यूनिकोड : भाषा प्रोद्योगिकी में हिंदी का नया औजार

■ राकेश शर्मा

कम्प्यूटर, मूल रूप से, नंबरों से सम्बंध रखते हैं। ये प्रत्येक अक्षर और वर्ण के लिए एक नंबर निर्धारित करके अक्षर और वर्ण संग्रहित करते हैं। यूनिकोड का आविष्कार होने से पहले, ऐसे नंबर देने के लिए सैंकडों विभिन्न संकेत लिपि प्रणालियां थीं। किसी एक संकेत लिपि में पर्याप्त अक्षर नहीं हो सकते हैं : उदाहरण के लिए, यूरोपीय संघ को अकेले ही, अपनी सभी भाषाओं को कवर करने के लिए अनेक विभिन्न संकेत लिपियों की आवश्यकता होती है। अंग्रेजी जैसी भाषा के लिए भी, सभी अक्षरों, विराम चिन्हों और सामान्य प्रयोग के तकनीकी प्रतीकों हेतु एक ही संकेत लिपि पर्याप्त नहीं थी।

ये संकेत लिपि प्रणालियां परस्पर विरोधी भी हैं। इसीलिए, दो संकेत लिपियां दो विभिन्न अक्षरों के लिए, एक ही नंबर प्रयोग कर सकती हैं, अथवा समान अक्षर के लिए विभिन्न नम्बरों का प्रयोग कर सकती हैं। किसी भी कम्प्यूटर (विशेष रूप से सर्वर) को विभिन्न संकेत लिपियां संभालनी पड़ती है; फिर भी जब दो विभिन्न संकेत लिपियों अथवा प्लैटफॉर्मों के बीच डाटा भेजा जाता है तो उस डाटा के हमेशा खराब होने का जोखिम रहता है।
यूनिकोड प्रत्येक अक्षर के लिए एक विशेष नम्बर प्रदान करता है –

चाहे कोई भी प्लैटफॉर्म हो,

चाहे कोई भी प्रोग्राम हो,

चाहे कोई भी भाषा हो

देवनागरी लिपि में यूनिकोड की विशेषताएं

देवनागरी यूनिकोड की परास (रेंज) 0900 से 097F तक है। (दोनो संख्याएं षोडषाधारी हैं)

क्ष, त्र एवं ज्ञ के लिये अलग से कोड नहीं है। इन्हें संयुक्त वर्ण मानकर अन्य संयुक्त वर्णों की भांति इनका अलग से कोड नहीं दिया गया है।

इस रेंज में बहुत से ऐसे वर्णों के लिये भी कोड दिये गये हैं जो सामान्यतः हिन्दी में व्यवहृत नहीं होते। किन्तु मराठी, सिन्धी, मलयालम आदि को देवनागरी में सम्यक ढंग से लिखने के लिये आवश्यक हैं।

नुक्ता वाले वर्णों (जैसे ज़) के लिये यूनिकोड निर्धारित किया गया है। इसके अलावा नुक्ता के लिये भी अलग से एक यूनिकोड दे दिया गया है। अतः नुक्तायुक्त अक्षर यूनिकोड की दृष्टि से दो प्रकार से लिखे जा सकते हैं – एक बाइट यूनिकोड के रूप में या दो बाइट यूनिकोड के रूप में। उदाहरण के लिये ज़ को ' ज ' के बाद नुक्ता (μ) टाइप करके भी लिखा जा सकता है।

देवनागरी के यूनिकोड चिह्न

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
U+090x		◌ँ	◌ं	◌ः	ऄ	अ	आ	इ	ई	उ	ऊ	ऋ	ऌ	ऍ	ऎ	ए
U+091x	ऐ	ऑ	ऒ	ओ	औ	क	ख	ग	घ	ङ	च	छ	ज	झ	ञ	ट
U+092x	ठ	ड	ढ	ण	त	थ	द	ध	न	ऩ	प	फ	ब	भ	म	य
U+093x	र	ऱ	ल	ळ	ऴ	व	श	ष	स	ह			◌़	ऽ	◌ा	◌ि
U+094x	◌ी	◌ु	◌ू	◌ृ	◌ॄ	◌ॅ	◌ॆ	◌े	◌ै	◌ॉ	◌ॊ	◌ो	◌ौ	◌्		
U+095x	ॐ	◌॑	◌॒	◌॓	◌॔				क़	ख़	ग़	ज़	ड़	ढ़	फ़	य़
U+096x	ॠ	ॡ	◌ॢ	◌ॣ	।	॥	०	१	२	३	४	५	६	७	८	९
U+097x	॰		ॲ									ॻ	ॼ		ॾ	ॿ

UTF-8, UTF-16 और UTF-32

1. यूनिकोड का मतलब है सभी लिपिचिह्नों की आवश्यकता की पूर्ति करने में सक्षम 'एकसमान मानकीकृत कोड' ।

2. पहले सोचा गया था कि केवल 16 बिट के माध्यम से ही दुनिया के सभी लिपिचिह्नों के लिये अलग–अलग कोड प्रदान किये जा सकेंगे। बाद में पता चला कि यह कम है। फिर इसे 32 बिट कर दिया गया। अर्थात इस समय दुनिया का कोई संकेत नहीं है जिसे 32 बिट के कोड में कहीं न कहीं जगह न मिल गयी हो।

3. 8 बिट के कुल 2पर घात 8 = 256 अलग–अलग बाइनरी संख्याएं बन सकती हैं; 16 बिट से 2 पर घात 16 = 65536 और 32 बिट से 4294967296 भिन्न (distinct) बाइनरी संख्याएं बन सकती हैं।

4. यूनिकोड के तीन रूप प्रचलित हैं। UTF–8, UTF–16 और UTF–32.

5. इनमें अन्तर क्या है? मान लीजिये आपके पास दस पेज का कोई टेक्स्ट है जिसमें रोमन, देवनागरी, अरबी, गणित के चिन्ह आदि बहुत कुछ हैं। इन चिन्हों के यूनिकोड अलग–अलग होंगे। यहाँ ध्यान देने योग्य बात है कि कुछ संकेतों के 32 बिट के यूनिकोड में शुरू में शून्य ही शुन्य हैं (जैसे अंग्रेजी के संकेतों के लिये)। यदि शुरुआती शून्यों को हटा दिया जाय तो इन्हें केवल 8 बिट के द्वारा भी निरूपित किया जा सकता है और कहीं कोई भ्रम या कांफ्लिक्ट नहीं होगा। इसी तरह रूसी, अरबी, हिब्रू आदि के यूनिकोड ऐसे हैं कि शून्य को छोड़ देने पर उन्हें प्राय: 16 बिट = 2 बाइट से निरूपित किया जा सकता है। देवनागरी, जापानी, चीनी आदि को आरम्भिक शून्य हटाने के बाद प्राय: 24 बिट = तीन बाइट से निरूपित किया जा सकता है। किन्तु बहुत से संकेत होंगे जिनमें आरम्भिक शून्य नहीं होंगे और उन्हें निरूपित करने के लिये चार बाइट ही लगेंगे।

6. बिन्दु (5) में बताए गये काम को UTF–8, UTF–16 और UTF–32 थोड़ा अलग अलग ढंग से करते हैं। उदाहरण के लिये यूटीएफ–8 क्या करता है कि कुछ लिपिचिह्नों के लिये 1 बाइट , कुछ के लिये 2 बाइट, कुछ के लिये तीन और चार बाइट इस्तेमाल करता है। लेकिन UTF–16 इसी काम के लिये 16 न्यूनतम बिटों का इस्तेमाल करता है। अर्थात जो चीजें UTF-8 में केवल एक बाइट जगह लेती थीं वे अब 16 बिट==2 बाइट के द्वारा निरूपित होंगी। जो UTF–8 में 2 बाइट लेतीं थी यूटीएफ–16 में भी दो ही लेंगी। किन्तु पहले जो संकेतदि 3 बाइट या चार बाइट में निरूपित होते थे यूटीएफ–16 में 32 बिट=4 बाइट के द्वारा निरूपित किये जायेंगे। (आपके पास बड़ी–बडी ईंटें हों और उनको बिना तोड़े खम्भा बनाना हो तो खम्भा ज्यादा बड़ा ही बनाया जा सकता है।)

7. लगभग स्पष्ट है कि प्राय: UTF-8 में इनकोडिंग करने से UTF–16 की अपेक्षा कम बिट्स लगेंगे।

8. इसके अलावा बहुत से पुराने सिस्टम 16 बिट को हैंडिल करने में अक्षम थे। वे एकबार में केवल 8–बिट ही के साथ काम कर सकते थे। इस कारण भी UTF-8 को अधिक अपनाया गया। यह अधिक प्रयोग में आता है।

9. UTF-16 और UTF-32 के पक्ष में अच्छाई यह है कि अब कम्प्यूटरों का हार्डवेयर 32 बिट या 64 बिट का हो गया है। इस कारण UTF-8 की फाइलों को 'प्रोसेस' करने में UTF-16, UTF-32 वाली फाइलों की अपेक्षा अधिक समय लगेगा।

यूनिकोड से आया है बदलाव !

यूनिकोड, प्रत्येक अक्षर के लिए एक विशेष नंबर प्रदान करता है, चाहे कोई भी प्लैटफॉर्म हो, चाहे कोई भी प्रोग्राम हो, चाहे कोई भी भाषा हो। यूनिकोड स्टैंडर्ड को ऐपल, एच.पी., आई.बी.एम., जस्ट सिस्टम, माईक्रोसॉफ्ट, औरेकल, सैप, सन, साईबेस, यूनिसिस जैसी उद्योग की प्रमुख कम्पनियों और कई अन्य ने अपनाया है। यूनिकोड की आवश्यकता आधुनिक मानदंडों, जैसे एक्स.एम.एल., जावा, एकमा स्क्रिप्ट (जावा स्क्रिप्ट), एल.डी.ए.पी., कोर्बा 3.0, डब्ल्यू.एम.एल. के लिए होती है और यह आई.एस.ओ./आई.ई.सी. 10646 को लागू करने का अधिकारिक तरीका है। यह कई संचालन प्रणालियों, सभी आधुनिक ब्राउजरों और कई अन्य उत्पादों में होता है। यूनिकोड स्टैंडर्ड की उत्पति और इसके सहायक उपकरणों की उपलब्धता, हाल ही के अति महत्वपूर्ण विश्वव्यापी सॉफ्टवेयर टेक्नोलॉजी रुझानों में से हैं।

यूनिकोड को ग्राहक–सर्वर अथवा बहु–आयामी उपकरणों और वेबसाइटों में शामिल करने से, परंपरागत उपकरणों के प्रयोग की अपेक्षा खर्च में अत्यधिक बचत होती है। यूनिकोड से एक ऐसा अकेला सॉफ्टवेयर उत्पाद अथवा अकेला वेबसाइट मिल जाता है, जिसे री–इंजीनियरिंग के बिना विभिन्न प्लैटफॉर्मों, भाषाओं और देशों में उपयोग किया जा सकता है। इससे डाटा को बिना किसी बाधा के विभिन्न प्रणालियों से होकर ले जाया जा सकता है।

यूनिकोड के प्रयोग से क्या लाभ हैं ?

1. एकरूपता ।

2. कार्यालय के सभी कार्य कंप्यूटर पर हिंदी में आसानी से होते हैं जैसे वर्ड प्रोसेसिंग, डाटा प्रोसेसिंग, ई–मेल, वैबसाइट निर्माण आदि ।

3. हिंदी में बनी फाइलों का आसानी से आदान–प्रदान कर सकते हैं ।

4. हिंदी की–वर्ड पर गूगल या किसी अन्य सर्च इंजन में सर्च कर सकते हैं ।

कंप्यूटरों पर हिंदी में कार्य करने के लिए कौन–कोन से की–बोर्ड ले–आउट विकल्प हैं ?

- इनसक्रिप्ट (standardized by DOE in 1986)
- रेमिंग्टन
- फोनेटिक (विजुअल फोनेटिक ले–आउट)

इनसक्रिप्ट की–बोर्ड ले–आउट सीखने के लिए क्या कोई ट्यूटर उपलब्ध है ?

INSCRIPT टाइपिंग सीखने के लिए www.ildc.in साइट से **हिन्दी एवं अंग्रेजी के लिए आसान टंकण प्रशिक्षक** डाउनलोड करके एक घंटा प्रतिदिन के हिसाब से 8–10 दिनों में INSCRIPT टाइपिंग सीखी जा सकती है ।

यूनिकोड को सक्रिय कैसे किया जा सकता है ?

यूनिकोड enable करने की प्रक्रिया आप राजभाषा विभाग की साइट http://rajbhasha.nic.in से ई–उपकरण पर क्लिक करके ले सकते हैं ?

क्या किसी भी हिंदी फ़ॉन्ट की सामग्री को शतप्रतिशत शुद्धता के साथ यूनिकोड में बदला जा सकता है ?

हां, सूचना प्रोद्योगिकी विभाग की वेबसाइट http://ildc.in साइट से सार्वत्रिक हिन्दी फॉन्ट कोड एवं भंडारण कोड परिवर्तक फोंट कनवर्टर के माध्यम से नॉन यूनिकोड फोंट की सामग्री को यूनिकोड में बदला जा सकता है । यह सॉफ्टवेयर निशुल्क है ।

यूनिकोड कन्सॉर्शियम के बारे में

यूनिकोड कन्सॉर्शियम, लाभ न कमाने वाला एक संगठन है जिसकी स्थापना यूनिकोड स्टैंडर्ड, जो आधुनिक सॉफ्टवेयर उत्पादों और मानकों में पाठ की प्रस्तुति को निर्दिष्ट करता है, के विकास, विस्तार और इसके प्रयोग को बढ़ावा देने के लिए की गई थी। इस कन्सॉर्शियम के सदस्यों में, कम्प्यूटर और सूचना उद्योग में विभिन्न निगम और संगठन शामिल हैं। इस कन्सॉर्शियम का वित्तपोषण पूर्णतः सदस्यों के शुल्क से किया जाता है। यूनिकोड कन्सॉर्शियम में सदस्यता, विश्व में कहीं भी स्थित उन संगठनों और व्यक्तियों के लिए खुली है जो यूनिकोड का समर्थन करते हैं और जो इसके विस्तार और कार्यान्वयन में सहायता करना चाहते हैं।

यूनिकोड और इन्टरनेट पर हिन्दी साक्षरता

ज्ञान के क्षेत्र में मानवता की सबसे बड़ी उपलब्धि इंटरनेट है । इसके द्वारा कम्प्यूटर और टेलीफोन प्रौद्योगिकियों के संयोजन से सूचना और संचार के क्षेत्र में चमत्कारी परिवर्तन लाये गये हैं । आजकल यदि कोई व्यक्ति इंटरनेट का उपयोग नहीं कर सकता तो वह व्यवहारिक रूप से असाक्षर माना जाता है।

इंटरनेट के ज़रिए ई–मेल, ई–वाणिज्य, ई–शासन आदि संभव हो सके हैं । इंटरनेट पर हम एक वेबसाइट द्वारा अपने विचारों और गतिविधियों को प्रकाशित कर सकते हैं । दुनिया में कोई भी जानकारी इंटरनेट पर लगभग तुरंत किसी भी खर्चे के बिना प्राप्त की जा सकती है।

आजकल पूरी दुनिया में इंटरनेट का उपयोग हो रहा है लेकिन कुछ देशों में यह प्रयोग कम है और कुछ में ज़्यादा । भारत में 5 प्रतिशत जनसंख्या इंटरनेट का उपयोग करती है । यह अनुपात विकसित देशों की तुलना में, जहां इंटरनेट को 70 प्रतिशत से अधिक लोगों द्वारा इस्तेमाल किया जाता है, बहुत कम है।

भारत में अंग्रेजी को ही इंटरनेट की भाषा माना जाता है, लेकिन यह स्थिति यूनिकोड के माध्यम से बदल रही है । विश्व की अधिकांश भाषाओं को यूनिकोड के द्वारा इंटरनेट पर देखा जा सकता है । भारतीय भाषाओं, खासकर हिंदी, को भी अब यूनिकोड के माध्यम से इंटरनेट पर देखा जाता है।

हिन्दी भारत की राष्ट्र भाषा है और दुनिया की तीसरी सबसे अधिक बोली जाने वाली भाषा है । यह आवश्यक है कि इंटरनेट पर हिन्दी का प्रयोग और बढ़े । इस संबंध में उचित क़दम उठाना आवश्यक है अन्यथा हिन्दी बोलने वाले लोग सूचना, संचार और ज्ञान के क्षेत्र में पीछे रह जायेंगे।

इंटरनेट का उपयोग नहीं करने वाले हिन्दी भाषी लोगों को दो श्रेणियों में बाँटा जा सकता है । पहली श्रेणी में वे लोग हैं जो हिंदी में साक्षर हैं लेकिन इंटरनेट का उपयोग नहीं कर सकते । दूसरी श्रेणी में वे लोग आते हैं जो हिन्दी बिलकुल नहीं पढ़–लिख सकते ।

यूनिकोड के माध्यम से इंटरनेट का प्रयोग साक्षर लोगों के लिए बहुत आसान हो गया है । सूचना, समाचार, साहित्य आदि अब यूनिकोड हिन्दी में प्रचुर मात्रा में उपलब्ध हैं । हिन्दी से अन्य भाषाओं में तथा अन्य भाषाओं से हिन्दी में अनुवाद करने की सुविधाएं भी इंटरनेट पर उपलब्ध हैं।

हिन्दी भाषी अशिक्षित लोगों को रोमनागरी और सरल हिन्दी लिपि के माध्यम से आसानी से हिंदी लिखना–पढ़ना सिखाया जा सकता है । हिन्दी को सरल लिपि में सीखने के बाद वे सुगमता से इंटरनेट का उपयोग कर सकते हैं. इंटरनेट पर प्राप्त सरल उपकरण भी स्वयं सीखने के लिए प्रभावी सिद्ध हो सकते हैं।

राष्ट्रीय समुद्र विज्ञान संस्थान,
दोना पावला, गोवा 403 004