HindiTech
A Blind Double Peer Reviewed Bilingual Web-Research Journal

भाषा-प्रौद्योगिकी में यूनिकोड : भारतीय भाषाओं के संदर्भ में

ऋचा

rsrishti@gmail.com

भारतीय भाषा संस्थान, मैसूरु में सेवा। संप्रति क्रिस्ट विश्वविद्यालय में एसोसिएट प्रोफेसर

सत्येन्द्र कुमार अवस्थी

awasthisatendra@gmail.com

लखनऊ विश्वविद्यालय से पी-एच.डी. के बाद भारतीय भाषा संस्थान, मैसूरु में सेवारत

---

1. भूमिका

आधुनिक युग में भारत को तकनीक के क्षेत्र में महारत प्राप्त होती जा रही है। ज्ञान-विज्ञान के क्षेत्र में यह निरन्तर प्रगति कर रहा है। सूचना-प्रौद्योगिकी के क्षेत्र में भी यह विकास के पथ पर अग्रसर है। इस पथ को और अधिक सुदृढ़ एवं शक्तिशाली बनाने के लिए भाषा प्रौद्योगिकी को भी विकसित करना होगा। भाषा प्रौद्योगिकी के विकास के लिये सबसे पहला कदम है- भाषा को अत्याधुनिक बनाना। तत्कालीन समय में 'अत्याधुनिक' से तात्पर्य है भाषा का वेब-आधारित एवं प्रौद्योगिकी उपयोग। यूरोपीय तथा अन्य भाषाएँ इस विषय में अग्रणी हैं, किंतु भारतीय भाषाएँ अभी भी बहुत पिछड़ी हैं। इसके लिए भारतीय भाषाओं में

विषय वस्तु का निर्माण तथा इसका वेब रूपांतरण अनिवार्य है। यह रूपांतरण यूनिकोड में करना अत्यावश्यक है, क्योंकि इस कोड में होने से किसी भी विषयवस्तु का वेब-आधारित एवं भाषा-प्रौद्योगिकी उपयोग किया जा सकता है।

2. यूनिकोड

2.1.यूनिकोड है क्या?

यूनिकोड एक प्रकार का कैरेक्टर इन्कोडिंग मानक है, जिसे सूचना प्रौद्योगिकी के अन्तर्गत महत्त्वपूर्ण स्थान प्राप्त है। यूनिकोड का अर्थ है सार्वभौमिक कोड (Universal Code)। इसका प्रयोग विश्व स्तर पर सूचना के आदान-प्रदान के मानक के रूप में स्वीकार किया जा रहा है। यूनिकोड के विकास से पूर्व भारत सरकार ने एक मानक निर्मित किया, जिसे ISCII (Indian Script Code for Information Interchange) के नाम से जाना जाता है। इस मानक में विभिन्न लिपियों में प्रयुक्त समान कैरेक्टरों के लिए समान कोड-प्वाइंट निर्धारित किया जाता है। उदाहरणस्वरूप, देवनागरी 'क' (ka) और गुजराती 'ક' (ka) के लिए समान कोड-प्वाइंट है। इसके विपरीत, यूनिकोड में प्रत्येक कैरेक्टर को एक विलक्षण कोड-प्वाइंट प्रदान किया जाता है जिसके कारण दो भाषाओं को परस्पर प्रतिचित्रित करना (mapping) अत्यन्त सरल होता है।

2.2.यूनिकोड लिपि

यूनिकोड में लिपि का अर्थ है 'अक्षरों व अन्य लिखित संकेतों का समूह' जो एक या एक से अधिक लेखन प्रणालियों (writing systems) में पाठीय सूचना का प्रतिनिधित्व करता है। इसके अतिरिक्त यूनिकोड में कुछ प्रतीक (symbols) भी होते हैं। प्रत्येक लिपि में कुछ लिपि-विशेष विराम चिह्न (punctuation), संख्या सूचक चिह्न (numerals) तथा विशेषक चिह्न (diacritics) भी होते हैं। जब विभिन्न भाषाओं में समान लिपि का प्रयोग किया जाता है, तो प्रमुख रूप से विशेषक चिह्नों (diacritics) तथा अन्य चिह्नों में कुछ असमानताएँ दृष्टिगत होती हैं। उदाहरणस्वरूप, हिंदी और मराठी दोनों भाषाएँ देवनागरी लिपि का प्रयोग करती हैं। किंतु मराठी में 'ळ' का प्रयोग होता है जबकि हिंदी में यह विद्यमान नहीं है। अतः यह कह सकते हैं कि समान लिपि का प्रयोग करने वाली भाषाओं में कुछ विषमताएँ भी होती हैं। जैसे हिंदी,

संस्कृत, कोंकणी, मराठी, नेपाली इत्यादि भाषाएँ परिधीय असमानताओं के बावजूद एक ही लिपि देवनागरी का प्रयोग करती हैं। इसलिए यूनिकोड एक प्रकार की आधारभूत तकनीकी योजना है जो लिपियों का अमूर्तीकरण करता है। यूनिकोड की लचीली लिपियाँ, संयोजन चिह्न तथा परितुलन कलन विधियों (collation algorithms) के द्वारा विभिन्न लेखन प्रणालियों में विद्यमान विषमताओं को भी समर्थन देती हैं। इसके अतिरिक्त कुछ कैरेक्टरों जैसे मुद्रा (currency) संकेतों, प्रतीकों, संख्या सूचक तथा विराम चिह्नों का प्रयोग एक से अधिक लेखन प्रणालियों में होता है। इन परिस्थितियों में यूनिकोड उन्हें एक आम लिपि (common script) ISO 15924 कोड "Zyyy" के अंतर्गत परिभाषित करता है। यूनिकोड में कुल मिलाकर इस प्रकार के कैरेक्टरों की संख्या 6379 है।

2.3.यूनिकोड इन्कोडिंग

यूनिकोड 16-बिट इन्कोडिंग का प्रयोग करता है जो 65,000 कैरेक्टरों से भी अधिक (65536) के लिए कोड-प्वाइंट उपलब्ध कराते हैं। यूनिकोड दो प्रतिचित्रण (mapping) विधियों को परिभाषित करता है- यूनिकोड परिवर्तन स्वरूप UTF (Unicode Transformation Format) इन्कोडिंग, और यूनिवर्सल कैरेक्टर सेट UCS (Universal Character Set) इन्कोडिंग। यूनिकोड मानक एक विस्तार यंत्रावली उपलब्ध कराते हैं जो एक मिलियन तक के लिए इन्कोडिंग कर सकते हैं वस्तुतः यूनिकोड मानक 49194 करैक्टरों के लिए उपलब्ध कराता है। भारतीय भाषाओं के लिए यूनिकोड ISCII-91 (Indian Script Code for Information Interchange-91) का नहीं अपितु ISCII-88 का प्रयोग करता है जो नवीनतम सरकारी मानक है।

2.4.संसाधन एवं संग्रह संबंधी मुद्दे

संसाधित करने के लिए प्रारूप ऐसा होना चाहिए, जिसे आसानी से खोजा जा सके, संक्षिप्त किया जा सके और सुरक्षित रूप से प्रयोग में लाया जा सके। सभी सामान्य यूनिकोड इन्कोडिंग निश्चित आकार की कोड इकाई के कुछ प्रकार का उपयोग करते हैं। इनकोड किए जाने वाले प्रारूप व कोड इकाई पर निर्भर करके इन कोड इकाईयों में एक या अधिक इकाईयाँ यूनिकोड कोड इकाई का प्रतिनिधित्व करती हैं।

UTF-16 लोकप्रिय है, क्योंकि अधिकतर एप्लीकेशन प्रोग्रामिंग इंटरफेस उस समय के हैं, जब यूनिकोड 16-बिट के निश्चित आयाम का होता था। किंतु UTF-16 का उपयोग करने से कैरेक्टर 'बेसिक मल्टीलिंगुल प्लेन' (Basic Multilingual Plane) से परे हो जाते हैं जिससे उनके प्रबंधन में समस्याएँ आती हैं और इन समस्याओं को UTF-32 से भी हल नहीं किया जा सकता है। दूसरी तरफ UTF-8 का उपयोग करने से पुराने संयंत्रों द्वारा उत्पन्न इनपुट पर काम किया जा सकता है। इसके अतिरिक्ति UTF-8 स्ट्रिंग को विश्वसनीय तरीके से सरल अनुमानी अल्गोरिद्म (Heuristic Algorithm) के द्वारा जाना जा सकता है। UTF- 8किसी भी यूनिकोड कैरेक्टर को इनकोड कर सकता है। अलग-अलग भाषाओं में फ़ाइलें सही कोड-पृष्ठ या फ़ॉन्ट का चयन किए बिना ही सही ढंग से प्रदर्शित की जा सकती हैं। उदाहरणस्वरूप, एक ही पाठ(Text) में किसी विशेष कोड को डाले बिना या इन्कोडिंग को बदलने के लिए मैनुअल सेटिंग किए बिना चीनी व अरबी भाषा पर काम किया जा सकता है।

UTF-16 और UTF-32 बाइटोन्मुख (byte oriented) नहीं हैं, इसलिए इन्हें किसी बाइटोन्मुख नेटवर्क में भेजने या किसी बाइटोन्मुख फ़ाइल में संग्रह करने के लिए किसी बाइट क्रम का चयन आवश्यक है। UTF-8 में यह समस्या उत्पन्न नहीं होती क्योंकि यह बाइटोन्मुख है। विकार की स्थिति में कुछ इन्कोडिंग दूसरों की अपेक्षा बेहतर तरीके से संभल जाती हैं। UTF-8 और UTF-EBCDIC (Extended Binary Coded Decimal Interchange Code) इस मामले में सर्वोत्तम हैं चूँकि वे हर अगले कोड बिंदु के प्रारंभ में हमेशा पुनर्समक्रमिक हो जाते हैं। UTF-16 और UTF-32 अगले कोड बिंदु के प्रारंभ में पुनर्समक्रमिक होकर विकृत या परिवर्तित बाइटों को संभाल तो लेते हैं किंतु कुछ असंगत लुप्त और नकली बाइट आगामी सभी पाठ को दूषित कर देते हैं।

2.5. भारतीय लिपियों एवं भाषाओं में संगणन

भारतीय लिपियों एवं भाषाओं में संगणन का अर्थ है – वेब विकास, डेटाबेस प्रबंधन, कंप्यूटर अनुप्रयोगों का स्थानीकरण, सॉफ्टवेरों व इनपुट विधियाँ का निर्माण, ओसीआर (ऑप्टिकल कैरेक्टर रिकॉगनिशन), वर्तनी-परीक्षक, वाक से पाठ और पाठ से वाक अनुप्रयोगों का निर्माण। भाषा-सीमा के बिना मानव-मशीन संपर्क को सुगम बनाने के लिए सूचना प्रक्रमण सयंत्रों का विकास, बहुभाषी ज्ञान संसाधनों का निर्माण और

उनका अधिगम, तथा नवीन उपभोक्ता पदार्थों व सेवाओं से जोड़ने का प्रयत्न किया जा रहा है जो भारतीय बहुभाषी संदर्भ में अत्यन्त महत्त्वपूर्ण है।

अधिकांशतः भारतीय लिपियाँ आजकल कंप्यूटर और इंटरनेट पर काम करने के लिए यूनिकोड का प्रयोग कर रही हैं, जिसके कारण कंप्यूटर पर भारतीय भाषाओं में लेखन अत्यधिक सुगम हो गया है। इसके लिए कई विधियाँ हैं जिनमें से कुछ निम्नलिखित हैं-

1. **इंस्क्रिप्ट**

इंस्क्रिप्ट (इंडियन स्क्रिप्ट), भारतीय लिपियों के लिए मानक कुंजीपटल है। यह कंप्यूटर के लिए एक स्पर्श टाइपिंग कुंजीपटल लेआउट है। यह कुंजीपटल लेआउट भारतीय भाषाओं में लेखन हेतु मानकीकृत की गयी है। इसे सी-डैक द्वारा विकसित किया गया है। यह देवनागरी, बंगाली, गुजराती, गुरुमुखी, कन्नड़, मलयालम, उड़िया, तमिल और तेलुगु आदि सहित 12 भारतीय लिपियों के लिए मानक कुंजीपटल है। आजकल यह विन्डोज़ (2000, XP, Vista, 7), लिनक्स और मैकिन्टॉश सहित सभी प्रमुख ऑपरेटिंग सिस्टम में अंतर्निहित होता है। यह कुछ मोबाइल फोन में भी उपलब्ध है।

1. **ध्वन्यात्मक लिप्यंतरण**

यह एक टाइपिंग विधि है, जिसमें उपयोगकर्ता रोमन कैरेक्टर में लिखता है और यह ध्वन्यात्मक रूप से समकक्ष लिपि में परिवर्तित हो जाता है। इस प्रकार का रूपांतरण ध्वनि-पाठ संपादकों (phonetic text editors), शब्द संसाधकों (word processors) और सॉफ्टवेयर प्लग-इन के द्वारा किया जाता है। इसके अतिरिक्त कुछ इनपुट मेथड एडीटर (IME) भी उपलब्ध हैं जिनकी सहायता से किसी भी प्रकार के अनुप्रयोग में भारतीय भाषाओं में इनपुट किया जा सकता है। इनमें से प्रमुख हैं -पाठ या कोई वेबसाइट पता लिखें या किसी दस्तावेज़ का अनुवाद करें।

2. अँग्रेजी से हिंदी में अनुवाद

बराह आईएमई, इंडिक आईएमई, गूगल इंडिक ट्रांसलिटरेशन आईएमई और माइक्रोसॉफ्ट इंडिक भाषा इनपुट टूल इत्यादि।

3. हिंदी भाषा में वेबसाइटों की वर्तमान स्थिति एवं यूनिकोड

इस वर्ग के अंतर्गत हम हिंदी की वेबसाइटों की वर्तमान स्थिति एवं इसकी यूनिकोड में उपलब्धता पर विचार करते हैं। हिंदी वेब निर्देशिका (Hindi Website Directory) के अनुसार हिंदी की महत्त्वपूर्ण वेबसाइटें निम्नलिखित हैं–

ब्लॉग [296], समाचारपत्र [114]

- नवभारत टाईम्स [http://navbharattimes.indiatimes.com]
- दैनिक जागरण [http://www.jagran.com]
- बी.बी.सी. हिंदी खबरें [http://www.bbc.co.uk/hindi]
- अमर उजाला [http://www.amarujala.com]
- राष्ट्रीय सहारा [http://www.rashtriyasahara.com/Home.aspx]

हिंदीपत्रिकाएँ [97]

- गृह सहेली [http://www.grehsaheli.com/]
- साहित्य कुंज [http://www.sahityakunj.net]
- चंदामामा [http://www.chandamama.com/hindi]
- हिंदी नेस्ट [http://www.hindinest.com/]
- गीता प्रेस गोरखपुर [http://www.gitapress.org/hindi/homeH.htm]

भारत सरकार [24

- केन्द्रीय हिंदी निदेशालय [http://hindinideshalaya.nic.in]
- राष्ट्रीय ज्ञान आयोग [http://knowledgecommission.gov.in/hindi/default.asp]
- भारत सरकार राजभाषा विभाग [http://www.rajbhasha.gov.in/]
- केन्द्रीय हिंदी संस्थान [http://www.hindisansthan.org/hi/centers/centers.htm]

- भारतीय रेल [http://www.indianrailways.gov.in/hindi/hindex.htm]

वेब पोर्टल [58]

- तरकश [http://www.tarakash.com]
- शैक्षिक सॉफ़्टवेयर [http://educationalsoftware.wikidot.com/]
- कैफेहिंदी [http://cafehindi.com]
- स्वतन्त्र आवाज़ [http://www.swatantraawaz.com]

शिक्षा [36]

- जवाहरलाल नेहरू विश्वविद्यालय [http://www.jnu.ac.in/Hindi/]
- भारतीय तकनीकी संस्थान, दिल्ली [http://www.iitd.ac.in/hindi/index/html]
- एन.सी.पी.यू.एल. [http://www.urducouncil.nic.in/Hindiweb.htm]
- भारतीय भाषा संस्थान [http://www.ciil.org/default.aspx]

कुछ अन्य वेबसाइट

- क्षेत्रीय संबंधी [25]
- समाज संबंधी [25]
- ज्योतिष संबंधी [25]
- इंटरनेट संबंधी [24]
- मनोरंजन संबंधी [21]
- धर्म संबंधी [19]
- व्यापार संबंधी [19]
- व्यक्तिगत संबंधी [14]
- खेल संबंधी [12]
- व्यंजन संबंधी [9]
- पर्यटन संबंधी [7]

- ग्रामीण संबंधी [4]
- कला संबंधी [2]
- राजनीति संबंधी [2]

उपर्युक्त वेबसाइटों के निरीक्षण के उपरान्त यह ज्ञात होता है कि हिंदी के लगभग सभी ब्लॉग यूनिकोड में हैं। सर्वाधिक प्रचलित समाचार पत्र भी यूनिकोड में पाए जाते हैं। कतिपय प्रचलित पत्रिकाएँ यूनिकोड में उपलब्ध हैं। भारत सरकार की कुछ साइटें तो यूनिकोड में हैं, किंतु कुछ महत्त्वपूर्ण साइटें यूनिकोड में नहीं भी हैं यथा-

- हिन्दुस्तान एरोनॉटिक्स लिमिटेड http://www.hal-india.com/hindi/index_h.asp
- भारतीय स्टेट बैंक
  http://www.statebankofindia.com/user.htm?action=indexhindi#
- एन.सी.ई.आर.टी.(NCERT) [http://www.ncert.nic.in]

प्रमुख विश्वविद्यालय एवं संस्थानों की वेबसाइटें भी यूनिकोड में दृष्टिगत हैं जैसे जवाहरलाल नेहरू विश्वविद्यालय (JNU), दिल्ली विश्वविद्यालय (DU), भारतीय तकनीकी संस्थान (IITs), नेशनल कौंसिल फॉर प्रमोशन ऑफ उर्दू लैंग्वेज (NCPUL) और भारतीय भाषा संस्थान (CIIL) आदि। अन्य साइट जैसे भारतीय डाक अभी निर्माणाधीन हैं। कुछ प्रमुख वेब पोर्टल भी यूनिकोड में मिलते हैं। हिंदी विकीपीडिया भी यूनिकोड में उपलब्ध है।

4. यूनिकोड एवं अन्य भारतीय भाषाएँ

हिंदी में लगभग 300 वेबसाइटें दृष्टिगत होती हैं। इसमें अधिकांशतः महत्त्वपूर्ण वेबसाइट यूनिकोड में हैं। हिंदी की अपेक्षा अन्य भारतीय भाषाओं में यूनिकोड वेबसाइटों की संख्या न्यून है।

4.1. हिंदी की वेबसाइटों से अन्य भाषायी वेबसाइटों की मात्रात्मक एंव गुणात्मक तुलना

4.1.1. हिंदी–पंजाबी

पंजाब राज्य सरकार की सरकारी वेबसाइटें पंजाबी में नहीं हैं। किंतु पंजाबी भाषा के अधिकतर समाचारपत्र यूनिकोड में हैं-

- Rozanaspokesman http://www.rozanaspokesman.com/
- Chardhikala http://www.chardhikala.com/
- Ajit *http://newspaper.ajitjalandhar.com/*(*Image*)
- Jagbani http://www.jagbani.com/

4.1.2. हिंदी–बांग्ला

- बांग्ला भाषा के अधिकांशतः समाचारपत्र यूनिकोड में नहीं हैं- http://www.uttarbangasambad.com/, http://www.dailydesherkatha.com/
- कुछ प्रमुख बंगाली पत्रिकाएँ यूनिकोड में हैं -http://www.guruchandali.com /, http://www.ichchhamoti.org/, http://www.tilottamabangla.com
- किंतु सर्वाधिक पत्रिकाएँ यूनिकोड में नहीं हैं – http://www.shatorupa.com/, http://www.balaka.co.in/
- बांग्ला भाषा में ब्लॉग अधिकतर यूनिकोड में हैं –http://royesoye.blogspot.com/, http://santwanastar.blogspot.com/

4.1.3. हिंदी-कन्नड़:- कर्नाटक राज्य सरकार की सरकारी वेबसाइट का मुख्य पृष्ठ कन्नड़ यूनिकोड में हैं। कुछ यूनिकोड वेब पृष्ठ मिश्रित भाषा में हैं – http://web1.kar.nic.in/fcslpg1/

जहाँ तक समाचारपत्र और पत्रिकाओं का प्रश्न है, कन्नड़ भाषा में प्रजावाणी, उदयवाणी जैसे समाचारपत्र यूनिकोड में हैं – किंतु यूनिकोड पत्रिकाओं की संख्या न के बराबर है। इसके अतिरिक्ति कन्नड़ भाषा में सभी ब्लॉग यूनिकोड में हैं।

4.1.4. हिंदी–अन्य भाषाएँ

नेपाली, असमी, गुजराती, तमिल, तेलुगु, आदि भाषाओं में यूनिकोड वेबसाइटों की संख्या कम हैं तथा मणिपुरी, मैथिली, कोंकणी आदि भाषाओं में यूनिकोड वेबसाइटें नहीं है।

5. निष्कर्ष

यूनिकोड आंतरिक संस्करण (Internal Processing) और पाठ्य-संग्रह हेतु प्रभावी प्रणाली बन गयी है। इसका प्रयोग विशिष्टतः नवीन सूचना प्रसंस्करण व्यवस्था के निर्माण में किया जाता है। यह कम्प्यूटर प्रसंस्करण हेतु पाठ (Text) के निरूपण के लिए उपयुक्त है। विश्व में सभी लिपिबद्ध भाषाओं के लिए सभी अक्षरों को इनकोड करने की क्षमता यूनिकोड में विद्यमान है।

अतः भाषा प्रौद्योगिकी के क्षेत्र में यूरोपीय भाषाओं की तुलना में भारतीय भाषाओं की वर्तमान स्थिति को देखते हुए यह अनिवार्य है कि भारतीय भाषाओं में अतिशीघ्र विषयवस्तु का निर्माण तथा इसका यूनिकोड में रूपान्तरण किया जाए, जिससे भारतीय भाषाएँ विश्व की अन्य भाषाओं के समकक्ष तथा सार्वभैमिक बन सकें। इसका लाभ भाषाविज्ञान, तकनीक विज्ञान, बहुभाषा अध्ययन, अनुसंधान एवं व्यापार आदि क्षेत्रों को मिल सकेगा। बहुभाषी पाठ सामग्री (Text) पर कार्य करने वाले भी लाभान्वित होंगे। अन्य शब्दों में यह कह सकते हैं कि सम्पूर्ण शिक्षा जगत को अत्यन्त लाभ प्राप्त होगा।

संदर्भ सूची:

- http://dir.hinkhoj.com/
- http://en.wikipedia.org/wiki/Unicode
- http://en.wikipedia.org/wiki/InScript_keyboard
- http://unicode.org/standard/WhatIsUnicode.html
- http://en.wikipedia.org/wiki/Indian_Script_Code_for_Information_Interchange
- http://varamozhi.sourceforge.net/iscii91.pdf
- http://tdil.mit.gov.in
- http://java.about.com/od/programmingconcepts/a/unicode.htm

- http://publib.boulder.ibm.com/infocenter/db2luw/v8/index.jsp?topic=/com.ibm.db2.udb.doc/admin/c0004816.htm
- http://tlt.its.psu.edu/suggestions/international/bylanguage/southasia.html

Citation: ऋचा & अवस्थी, सत्येन्द्र कुमार (2011). भाषा-प्रौद्योगिकी में यूनिकोड : भारतीय भाषाओं के संदर्भ में, HindiTech: A Blind Double Peer Reviewed Bilingual Web-Research Journal, 2 (3), 10-20. URL: https://hinditech.in/bhasha-praudyogiki-men-unicod-bhartiy-bhashaon-ke-sandarbh-men/