदेह, वह मिस पाल ही थी। यह तो सैर मुझे पता था कि यह उन दिनों कुल्लू में ही कहीं रहती है, पर इस तरह अचानक उसमें भेंट हो जायेगी, यह नहीं सोचा था। और उसे सामने देखकर भी मुझे विश्वास नहीं हुआ कि वह स्थायी रूप से कुल्लू और मनाली के बीच उस छोटे-से गाँव में रहती होगी। जब वह दिल्ली से नौकरी छोड़कर आयी थी, तो लोगों ने उसके बारे में क्या-क्या नहीं सोचा था!

बस रायसन के डाकखाने के पास पहुँचकर रुक गयी। मिस पाल डाकखाने के बाहर बड़ी पोस्टमार्टर से कुछ बात कर रही थी। हाथ में वह एक पैला लिये थी। बस के रुकने पर न जाने किस बात के लिए पोस्टमार्टर को धन्यवाद देती हुई वह बस की तरफ मुड़ी। तभी मैं उतरकर उसके सामने पहुँच गया। एक आदमी के अचानक सामने आ जाने से मिस पाल थोड़ा अचकचा गयी, मगर मुझे पहचानते ही उसका चेहरा खुशी और उत्साह से खिल गया।

“रणजीत तुम ?” उसने कहा, “तुम यहाँ से टपक पड़े ?”

“मैं इस बस से मनाली से आ रहा हूँ।” मैंने कहा।

“अच्छा! मनाली तुम कब से आये हुए थे ?”

“आठ-दस दिन हुए, आया था। आज वापस जा रहा हूँ।”

“आज ही जा रहे हो ?” मिस पाल के चेहरे से आधा उत्साह गायब हो गया, “देखो, कितनी बुरी बात है कि आठ-दस दिन से तुम यहाँ हो और मुझसे मिलने की तुमने कोशिश भी नहीं की। तुम्हें यह तो पता ही था कि मैं आजकल कुल्लू में हूँ।”

“हाँ, यह तो पता था, पर यह नहीं पता था कि कुल्लू के किस हिस्से में हो। अब भी तुम अचानक ही दिखायी दे गयी, नहीं मुझे कहाँ से पता चलता कि तुम इस जंगल को आबाद कर रही हो ?”

“सचमुच बहुत बुरी बात है,” मिस पाल उलाहने के स्वर में बोली, “तुम इतने दिनों से यहाँ हो और मुझसे तुम्हारी भेंट हुई आज जाने के वक़्त...”

इंटरनेट जोर-जोर से हॉर्न बजाने लगा। मिस पाल ने कुछ चिढ़कर इंटरनेट की तरफ देखा और एकसाथ झिड़कने क्षमा माँगने के स्वर में कहा, “बस जी एक मिनट। मैं भी इसी बस से कुल्लू चल रही हूँ। मुझे

Search PDF

Finished searching for:
इंटरनेट

Total instances found:
4

New Search

Results:

- इंटरनेट जोर - जोर से हॉर्न बजाने लगा। इंटरनेट
- इंटरनेट इंटरनेट की तरफ देखा और एकसाथ
- इंटरनेट जोर और जोर से हॉर्न बजाने लगा। इंटरनेट
- इंटरनेट जोर ने बस स्टॉप पर ही पैर सामने

Done
Use Advanced Search Options
Find a word in the current PDF document

1 of 1

Searching डाइवर in Windows XP

The screenshot shows the Adobe Reader interface with a PDF document titled 'misspaal.pdf' open. The document content is in Hindi and discusses a character named Miss Paal. A search bar at the top right contains the text 'डाइवर'. The search results panel on the right shows three matches for the word 'डाइवर' in the text.

Adobe Reader - [misspaal.pdf]

File Edit View Document Tools Window Help

Save a Copy Search Select 115% Help Search Web

miss paal
मोहन राकेश

वह डूब से विश्वासी देती आकृति मिस पाल ही हो सकती थी। फिर भी विश्वास करने के लिए मैंने अपना बख्सा ठीक किया। निःसंदेह, वह मिस पाल ही थी। यह तो बैर मुझे पता था कि वह उन दिनों कुल्लू में ही कहीं रहती है, पर इस तरह अचानक उससे भेंट हो जायेगी, यह नहीं सोचा था। और उसे सामने देखकर भी मुझे विश्वास नहीं हुआ कि वह स्थायी रूप से कुल्लू और मनाली के बीच उस छोटे-से गाँव में रहती होगी। जब वह दिल्ली से नौकरी छोड़कर आयी थी, तो लोगों ने उसके बारे में क्या-क्या नहीं सोचा था!

बस रायसन के डाकखाने के पास पहुँचकर रुक गयी। मिस पाल डाकखाने के बाहर खड़ी पोस्टमास्टर से कुछ बात कर रही थी। हाथ में वह एक थैला लिये थी। बस के रुकने पर न जाने किस बात के लिए पोस्टमास्टर को धन्यवाद देती हुई वह बस की तरफ मुड़ी। तभी मैं उतरकर उसके सामने पहुँच गया। एक आदमी के अचानक सामने आ जाने से मिस पाल थोड़ा अचकचा गयी, मगर मुझे पहचानते ही उसका चेहरा सुजी और उसाह से खिल गया।

“रजनीत तूम?” उसने कहा, “तूम यहाँ से टपक पड़े?”

“मैं इस बस से मनाली से आ रहा हूँ।” मैंने कहा।

“अच्छा! मनाली तूम कब से आये हुए थे?”

“आठ-दस दिन हुए, आया था। आज वापस जा रहा हूँ।”

“आज ही जा रहे हो?” मिस पाल के चेहरे से आधा उसाह गायब हो गया, “देखो, कितनी बुरी बात है कि आठ-दस दिन से तूम यहाँ ही और मुझसे मिलने की तूमने कोजिज भी नहीं की। तुम्हें यह तो पता ही था कि मैं आजकल कुल्लू में हूँ।”

“हाँ, यह तो पता था, पर यह नहीं पता था कि कुल्लू के किस हिस्से में हो। अब भी तूम अचानक ही विश्वासी दे गयी, नहीं मुझे कहाँ से पता चलता कि तूम इस जंगल को आबाद कर रही हो?”

“सचमुच बहुत बुरी बात है,” मिस पाल उलाहने के स्वर में बोली, “तूम इतने दिनों से यहाँ हो और मुझसे तुम्हारी भेंट हुई आज जाने के वक्त...”

डाइवर जोर-जोर से हँसि बजाने लगा। मिस पाल ने कुछ चिदकर **डाइवर** की तरफ देखा और एकसाथ झिड़कने क्षमा माँगने के स्वर में कहा, “बस जी एक मिनट। मैं भी इतनी बस से कुल्लू चाल रही हूँ। मुझे कुल्लू की एक सीट दे दीजिए। थैक यू बेरी मच!” और फिर मेरी तरफ मुड़कर बोली, “तूम इस बस से उतरने का मतलब क्या है?”

Search PDF

Finished searching for:
\\u0921\\u094d\\u0930
\\u093e\\u0907\\u0935\\u0930

Total instances found:
4

New Search

Results:

- डाइवर जोर - जोर से हँसने बजाने लगा।
- डाइवर के जोर-जोर से हँसने बजाने से
- डाइवर जोर-जोर से हँसने बजाने से
- डाइवर ने बस उतरा कर दी। बस च

Done
Use Advanced Search Options
Find a word in the current PDF document

Problems in other Indic scripts – two-part matras

Meaning	Devanāgarī	Malayālam
	देवनागरी	മലയാലം
ma	म	മ
maa	मा	മാ
mi	मि	മി
mii	मी	മീ
me	मे	മേ
mo	मो	മോ

Requirements for Xe_{La}TeX

- ToUnicode must be implemented, otherwise PDF will hardly be searchable
- hindi.ldf cannot be used as such, definitions from devanagari.sty have to be used
- handling ligatures: क्त vs. क्त (difference probably only in अक्तूबर = October)
- Zero width joiner *must not* produce visible mark in PDF

Requirements for other applications

- Use of ICU, Pango or similar rendering engine
 - the source text must be written in some editor, support in *vim* is missing
 - text has to be entered into the search dialogue
 - support for cut & paste between applications
 - correct order of matras for sorting (MakeIndex for Devanāgarī not yet available)
- Correct handling of superscript repha:
 - वर्ष is sometimes displayed as वर्ष
 - in addition, characters may be overlapped, i. e. दर्जन is displayed as दर्ज

Future plans for Velthuis Devanāgarī for T_EX

- Reimplementation of the preprocessor in LUA
 - everything will be done in a single step while T_EXing
 - X₃T_EX could process old documents if LUA were implemented
- Conversion of Velthuis fonts to OpenType

Availability

- Velthuis Devanāgarī:
 - <http://devnag.sarovar.org> (CVS repository + releases)
 - CTAN
 - T_EXLive
- My tools: <http://icebearsoft.euweb.cz>

Acknowledgment

- Other developers of Velthuis Devanāgarī for T_EX, especially Anshuman Pandey for translating the captions into Hindi
- John Smith and Arnošt Štědrý for providing test files created by X₃T_EX
- Alexandr Babič for running the test under Ubuntu
- Petr Tomášek for explanation of topics related to font rendering in X
- \mathcal{C} S_{TUG} and TUG for financial support

الْحَطُّ يَبْقَى مَا نَابَعْدَكَ كَاتِبُهُ
وَكَانَ الْحَطُّ تَحْتَ الْأَرْضِ مِنْ فَوْقِ

From the textbook *Základy moderní spisovné arabštiny*
by Jiří Fleissig and Charif Bahbouh