

कंप्यूटर पर हिंदी भाषा संसाधन के लिए युनिकोड की भूमिका

----विजय कुमार मल्होत्रा, पूर्व निदेशक (राजभाषा), रेल मंत्रालय, भारत सरकार

<malhotravk@gmail.com>

यह एक ऐतिहासिक संयोग ही है कि कंप्यूटर का विकास सर्वप्रथम ऐसे देशों में हुआ, जिनकी भाषा मुख्यतः अंग्रेजी या रोमन लिपि पर आधारित कोई योरोपीय भाषा थी. कदाचित् यही कारण है कि रोमनेतर भाषाओं में कंप्यूटरसाधित भाषा संसाधन का कार्य कुछ देरी से आरंभ हुआ . इस बात में भी कोई संदेह नहीं कि रैखिक (linear) लिपि होने के कारण रोमन लिपि में सूचना संसाधन का कार्य अपेक्षाकृत सरल भी था. किंतु इस बात का कोई तकनीकी कारण नहीं है कि रोमन लिपि या अंग्रेजी भाषा कंप्यूटर संसाधन के लिए आदर्श समझी जाए . वस्तुतः कंप्यूटर की दो संकेतों की अपनी स्वतंत्र गणितीय भाषा है और उसी भाषा में वे हमारी भाषाओं को ग्रहण करके अपने सभी कार्य संपन्न करते हैं. इसलिए कंप्यूटर के लिए किसी भी भाषा को अपनाने में कोई तकनीकी बाधा नहीं है. वस्तुतः कंप्यूटर में सारी गणनाएँ केवल दो संकेतों (0 और 1) से होती हैं . केवल गणित ही नहीं , तार्किक कथनों को भी 'हाँ' या 'नहीं' के बीजगणित में ढाला जा सकता है . विभिन्न लिपियों के माध्यम से प्राकृतिक भाषाओं के पाठ के कुंजीयन के लिए द्वि-आधारी कोड (binary code) बनाए गए हैं. रोमन लिपि के कोड को आस्की -7 कोड (ASCII अर्थात् American Standard Code for Information Interchange) कहा जाता है. इस 7 अंकीय कोड में रोमन लिपि के सभी अक्षर , अंक और विराम चिह्न समाहित हो जाते हैं . फ्रांसीसी, जर्मन, इतालवी, पुर्तगाली, डच और स्पेनिश आदि रोमन-आधारित भाषाओं में प्रयुक्त सभी विशेषक चिह्नों (diacritic marks) को भी इसमें समाविष्ट किया गया है . जैसे A के लिए आस्की -7 कोड में 8 बिट की 0 और 1 की यह गणना 01000001के रूप में निर्धारित की गई है. जैसे ही इसे डीकोडित किया जाएगा, यह A में परिणत हो जाएगा.

जिस प्रकार आस्की-7 कोड एक ऐसी कोडिंग प्रणाली है, जो सभी रोमन-आधारित भाषाओं को कंप्यूटर के स्मृति कोश में भंडारण या स्टोरेज के लिए विकसित की गई है, ठीक उसी प्रकार कंप्यूटर में सभी भारतीय भाषाओं के भंडारण या स्टोरेज के लिए इस्की-8 कोडिंग प्रणाली (ISCI अर्थात् Indian Standard Code for Information Interchange) की परिकल्पना की गई है . वस्तुतः यह आस्की-7 का ही विस्तृत रूप है . सच तो यह है कि आस्की-7 भी मूलतः 8 बिट की कोडिंग प्रणाली ही है , लेकिन उसके अंतर्गत रोमन लिपि के छोटे और कैपिटल अक्षरों, उसके विराम चिह्नों तथा अन्य रोमन-आधारित भाषाओं में प्रयुक्त सभी विशेषक चिह्नों को केवल 7 बिट में ही समाहित कर लिया गया है. उदाहरण

के लिए A के लिए निर्धारित कोड 01000001 में पहला बिट 0 रखा गया है. इसी प्रकार सभी अक्षरों के कोडों में पहला अक्षर 0 ही है. इस 0 को मिलाकर A के लिए निर्धारित कोड 01000001 की गणना भी 8 बिट की हो जाती है. हिंदी और अन्य भारतीय भाषाओं के लिए कोडिंग प्रणाली निर्धारित करते समय आई .आई.टी, कानपुर के वैज्ञानिकों के सामने निम्नलिखित दो बातें बहुत स्पष्ट थीं. पहली बात तो यह थी कि भारतीय भाषाओं के लिए निर्धारित की जाने वाली कोडिंग प्रणाली में रोमन लिपि के समावेश की सुविधा भी सहज रूप में उपलब्ध होनी चाहिए . साथ ही दूसरी बात यह थी कि रोमन लिपि के लिए विकसित क्वेटी कुंजीपटल में ही हिंदी और अन्य भारतीय लिपियों में टाइप करने की सुविधा उपलब्ध कराई जानी चाहिए , ताकि रोमन लिपि और भारतीय भाषाओं में परस्पर सह-अस्तित्व की भावना विकसित हो सके अर्थात् हिंदी और अंग्रेजी में साथ-साथ काम किया जा सके.

यदि हम भारतीय भाषाओं के परिदृश्य को देखें तो पाएँगे कि भारत में 1650 से अधिक भाषाएँ और बोलियाँ प्रचलित हैं ,जिनमें से 22 भारतीय भाषाएँ भारत के संविधान की 8 वीं अनुसूची में राष्ट्रीय भाषाओं के रूप में परिगणित की गई हैं . फिर भी यह देश एकभाषिक क्षेत्र (Linguistic Zone) है. एकभाषिक क्षेत्र की मूल परिकल्पना को समझने के लिए यदि हम विश्व-भर की सभी प्राकृतिक भाषाओं के मूल की ओर दृष्टि डालें तो पाएँगे कि विश्व-भर की भाषाओं में दो स्पष्ट लक्षण दिखाई पड़ते हैं : सार्वभौमिक लक्षण और भाषा-विशिष्ट लक्षण. सार्वभौमिक लक्षण वे हैं जो हिंदी , तमिल, अंग्रेजी, चीनी और अरबी जैसी विभिन्न भाषा-परिवारों से जुड़ी भाषाओं में भी समान रूप से पाए जाते हैं. उदाहरण के लिए, 'खाया' एक सकर्मक क्रिया है , जो खाए जाने के लिए एक कर्म और खाना क्रिया संपन्न करने के लिए एक कर्ता की आकांक्षा करती है और साथ ही यह भी अपेक्षा करती है कि उसका कर्ता सजीव हो . इस क्रिया की मूल संरचना में ये सभी सार्वभौमिक लक्षण दिखाई पड़ते हैं. इसप्रकार इसकी सकर्मकता विश्व की सभी भाषाओं में समान है,लेकिन कर्ता के साथ 'ने' का प्रयोग हिंदी का भाषा-विशिष्ट लक्षण है. कुछ ऐसे भी लक्षण होते हैं जो भाषा-वर्ग विशिष्ट होते हैं. उदाहरण के लिए, एक विशेष वाक्य सँचे में कर्ता के साथ 'को' का प्रयोग सभी भारतीय भाषाओं में समान रूप से पाया जाता है ; जैसे, हिंदी में 'राम को बुखार है', मराठी में 'रामला ताप आहे', तमिल में, 'रामक्कु ज्वरम्', मलयालम में ' रामन्नु पनियानु ', कन्नड़ में 'रामनिगे ज्वर दिगे', बँगला में 'रामेर ताप आछे' और अंग्रेजी में इसका अनुवाद होगा, 'Ram has a fever'. अंग्रेजी में आप देखेंगे कि 'को' का वाचक कोई परसर्ग या पूर्वसर्ग नहीं है. इससे स्पष्ट होता है कि भारत एकभाषिक क्षेत्र है. यदि हम इन लक्षणों के विश्लेषण के लिए भाषा प्रौद्योगिकी का उपयोग करें तो हम

भारतीय भाषाओं में कंप्यूटर साधित स्वयं भाषा शिक्षक, ऑटो-करेक्ट, ग्रामर चैकर और मशीनी अनुवाद जैसे अत्यंत जटिल भाषिक उपकरणों का विकास भी कर सकते हैं.

भारत की अनेक भाषाओं की लिपियाँ तो एक-दूसरे से इतनी भिन्न दिखती हैं कि उनमें समानता के अंतर्निहित सूत्र को खोजना भी सरल नहीं है . उदाहरण के लिए आर्यभाषाओं और द्रविड़ भाषाओं में इतना अंतर दिखाई पड़ता है कि यह विश्वास करना कठिन हो जाता है कि सभी भारतीय भाषाओं की लिपियाँ (उर्दू को छोड़कर) का उद्गम ब्राह्मी लिपि के समान स्रोत से हुआ है . यद्यपि इन भाषाओं के अक्षर ऊपरी तौर पर भिन्न-भिन्न दिखाई पड़ते हैं, लेकिन इनकी वर्णमाला और लिपि समान उच्चारण पद्धति पर आधारित है . उर्दू इसका अपवाद है , क्योंकि इसकी लिपि फ़ारसी- अरबी लिपि पर आधारित है.

यदि ऐतिहासिक दृष्टि से देखें तो अशोक काल से ही हमें उत्तर और दक्षिण भारत में ब्राह्मी लिपि का व्यापक उपयोग मिलने लगता है . चौथी शताब्दी के उत्तरार्ध में ब्राह्मी दो शैलियों में विभक्त हो गई , उत्तरी शैली और दक्षिणी शैली. उत्तर भारत की सभी परवर्ती लिपियाँ ब्राह्मी की उत्तरी शैली से और दक्षिणी भारत की सभी परवर्ती लिपियाँ ब्राह्मी की दक्षिणी शैली से विकसित हुईं. कालांतर में इन लिपियों में इतना अंतर आ गया कि बिना सीखे उत्तर वालों के लिए दक्षिण की किसी लिपि को पढ़ना संभव न रहा और इसीप्रकार बिना सीखे दक्षिण वालों के लिए उत्तर भारत की किसी लिपि को पढ़ना संभव न रहा . इस अंतर का प्रमुख कारण कदाचित् लेखन- सामग्री की भिन्नता भी रहा है . दक्षिण भारत में ताड़ वृक्षों की बहुलता के कारण लेखन- सामग्री के रूप में ताड़पत्रों का उपयोग किया जाता था और उस पर सीधे लिखने से ताड़पत्रों के फट जाने की आशंका रहती थी. इसलिए दक्षिण भारत में नोकदार सूखी कलम से वृत्ताकार रूप में लिखने की परंपरा विकसित हुई. इसके विपरीत उत्तर भारत में चपटी और खोखली कलम से वृत्त के बजाय ऊपर-नीचे की मात्राओं के साथ लिखने की परंपरा विकसित हुई. चौथी शताब्दी में ब्राह्मी की उत्तरी शैली से 'गुप्त लिपि' का जन्म हुआ . इसके अनेक वर्णों की आकृति देवनागरी के वर्णों से मिलती- जुलती है. गुप्त लिपि के वर्णों में शिरोरेखा का प्रयोग स्पष्ट रूप से मिलता है . छठी शताब्दी तक आते-आते गुप्त लिपि कुटिल लिपि में परिवर्तित हो गई और इसी कुटिल लिपि से अन्य अनेक आर्यभाषिक लिपियों के साथ- साथ देवनागरी लिपि का भी विकास हुआ . यह लिपि अपने उद्भव काल से ही सारे देश में प्रयुक्त होने लगी थी . इस प्रकार ब्राह्मी लिपि का वास्तविक राष्ट्रीय स्वरूप देवनागरी के रूप में अवतरित हुआ . आज भी विश्व भर में प्रकाशित संपूर्ण संस्कृत साहित्य अधिकांशतः देवनागरी लिपि में ही पाया जाता है . हिंदी और इसकी अन्य अनेक क्षेत्रीय बोलियाँ

देवनागरी में ही लिखी जाती हैं. संस्कृत के अलावा नेपाली, मराठी और कोंकणी भाषा की आधिकारिक लिपि भी देवनागरी ही है. दक्षिण में इसे नंदिनागरी कहते हैं.

ऐतिहासिक रूप में इसकी प्रामाणिक जानकारी सन् 1837 में मिली. आई.आई.टी. कानपुर के कंप्यूटर वैज्ञानिकों को सभी भारतीय भाषाओं के लिए जिस्ट प्रौद्योगिकी के आधार पर समान कुंजीपटल का विकास करते हुए इस तथ्य का व्यावहारिक अनुभव हुआ. पहली बार इसका सार्वजनिक प्रदर्शन सन् 1983 में नई दिल्ली में आयोजित विश्व हिंदी सम्मेलन में किया गया . क्वेटी कुंजीपटल पर सभी भारतीय भाषाओं को समेटना अपने आप में एक जटिल कार्य था , लेकिन ब्राह्मी लिपि से उद्भव के फलस्वरूप भारतीय लिपियों के समान ध्वन्यात्मक स्वरूप के कारण यह कार्य अत्यंत वैज्ञानिक रूप में संपन्न हो गया . सभी भारतीय भाषाओं के लिए जिस्ट (GIST अर्थात् Graphics and Intelligence based Script Technology) प्रौद्योगिकी के आधार पर समान ध्वन्यात्मक कुंजीपटल और समान कोडिंग प्रणाली का विकास किया गया.

भारतीय लिपियाँ अपने स्वरूप में अक्षरात्मक हैं , लेकिन उनकी वर्णमाला का क्रम ध्वन्यात्मक है और ब्राह्मी लिपि से उद्भव के कारण उनकी विरासत भी एक ही है . कुछ लिपियों में मामूली-सा अंतर होने के कारण कुछ अक्षर अतिरिक्त हैं और कुछ लिपियों में कुछ अक्षर कम हैं. 1986-88 में विकसित ISCI (Indian Standard Code for Information Interchange) कोड में परिवर्धित देवनागरी के अंतर्गत इस पक्ष का भी ध्यान रखा गया और इसे भारतीय मानक ब्यूरो ने मानक के रूप में स्वीकार कर लिया था, लेकिन जब कंप्यूटर के उ पयोग का सवाल आया तो भारतीय भाषाओं में डेटा प्रविष्टि के अनेक विकल्प सामने थे और यही चिंता की बात थी . भारतीय भाषाओं में डेटा प्रविष्टि के लिए डिफॉल्ट विकल्प INSCRIPT (INdian SCRIPT) लेआउट है. इस लेआउट में मानक 101 कुंजीपटल का उपयोग किया जाता है. वर्णों की मैपिंग इस प्रकार से की गई है कि यह सभी भारतीय भाषाओं (बाएँ से दाईं ओर लिखी जाने वाली भाषाओं) के लिए समान कुंजीपटल बन जाता है . इसका प्रमुख कारण यही है कि भारतीय भाषाओं के वर्णों का समुच्चय समान है . हम भारतीय भाषाओं की वर्णमाला के वर्णों को व्यंजन , स्वर, अनुनासिक और संयुक्ताक्षरों में विभाजित कर सकते हैं. प्रत्येक व्यंजन विशिष्ट ध्वनि और स्वर का संयोजन होता है . स्वर शुद्ध ध्वनियों को दर्शाता है . अनुनासिक वे नासिक्य ध्वनियाँ होती हैं, जिनका उच्चारण स्वर के साथ किया जाता है. संयुक्ताक्षर दो या अधिक वर्णों का संयोजन होता है. भारतीय भाषाओं की वर्णमाला की तालिका को स्वर और व्यंजन में विभाजित किया जाता है. स्वर दो प्रकार के होते हैं, दीर्घ और लघु. व्यंजनों को अनेक वर्णों में विभाजित किया जाता है . INSCRIPT लेआउट में यह व्यवस्था प्रतिबिंबित होती है. इसीलिए इसकी व्यवस्था बहुत सरल होती है . इन्स्क्रिप्ट ले आउट में सभी स्वरों को

कुंजीपटल के बाईं ओर रखा गया है और व्यंजनों को दाईं ओर. यह व्यवस्था इसप्रकार से की गई है कि प्रत्येक वर्ग को दो कुंजियों में विभाजित कर दिया गया है . इस प्रकार इन भाषाओं के समान अकारादि क्रम के कारण ही सभी भारतीय भाषाओं के लिए समान कुंजीपटल और समान कोड विकसित किया जा सका है और सभी भारतीय भाषाओं के लिए समान कोडिंग के कारण ही भारतीय लिपियों में परस्पर लिप्यंतरण की सुविधा भी सहज ही उपलब्ध हो जाती है. चूँकि ISCI में रोमन लिपि को भी समाहित किया गया है, इसलिए इंडिक लिपियों अर्थात् भारतीय भाषाओं की लिपियों से रोमन लिपि में भी लिप्यंतरण किया जा सकता है.

माइक्रोसॉफ्ट ने हाल ही में 'विंडोज /ऑफिस हिंदी' के रूप में एक ऐसी ऑपरेटिंग प्रणाली और ऑफिस पैकेज का विकास किया है , जिसमें हिंदी में कुंजीयन या टाइप करने के लिए उस बारहखड़ी का उपयोग किया है , जिसका प्रयोग हिंदी सीखने- सिखाने के लिए शताब्दियों से भारत के गाँव- गाँव में किया जाता रहा है . इसमें संदेह नहीं कि देवनागरी लिपि आज भी विश्व की सर्वाधिक वैज्ञानिक लिपि मानी जाती है ,लेकिन हिंदीभाषियों के लिए हिंदी में टाइप करना आज भी टेढ़ी खीर है .इसलिए माइक्रोसॉफ्ट ने एक वैकल्पिक कुंजीपटल के रूप में वेबदुनिया की सहायता से एक ऐसी IME (Input Method Editor) का विकास किया गया है ,जिसमें रोमन लिपि की सहायता से ध्वन्यात्मक रूप में हिंदी के पाठ को टाइप किया जा सकता है . वस्तुतः यह कुंजीपटल उन लोगों के लिए अधिक उपयोगी है, जो रोमन लिपि में पहले से ही टाइप करना जानते हैं . इसकी एक झलक निम्नलिखित भाषापट्टी से देखी जा सकती है :

Help																	
m	ma	maa	mi	mee	mu	moo	me	mai	mo	mau	m'	mH	mM	mO	mA	mRa	mra
म्	म	मा	मि	मी	मु	मू	मे	मै	मो	मौ	मं	मः	मैं	मों	मं	मृ	म्र

इसमें 'म' अक्षर से बनने वाले सभी शब्दों को बिना हिंदी टाइप जाने भी रोमन कुंजीपटल की सहायता से हिंदी में टाइप किया जा सकता है . यहाँ तक कि अगर आप 'म्' भी लिखना चाहते हैं तो भी इस पट्टी को देखकर पता लगा सकते हैं कि 'ऋ'की मात्रा कैपिटल R से लिखी जा सकती है .आप जो भी शब्द टाइप करना चाहते हैं ,उसका पहला अक्षर रोमन लिपि में टाइप करें .जैसे आप 'भारत' लिखना चाहते हैं तो आप जैसे ही 'b' टाइप करेंगे तो 'ब' की बारहखड़ी स्क्रीन पर आ जाएगी और आपको स्क्रीन पर 'भ' टाइप करने के लिए 'bh' दिखाई पड़ेगा. फिर 'आ' की मात्रा के लिए aa और 'भारत' लिखने के लिए रोमन लिपि में 'bhaarat' टाइप करें. इसप्रकार यह विधि पूरी तरह से ध्वन्यात्मक है. आप जिस क्रम से बोलते हैं , उसी क्रम से टाइप भी करेंगे . उदाहरण के लिए हिंदी में इ की मात्रा लिखी तो पहले जाती है , लेकिन उसका उच्चारण बाद में होता है . उदाहरण के लिए यदि आपको 'स्त्रियाँ' लिखना है तो आप 'striyaa^' टाइप करेंगे.

हिंदी की यह विशेषता है कि इसमें अल्पप्राण और महाप्राण का युग्म साथ-साथ रहता है. ध्वन्यात्मक लिप्यंतरण (Phonetic Transliteration) नामक इस कुंजीपटल से 'क' लिखने के लिए आप 'k' टाइप करते हैं और 'ख' लिखने के लिए 'kh'.'ग' लिखने के लिए 'g' और 'घ' के लिए 'gh' टाइप करते हैं. इसप्रकार दूसरी या विदेशी भाषा के रूप में हिंदी सीखने वालों को 'h' के माध्यम से महाप्राणत्व का बोध हो जाता है . हिंदी में यह ध्वनि अर्थभेदक है. भारत में भी अहिंदीभाषियों के लिए विशेषकर तमिलभाषियों के लिए यह विधि हिंदी सीखने में काफ़ी मददगार सिद्ध हुई है.

इसके अलावा IME (अर्थात् Input Method Editor) के अंतर्गत टाइपराइटर मोड के रूप में प्रचलित टाइपराइटर के विभिन्न ले आउट भी रखे गए हैं . यद्यपि यह प्रवृत्ति मानकीकरण की दिशा में बहुत उचित नहीं कही जा सकती , लेकिन हिंदी जगत् की जमीनी सचाई को देखते हुए यह समझौता भी कंप्यूटर कंपनियों के लिए आवश्यक हो गया था. यदि मानकीकरण की दृष्टि से देखा जाए तो हिंदी कंप्यूटिंग के लिए इन्स्क्रिप्ट ले आउट या कुंजीपटल भारतीय मानक ब्यूरो द्वारा मानकीकृत कुंजीपटल है और यही स्थिति कमोबेश उर्दू को छोड़कर अन्य भारतीय भाषाओं की भी है , किंतु यदि व्यावहारिक दृष्टि से देखा जाए तो टाइपराइटर का कुंजीपटल हिंदी कंप्यूटिंग के लिए सर्वाधिक लोकप्रिय ले आउट या कुंजीपटल है. इसका प्रमुख कारण यह है कि जो टाइपिस्ट या लेखक सीधे टाइपराइटर से कंप्यूटर के क्षेत्र में आए हैं , उन्हें आज भी हिंदी कंप्यूटिंग के लिए टाइपराइटर कुंजीपटल अधिक सरल लगता है, लेकिन हिंदी टाइपराइटर के कुंजीपटल के क्षेत्र में व्याप्त अराजकता के प्रति भी हम उदासीन नहीं बने रह सकते . माइक्रोसॉफ्ट के ऑफिस हिंदी 2003 के अंतर्गत IME के अंतर्गत आज भी चार प्रकार के टाइपराइटरों के कुंजीपटल उपलब्ध हैं. इसके अलावा हमें यह भी देखना होगा कि हिंदी में टाइपराइटर का कुंजीपटल यांत्रिक सुविधा की दृष्टि से बनाया गया था , जबकि कंप्यूटर के लिए विकसित इन्स्क्रिप्ट कुंजीपटल भारतीय भाषाओं की ध्वन्यात्मक व्यवस्था के आधार पर विकसित किया गया था. प्रयत्नलाघव की दृष्टि से भी यह टाइपराइटर की तुलना में अधिक सटीक और तार्किक है. उदाहरण के लिए हिंदी में टाइपराइटर मोड पर 'औ' को टाइप करने के लिए तीन कुंजियाँ दबानी पड़ती हैं , जबकि हिंदी स्वर की मूल ध्वनि होने के कारण इन्स्क्रिप्ट कुंजीपटल से 'औ' को सिर्फ़ एक ही कुंजी दबाकर अर्थात् एक ही स्ट्रोक में टाइप किया जा सकता है.

वर्तमान परिदृश्य में , इंडिक अर्थात् भारतीय भाषाओं के अधिकांश उपयोगकर्ता सिस्टम और फ्रॉन्ट की असंगतता के कारण आज भी अमानक फ्रॉन्ट का उपयोग कर रहे हैं और ई-मेल, गपशप(चैट), टैम्पलेट, ऑटो टेक्स्ट, थिसॉरस, स्पेलचैक जैसे अनुप्रयोगों का इंडिक भाषाओं में उपयोग करने में हिचकिचाते हैं . यही कारण है कि कंप्यूटर पर

हिंदी के उपयोगकर्ता आज भी शब्दसंसाधन तक ही सीमित हैं . शब्दसंसाधन के अंतर्गत भी वे कंप्यूटर पर हिंदी में टाइप करने मात्र को ही हिंदी कंप्यूटिंग समझने लगते हैं . बहुत ही कम उपयोगकर्ता ऐसे हैं जो हिंदी और अन्य भारतीय भाषाओं में , पावर पॉइंट, एक्सेल और एक्सेस आदि का उपयोग करते हैं. इसका मुख्य कारण अब तक तो यही था कि इंडिक भाषाओं में विभिन्न सिस्टमों के आरपार कोई समान मानक प्रचलित नहीं था. इस दिशा में भारत सरकार द्वारा अनुमोदित भारतीय भाषाओं में कंप्यूटिंग के लिए ISCII कोडिंग प्रणाली एक अच्छी शुरुआत थी, लेकिन विश्वीकरण के इस युग में विविध प्रकार के प्लेटफॉर्म, फ्रॉन्ट और सिस्टम के बावजूद आवश्यकता एक ऐसी मानक कोडिंग प्रणाली की है, जिसके अंतर्गत विश्व की सभी भाषाएँ सह- अस्तित्व की भावना के साथ रह सकें . इन समस्याओं का एकमात्र समाधान है, युनिकोड. इसलिए हमारा प्रयास यह होना चाहिए कि इंडिक भाषाओं के उपयोगकर्ताओं को युनिकोड में भाषा कंप्यूटिंग के लाभों से अवगत कराया जाए . युनिकोड में भारतीय भाषाओं को ISCII के आधार पर ही एन्कोड किया गया है.

कंप्यूटिंग समुदाय के अधिकांश लोगों ने युनिकोड को बहुप्ले टफॉर्म-अंतर्राष्ट्रीयकरण से संबंधित समस्याओं के लिए आदर्श समाधान बताते हुए उसकी प्रशंसा की है . विश्वकर्मा के अधिकांश सॉफ्टवेयर विकासकर्ताओं ने अपने सॉफ्टवेयर को युनिकोड के अनुरूप घोषित कर दिया है . ये हैं: IBM, Microsoft, Oracle, Sybase, Unisys, Apple, Bell Labs, Compaq, GNU/Linux, Sun, SCO, Hewlett Packard, Netscape, Ericsson और Novell.और भी अधिकाधिक अनुप्रयोग युनिकोड के अनुरूप घोषित होते जा रहे हैं . विशेषकर इंटरनेट के प्रचार-प्रसार से अब यह आशा की जा सकती है कि युनिकोड इस बहुभाषी विश्वे में सर्वमान्य मानक के रूप में स्थापित हो जाएगा.

युनिकोड ASCII की सरलता और एकरूपता पर आधारित है , लेकिन यह ASCII की लेटिन वर्णमाला को ही एन्कोड करने की सीमित क्षमता को पार करके उससे कहीं आगे चला जाता है. युनिकोड मानक, विश्व की लिखित भाषाओं के सभी वर्णों को एन्कोड करने की क्षमता रखता है . इसमें 16-बिट की एन्कोडिंग का उपयोग किया गया है और यह 65,000 से अधिक कोड पॉइंट प्रदान करता है . वर्ण कोडिंग को सरल और दक्ष बनाने के लिए युनिकोड मानक प्रत्येक वर्ण को 16-बिट का विशिष्ट मान प्रदान करता है और जटिल मोड या इस्केप कोड का उपयोग नहीं करता.

यद्यपि 65,000 वर्ण, विश्वकी महत्वपूर्ण भाषाओं में प्रयुक्त अधिकांश हजारों वर्णों को एन्कोड करने के लिए पर्याप्त हैं , युनिकोड मानक और ISO 10646, UTF-16 नामक एक ऐसा विस्तार तंत्र प्रदान करते हैं , जिसकी सहायता से इस्केप कोड का उपयोग किए बिना ही एक मिलियन से अधिक वर्णों को एन्कोड किया जा सकता है .यह

विश्व की सभी ऐतिहासिक लिपियों सहित संपूर्ण एन्कोडिंग की वर्ण संबंधी सभी ज्ञात आवश्यकताओं की पूर्ति के लिए पर्याप्त है.

युनिकोड निजी उपयोग के लिए भी कुछ कोड मान सुरक्षित रखता है , जिनका उपयोग सॉफ्टवेयर और हार्डवेयर विकासकर्ता आंतरिक रूप में अपने वर्णों और प्रतीकों के निर्धारण के लिए कर सकते हैं .UTF-8, HTML और इसी प्रकार के प्रोटोकॉल के लिए लोकप्रिय हैं. UTF-8 सभी युनिकोड वर्णों को बाइट की चर लंबाई एन्कोडिंग में रूपांतरित करने का एक माध्यम है. इसके लाभ ये हैं कि युनिकोड वर्ण , चिर-परिचित ASCII सेट के अनुरूप हैं और इनका बाइट- मान भी ASCII के समान है और सॉफ्टवेयर में भारी फेर-बदल के बिना ही UTF-8 में रूपांतरित युनिकोड वर्णों की उपयोग अधिकांश वर्तमान सॉफ्टवेयरों के साथ भी किया जा सकता है.

UTF-16 ऐसे अनेक परिवेशों में लोकप्रिय है , जिनमें स्टोरेज के किफ़ायती उपयोग के साथ वर्णों की दक्षतापूर्वक पहुँच को संतुलित करने की आवश्यकता रहती है . यह समुचित रूप में संक्षिप्त है और भारी से भारी वर्ण भी अकेले 16-बिट के कोड यूनिट में समाहित हो जाते हैं , जबकि अन्य सभी वर्णों तक 16-बिट कोड यूनिट के युग्म के ज़रिए पहुँच होती है. जहाँ नियत चौड़ाई को छोड़कर मेमोरी स्पेस की कोई चिंता नहीं है और वर्ण तक अकेले कोड यूनिट के ज़रिए पहुँच अपेक्षित है , वहाँ UTF-32 काफ़ी लोकप्रिय है. UTF-32 के उपयोग के समय अकेले 32-बिट के कोड यूनिट में प्रत्येक युनिकोड वर्ण को एन्कोड किया जाता है . सभी तीनों एन्कोडिंग फ़ॉर्मों को प्रत्येक वर्ण के लिए अधिक से अधिक 4 बाइट (या 32-बिट) के डेटा की आवश्यकता होती है.

युनिकोड में भारतीय लिपियों को कोड स्पेस U+0900 से U+0D7F तक आबंटित किया गया है. युनिकोड कन्सोर्शियम बड़े कंप्यूटर निगमों ,सॉफ्टवेयर विकासकर्ताओं , डेटाबेस विक्रेताओं,अंतर्राष्ट्रीय एजेन्सियों और विभिन्न उपयोगकर्ता समूहों का एक संगठन है और इसकी स्थापना 1991 में की गई थी .सूचना प्रौद्योगिकी मंत्रालय , भारत सरकार, युनिकोड कन्सोर्शियम का पूर्णकालिक सदस्य है.

युनिकोड कन्सोर्शियम एक अलाभकारी संगठन है और इसकी स्थापना युनिकोड मानक के विकास और प्रचार- प्रसार के लिए की गई थी . युनिकोड मानक, आधुनिक सॉफ्टवेयर उत्पादों और मानकों में पाठ को दर्शाने के लिए निर्धारित किया गया है. कन्सोर्शियम के सदस्य, कंप्यूटर और सूचना प्रोसेसिंग उद्योग के निगमों और संगठनों के व्यापक स्पैक्ट्रम का प्रतिनिधित्व करते हैं.

युनिकोड और अंतर्राष्ट्रीय मानक संगठन (ISO) ने 1992 में युनिकोड मानक संस्करण 1.0 नामक अंतर्राष्ट्रीय मानक कोड निर्धारित और प्रचारित करने के लिए आपस में सहयोग कर लिया था. ISO 10646 की तुलना में युनिकोड संस्करण 1.0 के कुछ वर्णों

के नाम थोड़े-से अलग थे. युनिकोड संस्करण 2.0 में इन नामों को ISO 10646 के समान कर दिया गया. संस्करण 3.0 फ़रवरी 2000 में प्रकाशित किया गया. अब युनिकोड 32-बिट के अधिक व्यापक ISO10646 का 16-बिट फ़ॉर्म है.

युनिकोड कन्सोर्शियम, युनिकोड मानक को कार्यान्वित करने के लिए अनुरूप होने के कारण इन एन्कोडिंग फ़ॉर्मों का पूरी तरह से अनुमोदन करता है. समग्रता में, युनिकोड मानक, संस्करण 3.0 विश्व की सभी वर्णमालाओं, भावचित्रों और प्रतीक संकलनों के 49,194 वर्णों के कोड प्रदान करता है. ये सब कोड पहले कोड स्पेस के क्षेत्र 64K के वर्णों को समाहित कर लेते हैं, जिन्हें संक्षेप में BMP या मूल बहुभाषी प्लेन (basic multilingual plane) कहा जाता है.

हिंदी के व्यापक प्रचार-प्रसार में युनिकोड की सुविधा क्रांतिकारी परिवर्तन ला सकती है. आज विश्व की सभी लिखित भाषाओं के लिए युनिकोड नामक विश्वव्यापी कोड का उपयोग, माइक्रोसॉफ़्ट, आई.बी.एम., लाइनेक्स, ओरेकल जैसी विश्व की लगभग सभी कंप्यूटर कंपनियों द्वारा किया जा रहा है. यह कोडिंग सिस्टम फ़ॉन्ट्समुक्त, प्लेटफ़ॉर्ममुक्त और ब्राउज़रमुक्त है. विंडोज़ 2000 या उससे ऊपर के सभी पी सी युनिकोड को सपोर्ट करते हैं, इसलिए युनिकोड आधारित फ़ॉन्ट का उपयोग करने से न केवल हिंदी को आज विश्व की उन्नत भाषाओं के समकक्ष रखा जा सकता है, बल्कि इसकी सहायता से निर्मित वेबसाइट में खोज आदि अधुनातन सुविधाएँ भी सहजता से ही उपलब्ध हो सकती हैं.

भारत के महामहिम राष्ट्रपति ने 14 सितंबर, 2006 को हिंदी दिवस पर केंद्रीय हिंदी संस्थान द्वारा विज्ञान भवन, नई दिल्ली में आयोजित विशेष समारोह में अपने भाषण में युनिकोड के महत्व पर प्रकाश डालते हुए कहा था: 'विश्वभूके अनेक हिस्सों में हिंदी भाषा आसानी से बोली जा सके इसके लिए इंटरनेट पर हिंदी साहित्य का युनिकोड स्वरूप उपलब्ध करवाना होगा.'

इस समय भारत सरकार का संभवतः एक ही मंत्रालय ऐसा है जिसने इस दिशा में ठोस कदम उठाकर एक मिसाल कायम की है. विदेश मंत्रालय ने युनिकोड के महत्व को समझकर अपनी वेबसाइट <www.mea.gov.in> युनिकोड में निर्मित करके गौरवपूर्ण शुरुआत की है. विश्व भर में फैले भारतवंशी लोगों ने इसका खुले दिल से स्वागत किया है, इसलिए आवश्यकता इस बात की है कि भारत सरकार के सभी सरकारी कार्यालयों और उपक्रमों को आदेश दिए जाएँ कि वे अपनी कंप्यूटर प्रणालियों में युनिकोड आधारित फ़ॉन्ट को डाउनलोड करने की व्यवस्था करें और अपनी वेबसाइट भी युनिकोड आधारित फ़ॉन्ट से निर्मित करें ताकि उनमें ई-मेल, चैट और खोज आदि की सुविधा सहज रूप में उपलब्ध हो सके. हाल ही में प्रभा साक्षी नामक हिंदी समाचार पोर्टल के प्रबंध

संपादक श्री बालेंदु दाधीच के निर्देशन में 13-15 जुलाई, 2007 को न्यूयॉर्क में आयोजित किए जाने वाले 8 वें विश्व हिंदी सम्मेलन के लिए एक विशेष वेबसाइट का निर्माण किया है. www.vishwahindi.com नाम से लॉन्च की गई यह द्विभाषी वेबसाइट पूर्णतः युनिकोड पर आधारित है . इसके अंतर्गत दी गई सूचनाएँ आवश्यकतानुसार स्थितिपरक (Static), गतिशील (dynamic) और अंतःक्रियात्मक (Interactive) हैं. युनिकोड आधारित कोडिंग प्रणाली का सबसे बड़ा लाभ यह है कि कोई भी उपयोगकर्ता हिंदी या विश्व की किसी भी भाषा में निर्मित वेबसाइट की सामग्री को बिना फ्रॉन्ट डाउनलोड किए ही पढ़ सकता है. दैनिक भास्कर, राजस्थान पत्रिका आदि हिंदी के कुछ महत्वपूर्ण समाचारपत्रों और अनुभूति.कॉम जैसी ई-पत्रिकाओं ने अपनी वेबसाइट डाइनेमिक फ्रॉन्ट या वेबफ्रॉन्ट की सहायता से निर्मित की है . ऐसी स्थिति में ये वेबसाइट भी बिना फ्रॉन्ट डाउनलोड किए खुल तो जाती हैं और आप वेब सामग्री हिंदी में पढ़ भी लेते हैं, लेकिन इसे न सो सहेजा जा सकता है और न ही ऑफ लाइन में पढ़ा जा सकता है . और फिर खोज का तो प्रश्न ही नहीं उठता . मेरे विचार में खोज एक ऐसी सुविधा है , जिसके माध्यम से किसी भी मूल शब्द या कीवर्ड को लेकर उपयोगकर्ता वेबसाइट में गोता लगाकर मोती चुन सकता है अर्थात् वाँछित सूचना पा सकता है. यह प्रसन्नता की बात है कि जैसे-जैसे इस दिशा में जागरूकता बढ़ रही है , हिंदी की वेबसाइटों को युनिकोड में परिवर्तित किया जा रहा है. हाल ही में पहले बीबीसी.हिंदी, वॉइस ऑफ अमेरिका, युनिवार्ता, पी.टी.आई (हिंदी), फिर नवभारत टाइम्स और अब msn/india/hindi और दैनिक जागरण की वेबसाइट भी युनिकोड में परिवर्तित हो गई है.

इसके अलावा , वागर्थ, गीता प्रेस, गोरखपुर, विकिपीडिया, अन्यथा जैसी वेबसाइट भी युनिकोड में ही निर्मित की गई हैं . युनिकोड की केवल एक ही सीमा है कि यह विंडोज 98 को सपोर्ट नहीं करता अर्थात् यदि आपके कंप्यूटर पर विंडोज 98 स्थापित है तो आप युनिकोड – समर्थित फ्रॉन्ट को पढ़ नहीं सकते . विंडोज 2000 या उसके ऊपर की कोई भी प्रणाली युनिकोड को सपोर्ट करती है.

एक और भी समस्या है , जिसका समाधान युनिकोड आधारित फ्रॉन्ट का उपयोग शुरू करने से पहले करना होगा . अब तक हिंदी में टू-टाइप फ्रॉन्ट का ही बोल-बाला था . टू-टाइप फ्रॉन्ट 8 बिट या 7 बिट प्रणाली पर आधारित होता है और अगर आप युनिकोड (जो 16 बिट प्रणाली पर आधारित है और जो ओपन टाइप फ्रॉन्ट होता है) पर आधारित फ्रॉन्ट की सहायता से अपनी वेबसाइट का निर्माण करना चाहते हैं तो टू-टाइप फ्रॉन्ट में पहले ही निर्मित सामग्री को युनिकोड में परिवर्तित करने के लिए कन्वर्टर का मदद लेनी होगी. इस कार्य के लिए भारत सरकार द्वारा विकसित 'परिवर्तन' नामक कन्वर्टर और माइक्रोसॉफ्ट द्वारा विकसित 'टीबीआईएल' नामक कन्वर्टर काफ़ी उपयोगी सिद्ध हो सकता

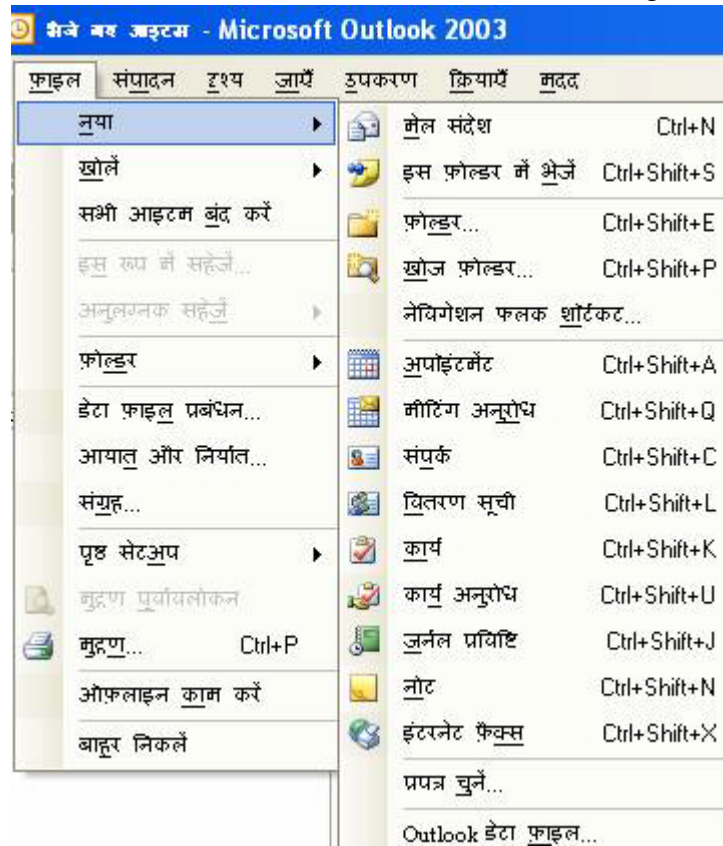
है.इसके अलावा भारत सरकार ने सी डैक के सहयोग से युनिकोड –आधारित ओपन टाइप में अनेक युनिकोड-आधारित फॉन्ट जनता के लिए निःशुल्क उपलब्ध करा दिए हैं.

जैसा कि ऊपर स्पष्ट किया गया है कि कंप्यूटर के क्षेत्र में अब तक हिंदी का प्रयोग फ्रॉन्ट स्तर पर ही किया जाता था और इसका उपयोग भी शब्दसंसाधन तक ही सीमित था,लेकिन ऑफिस हिंदी के आगमन से हिंदी सच्चे अर्थों में कंप्यूटर के क्षेत्र में विश्व की अन्य विकसित भाषाओं के समकक्ष आ गई है .वस्तुतः माइक्रोसॉफ्ट ने पहली बार हिंदी को ऑपरेटिंग सिस्टम में समाहित किया है,जिसके फलस्वरूप कंप्यूटर संबंधी सभी अनुप्रयोगों में हिंदी का प्रयोग सहजता और सरलता से किया जा सकता है . कंप्यूटर में अंग्रेजी के व्यापक उपयोग के कारण आम लोगों में यह धारणा बन गई थी कि अंग्रेजी ही कंप्यूटर की मूल भाषा है .इस धारणा की पुष्टि इस कारण से भी होने लगी थी किसी भी सॉफ्टवेयर पर फ़ाइल का नाम केवल अंग्रेजी में ही रखा जा सकता था .अकारादि क्रम से शब्दों की अनुक्रमणिका तक नहीं बनाई जा सकती थी और यही कारण है कि हिंदी में कोश निर्माण जैसे श्रमसाध्य और समयसाध्य कार्य के लिए भी कंप्यूटर का उपयोग मात्र टाइपिंग के लिए ही किया जा सकता था.

आइए अब हम पूर्णतः यूनिकोड-आधारित ऑफिस हिंदी नामक एक अत्यंत लोकप्रिय सिस्टम की विशेषताओं की झलक पाने से पहले उसकी पृष्ठभूमि पर एक नज़र डालें . विंडोज ऑपरेटिंग सिस्टम का निर्माण और विकास माइक्रोसॉफ्ट कॉर्पोरेशन ने किया है और इसके अंतर्गत एम . एस.ऑफिस के विभिन्न अनुप्रयोगों का विकास भी माइक्रोसॉफ्ट के तत्वावधान में ही किया गया है,किंतु आरंभ में विश्वअकी कुछेक भाषाओं को ही इसमें स्थान दिया गया था . कदाचित् इसका कारण यह भी था कि भारत में हिंदी को संवैधानिक मान्यता मिलने के बावजूद हिंदी को उचित स्थान प्रदान नहीं किया गया था और अंग्रेजी का वर्चस्व इस कदर बढ़ गया था कि विदेश में रहने वाले लोगों को ऐसा लगने लगा था कि भारत की प्रमुख राजभाषा भी कदाचित् अंग्रेजी ही है ,लेकिन यह एहसास होने पर कि विश्वर में तीसरे स्थान पर बोली जाने वाली भाषा आज भी हिंदी है और इसकी और अधिक उपेक्षा से माइक्रोसॉफ्ट के व्यापारिक हितों को भी नुकसान पहुँच सकता है ,माइक्रोसॉफ्ट ने वर्ष 1998 में वर्ड 2000 के दक्षिण पूर्वशिया संस्करण में हिंदी को सीमित स्थान देकर इसकी शुरुआत की . उस समय तक भारत की अनेक कंपनियों ने हिंदी और अनेक भारतीय भाषाओं में विभिन्न प्रकार के फ़ॉन्ट बनाने का काम शुरु कर दिया था.यही कारण है कि आरंभ में माइक्रोसॉफ्ट की इस शुरुआत पर लोगों ने इतना ध्यान नहीं दिया,किंतु कुछ समय के बाद ऑफिस एक्स पी के लोकार्पण के बाद कंप्यूटर जगत् में एक हलचल सी मच गई . ऑफिस एक्स पी के माध्यम से पहली बार हिंदी को

ऑपरेटिंग सिस्टम में समाहित किया गया था और उपयोगकर्ता सरलता और सहजता के साथ कंप्यूटर पर अपने सभी कार्य हिंदी में संपन्न कर सकते थे. यहाँ तक कि अपनी फ़ाइलों के नाम भी हिंदी में लिख सकते थे ,किंतु ये सभी कार्य विंडोज़ 2000 और उससे ऊपर के सिस्टम में ही संभव थे. इसका प्रमुख कारण यह था कि माइक्रोसॉफ़्ट ने यूनिकोड के विश्वव्यापी मानक को अपनाया था ,जिसका न्यूनतम आधार विंडोज़ 2000 है. इसके विपरीत बाज़ार में उपलब्ध अधिकाँश हिंदी फ़ॉन्ट विंडोज़ 1998 पर भी चल सकते थे.

इस पृष्ठभूमि के बाद , आइए अब हम सबसे पहले ऑफ़िस हिंदी की उन विशेषताओं की झलक देखें,जिनके कारण यह ऑफ़िस हिंदी के नाम से लोकप्रिय होने लगा है. वस्तुतः माइक्रोसॉफ़्ट ने अब ऑफ़िस हिंदी के साथ- साथ विंडोज़ हिंदी का भी लोकार्पण कर दिया है. इसकी सबसे महत्वपूर्ण परिणति यह हुई है कि अब विंडोज़ हिंदी का संपूर्ण इंटरफ़ेस ही हिंदी में रूपांतरित हो गया है अर्थात् इसका मेनू,सबमेनू,सहायता आदि भी हिंदी में हो गए हैं.इससे अंग्रेज़ी न जानने वाले उपयोगकर्ताओं के लिए काफ़ी सुविधा हो गई है.



इसी प्रकार अब फ़ाइलों के नाम भी हिंदी में रखे जा सकते हैं :



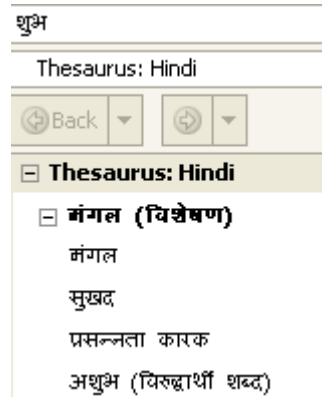
हिंदी और अंग्रेजी में तथा हिंदी के विभिन्न कुंजीपटलों के बीच स्विच करने के लिए एक भाषा पट्टिका दी गई है.



इस पट्टिका की सहायता से उपयोगकर्ता हिंदी के विभिन्न कुंजीपटलों में से अपनी पसंद के कुंजीपटल का चुनाव कर सकते हैं.

इनमें प्रमुख कुंजीपटल हैं : रेमिंगटन, इन्स्क्रिप्ट और ध्वन्यात्मक लिप्यंतरण कुंजीपटल . रेमिंगटन कुंजीपटल उन उपयोगकर्ताओं के लिए सहज है जो टाइपराइटर पर काम करना के अभ्यस्त रहे हैं . इन्स्क्रिप्ट कुंजीपटल कंप्यूटर का तर्कसंगत कुंजीपटल है ,जिसे देवनागरी के वर्णक्रम के अनुरूप बनाया गया है. मूलतः : कंप्यूटर पर हिंदी में काम करने वाले उपयोगकर्ताओं के लिए यह बहुत उपयुक्त है. ध्वन्यात्मक लिप्यंतरण कुंजीपटल रोमन लिपि के माध्यम से काम करने वाले उपयोगकर्ताओं के लिए आदर्श है .इसमें हिंदी की बारहखड़ी की पद्धति का पूरा उपयोग किया गया है.

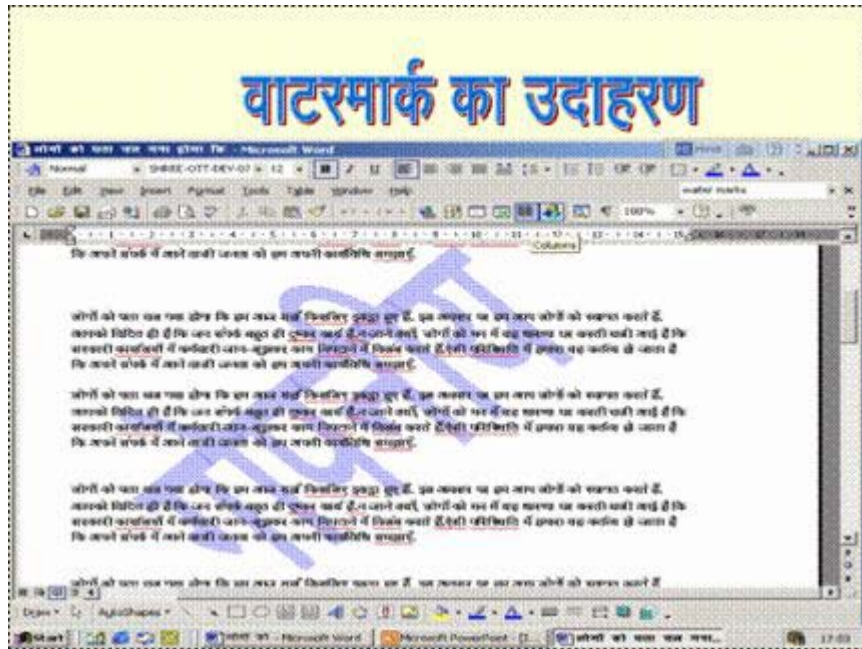
ऑटो करेक्ट ऑफिस हिंदी की अन्यतम विशेषता है. ऑटो करेक्ट और स्पेल चेकर में मुख्य अंतर यही है कि स्पेल चेकर केवल अशुद्धियों को रेखांकित करता है और ऑटो करेक्ट उन्हें ठीक भी कर देता है. हिंदी में थिसॉरस का प्रवेश भी पहली बार ऑफिस हिंदी में ही किया गया है.दाहिने क्लिक करके आप किसी भी शब्द के पर्याय ,विलोम और संबद्ध शब्दों को देख सकते हैं.



हिंदी में अकारादि क्रम से अनु क्रमणिका तैयार करने का कार्य सॉर्टिंग के माध्यम से वर्ड,एक्सेस और एक्सेल के माध्यम से सहजता से किया जा सकता है. इसका उपयोग कोश निर्माण ,पुस्तकालय और वरीयता सूची आदि के लिए किया जा सकता है.



वाटर मार्क या जलचिह्न का उपयोग करेंसी नोट या गोपनीय दस्तावेजों में किया जाता है.



इससे यह स्पष्ट है कि युनिकोड के आगमन से और ऑफिस हिंदी जैसे पैकेजों के लोकार्पण से आज कंप्यूटर के क्षेत्र में हिंदी विश्व की सभी विकसित भाषाओं के समकक्ष आ गई है.

