

ইউনিকোড এবং বাংলা কম্পিউটিং

■ মোঃ মারুফ হোসেন

সহ-সম্পাদক

মাসিক কম্পিউটার টুমরো

বাংলা কম্পিউটিং বলতে কী বোঝায় ?

অনেকেই কম্পিউটিং, বিশেষত বাংলা কম্পিউটিং বলতে বাংলা টাইপিংকেই বোঝেন। কিন্তু এটি আসলে বাংলা কম্পিউটিংয়ের অনেকগুলো ব্যবহারের মাত্র একটি। বাংলা কম্পিউটিং হলো কম্পিউটার ব্যবহার ও পরিচালনার প্রতিটি স্তরে যাবতীয় ইনপুট, আউটপুট হিসেবে বাংলার ব্যবহার, অর্থাৎ সবকিছুই বাংলায় সাধিত হবে। যেখানে কম্পিউটার চালু করার পর ইন্টারফেসের সব কথা বাংলায় থাকবে, কমান্ড বাংলায় দেয়া হবে, পাসওয়ার্ড দেয়া, ডাটাবেজে ডাটা দেয়া, এমনকি ডাটাবেজে ডাটা সংরক্ষণ, সার্চিং, সার্চিং সবই বাংলায় হবে—

সেটাই বাংলা কম্পিউটিং।

আর তাই বাংলা কম্পিউটিংয়ে বাংলা কী-বোর্ডের চেয়েও বেশি গুরুত্বপূর্ণ প্রশ্ন হলো বাংলা ক্যারেঙ্কার সেট। কেননা, কী-বোর্ডই কম্পিউটারে ইনপুট দেয়ার একমাত্র পদ্ধতি নয়— কম্পিউটারে বর্ণ ইনপুট দেয়ার জন্য অপটিক্যাল ক্যারেঙ্কার রিডার (ওসিআর), স্পিচ রিকগনিশন ইত্যাদি রয়েছে। কিন্তু কম্পিউটারে সরাসরি বাংলাকে প্রসেস করতে চাইলে একটি সর্বজন স্বীকৃত স্ট্যান্ডার্ড এবং সর্বোপরি বিশ্ব স্বীকৃত এনকোডিং সিস্টেম বা বর্ণ সংকেতায়ন ব্যবস্থা থাকতে হবে। আর বর্তমান বিশ্বে সবচে' বেশি স্বীকৃত ব্যবস্থাটি হলো ইউনিকোড।

ইউনিকোড কী ?

আমরা একটা বিষয় জানি যে, কম্পিউটার মূলত সংখ্যা নিয়ে কাজ করে। এমনকি বর্ণ ও অন্যান্য চিহ্নের জন্যও কম্পিউটার একটি সংখ্যামান প্রদান করে। ইউনিকোডের আগে এরকম হাজারটা ক্যারেঙ্কার এনকোডিং স্ট্যান্ডার্ড ছিল— যার একেকটির মান ছিল একেক রকম। তাছাড়া স্ট্যান্ডার্ডগুলোর আকারও এত বড় ছিল না যে, একটি একক স্ট্যান্ডার্ডে সব বর্ণ যুক্ত করা সম্ভবপর হবে। যেমন— শুধু ইউরোপিয়ান ইউনিয়নের বিভিন্ন ভাষার জন্যই দরকার হতো একাধিক স্ট্যান্ডার্ড। এমনকি ইংরেজির জন্যও একটি একক এনকোডিংয়ে সমস্ত বর্ণ, যতি চিহ্ন ও সাধারণ কাজে ব্যবহারের বিশেষ চিহ্নগুলো সংরক্ষণ করা সম্ভব ছিল না।

এই এনকোডিং সিস্টেমগুলো একটি অন্যটির সাথে বিভিন্ন ধরনের সমস্যার সৃষ্টি করে। তাছাড়া একই বর্ণ একাধিক এনকোডিং স্ট্যান্ডার্ডে একেকটি মানের হতে পারে। কিংবা একটি মানের জন্য একাধিক কম্পিউটার স্ট্যান্ডার্ডে একাধিক বর্ণও থাকতে পারে। ফলে একটি কম্পিউটার (বিশেষত সার্ভার)—কে সব ধরনের এনকোডিং স্ট্যান্ডার্ড সাপোর্ট দিতে হলে সেই কম্পিউটারের জন্য তা খুবই সমস্যার হতে পারে। আর তার ফলস্বরূপ ডাটা করাষ্ট করার ঝুঁকিও অনেক বেড়ে যায়।

অন্যদিকে, ইউনিকোড হলো একটি একক স্ট্যান্ডার্ড সিস্টেম, যা প্র্যাটফর্ম, প্রোগ্রাম ও ভাষার জটিলতা মুক্ত। আর বর্তমানে ইউনিকোড ব্যবহার করছে আইটি অঙ্গনের সমস্ত রথী-মহারথীরাই। তাই ইউনিকোড ভিত্তিক সফটওয়্যার নির্মাণ করছে এপল, এইচপি, আইবিএম, মাইক্রোসফট, ওরাকল, স্যাপ, সনি, সিবজ, ইউনিসিস সহ সমস্ত নামি-দামি নির্মাতারা। শুধু তাই নয়, বর্তমান সময়ের অন্যান্য স্ট্যান্ডার্ড টেকনোলজি XML, Java, ECMAScript (JavaScript), LDAP, CORBA3.0 কিংবা WML-এও এটি ব্যবহৃত হচ্ছে। আর ISO/IEC 10646-এর অফিসিয়াল ব্যবহার পদ্ধতিই হলো ইউনিকোড। তাছাড়া বর্তমান সময়ের সমস্ত অপারেটিং সিস্টেম, ব্রাউজার সহ অন্যান্য সফটওয়্যারেও এটি ব্যবহৃত হচ্ছে। আর দিন দিন ইউনিকোড স্ট্যান্ডার্ডের ব্যবহার বৃদ্ধি পাওয়া এটাই প্রমাণ করে যে, ভবিষ্যতের একক স্ট্যান্ডার্ড হিসেবেই টিকে থাকবে ইউনিকোড। তাছাড়া একাধিক ক্যারেঙ্কার সেটের পরিবর্তে একটি একক স্ট্যান্ডার্ড ক্যারেঙ্কার সেট থাকার ফলে ক্লায়েন্ট সার্ভার কিংবা n-Tier এপ্লিকেশন ও ওয়েবসাইটের খরচ অনেক কমে যাবে। এবং ডাটা করাষ্টের সমস্যা ছাড়াই এটি ডাটার আদান-প্রদানে সমর্থ হবে।

আর সে জন্যই ইউনিকোড হচ্ছে বিশ্বজুড়ে স্বীকৃত একটি আন্তর্জাতিক বর্ণ সংকেতায়ন ব্যবস্থা (International Character Encoding Standard)। বিশ্বের প্রায় প্রতিটি ভাষার জন্যই ইউনিকোডের রয়েছে একটি নির্দিষ্ট মানদণ্ড। মূলত ইউনিকোডের কল্যাণেই আজ বিশ্বের প্রতিটি ভাষাকে একটি একক মানদণ্ডে নিয়ে আসা সম্ভব হয়েছে। তবে অনেকেই মনে করেন, ইউনিকোড হচ্ছে— বিভিন্ন ভাষার অসংখ্য হরফের একটি বিশাল তালিকা— যা ANSI-র একটি বর্ধিত সংস্করণ। কার্যত এ ধারণাটি পুরোপুরি সত্য নয়। যদিও ইউনিকোড ANSI-র বর্ধিত সংস্করণ, তারপরও এটি শুধুই একটি তালিকা নয়। ANSI-র চেয়ে এটি অনেক বেশি সুশৃঙ্খল ও উন্নত। যে-কোনো ভাষা/লিপি'র বর্ণমালাকে ইউনিকোড-এ অন্তর্ভুক্তির সময় শুধু বর্ণমালাটিকেই অন্তর্ভুক্ত করা হয় না, বরং সংশ্লিষ্ট ভাষার বর্ণ সমূহের প্রকৃতি, আচরণ, মাত্রার ব্যবহার, ব্যাকরণগত প্রয়োগ এবং শ্রেণীবিন্যাস (সিটিং)-এর বিষয়টিকেও বিশেষভাবে বিবেচনা করা হয়। আর বিশেষ বিবেচনার কারণেই এক একটি ভাষার লিপি ইউনিকোডে অন্তর্ভুক্ত হতে ৬ মাস থেকে কয়েক বছর পর্যন্ত সময় লাগে।

ইউনিকোডের সংক্ষিপ্ত ইতিহাস

ইউনিকোড স্ট্যান্ডার্ড নিয়ে প্রথম কাজ করে ক্যালিফোর্নিয়া ভিত্তিক ইউনিকোড কনসোর্টিয়াম। তারা ১৯৯১ সালে ইউনিকোডের প্রথম সংস্করণ প্রকাশ করে। পরবর্তীতে তারা ISO/IEC 10646 Universal Multiple-Octet Coded Character Set (UCS) এর সাথে মিলিতভাবে কাজ করে ১৯৯৩ সালে ইউনিকোড ১.১ প্রকাশ করে। এরপর ১৯৯৬ সালে ইউনিকোড ২.০ প্রকাশিত হয়, যাতে ব্রাহ্মীলিপির অন্তর্ভুক্ত বিভিন্ন লিপির সাথে বাংলাও যুক্ত হয়। এরপর ১৯৯৮ সালে ইউনিকোড ২.১ এবং ১৯৯৯ সালে ইউনিকোড ৩.০ প্রকাশিত হয়। এই সবগুলো সংস্করণেই বর্ণ ছিল BMP অংশে। আর গত মাসে প্রকাশিত হয়েছে ইউনিকোডের সর্বশেষ সংস্করণ ইউনিকোড ৪.০।

ইউনিকোড স্ট্যান্ডার্ড

বিশ্বব্যাপী সকল ভাষার সকল বর্ণমালাকে একটিমাত্র ক্যারেক্টার কোডিংয়ের আওতায় নিয়ে যে ক্যারেক্টার কোডিং স্ট্যান্ডার্ড তৈরি করা হয়েছে সেটিই হলো ইউনিকোড স্ট্যান্ডার্ড। ইউনিকোড শুধুমাত্র বর্ণটিকেই সংরক্ষণ করে না, বরং তার প্রেসেটিং ও প্রদর্শন পদ্ধতিও সংরক্ষণ করে। তাই বর্তমান সময়ের সমস্ত আধুনিক ভাষার লিখিত রূপই এটি সাপোর্ট করে। পাশাপাশি এটি অনেক প্রাচীন ও ঐতিহাসিক ভাষাই সাপোর্ট করে।

ইউনিকোড স্ট্যান্ডার্ডের একাধিক সংস্করণ রয়েছে। আর এর প্রতিটি সংস্করণ একটির সাথে অন্যটি কম্প্যাটবিল। শুধু ইউনিকোডের পারস্পরিক সংস্করণই নয়, এটি আন্তর্জাতিক স্ট্যান্ডার্ড ISO/IEC 10646-এর সাথেও সমঝোতা করতে পারে। উদাহরণস্বরূপ বলা যায় ইউনিকোড ৩.০ একই বর্ণ ও এনকোডিং পয়েন্ট ব্যবহার করে ISO/IEC 10646-1: 2000-এর। একইসাথে ইউনিকোড ৩.১ ISO/IEC 10646-2:2001-এর বর্ণ ও এনকোডিং পয়েন্ট ব্যবহার করে। তবে, এতে শুধু বর্ণ ও এনকোডিং পয়েন্টই থাকে না, বরং এতে বর্ণ সম্পর্কিত অনেক অতিরিক্ত তথ্য থাকে। ফলে বর্ণটি কিভাবে সংরক্ষিত হবে এবং ক্রমে কিভাবে প্রদর্শিত হবে তাও সংরক্ষিত হয় ইউনিকোড স্ট্যান্ডার্ডে। তাছাড়া ইউনিকোড স্ট্যান্ডার্ড মেনে চলা যে-কোনো প্রোগ্রাম ISO/IEC 10646-এ চলবে কোনো রকম বাড়াতি প্রোটোকল ছাড়া।

ইউনিকোড ব্যবহার করার ফলে বহুভাষা ব্যবহার করে তৈরি করা সাধারণ টেক্সটফাইল তৈরি করা ও বহন করা সহজ হয়ে যাচ্ছে। বিশেষ করে যেখানে বহুভাষা ব্যবহার করে কাজ করা অত্যন্ত জরুরি— সেখানে এই স্ট্যান্ডার্ড অত্যন্ত জরুরি। ফলে ব্যবসায়ী, ভাষাবিদ, গবেষক, বিজ্ঞানী এমনকি টেকনিক্যাল লেখক যাদের একাধিক ভাষার বর্ণ ব্যবহার করে কাজ করতে হয় তাদের জন্য অত্যন্ত উপযোগী। তাছাড়া গণিতবিদ ও টেকনিশিয়ান, যাদের অসংখ্য গাণিতিক চিহ্ন নিয়ে কাজ করতে হয়, তারাও ইউনিকোডকে দারুণ কার্যকরী হিসেবে চিহ্নিত করবেন।

ইউনিকোড স্ট্যান্ডার্ডটি তৈরি ভিত্তি হিসেবে এসকি ব্যবহার করা হয়েছে। কেননা, এসকি সহজ ও ধারাবাহিক। কিন্তু ইউনিকোড এসকি'র সীমাবদ্ধতাকে ছাড়িয়ে গেছে প্রায় সবক্ষেত্রেই। এসকি যেখানে শুধুমাত্র ল্যাটিন বর্ণমালা নিয়ে কাজ করত, ইউনিকোড সেখানে বিশ্বের সব ধরনের ভাষা নিয়েই কাজ করছে। ক্যারেক্টার কোডিং সহজ ও কার্যকরী রাখার জন্য ইউনিকোড স্ট্যান্ডার্ড প্রতিটি বর্ণকে একটি একক সাংখ্যমান এবং একটি নাম প্রদান করে।

তবে ইউনিকোড স্ট্যান্ডার্ডের মূল উদ্দেশ্য ছিল একটি ১৬ বিটের এনকোডিংয়ের মাধ্যমে ৬৫ হাজারেরও বেশি বর্ণকে কোড পয়েন্ট দেয়া। কেননা, এই ৬৫ হাজার বর্ণের মাধ্যমেই বিশ্বের প্রধান প্রধান ভাষার বর্ণমালা সমূহকে এনকোডিং কোড পয়েন্ট দেয়া সম্ভব। বর্তমানে ইউনিকোড স্ট্যান্ডার্ড ও ISO/IEC 10646 তিন ধরনের এনকোডিং ফর্ম সাপোর্ট করছে যার মধ্যে বর্ণমালার মাত্র একটি সাধারণ ঠাণ্ডার থাকলেও সেটি কয়েক লাখ বর্ণ তৈরি করতে পারে। আর বর্তমানের এই পদ্ধতিই প্রচলিত সব ক্যারেক্টার এনকোডিংয়ের জন্য যথেষ্ট বলে দাবি করে ইউনিকোড কনসোর্টিয়াম। কেননা, এই পদ্ধতিতেই বিশ্বের সব ঐতিহাসিক লিপিকে সাপোর্ট দেয়া সম্ভব।

ইউনিকোড স্ট্যান্ডার্ডে যে ধরনের বর্ণ থাকতে পারে

বর্তমান বিশ্বের যতগুলোর ভাষার লিখিত রূপ রয়েছে তার সবগুলোই সংযুক্ত করা হয়েছে ইউনিকোডে। এসব লিপির মধ্যে ইউরোপিয়ান লিপি, মধ্যপ্রাচ্যের ডান থেকে বামে লেখার লিপিও এশিয়ার অনেকগুলো লিপি রয়েছে।

ইউনিকোড স্ট্যান্ডার্ডে শুধু বর্ণমালাই নয়, সব লিপিরই নিজস্ব যেসব বিরাম চিহ্ন, বৈশিষ্ট্য সূচক চিহ্ন, গাণিতিক চিহ্ন, টেকনিক্যাল চিহ্ন, তীর চিহ্ন ও ডিগ্ব্যাটস আছে তা অন্তর্ভুক্ত করা হয়েছে। আর সব ধরনের চিহ্নের জন্যই এতে থাকছে একটি করে কোড পয়েন্ট। আবার অনেক সময় বিভিন্ন বৈশিষ্ট্য সূচক চিহ্নের মাধ্যমে, যেমন টিন্ট (-) অন্যান্য মূল বর্ণ মিলিয়ে পরিবর্তিত বর্ণ তৈরির পদ্ধতিও সংরক্ষণ করে ইউনিকোড স্ট্যান্ডার্ড। যেমন ~ ও n একত্রিত করে তৈরি হচ্ছে ñ।

ইউনিকোড স্ট্যান্ডার্ড ৩.২ বর্তমানে বিশ্বের বহু প্রধান ভাষার ৯৫,২২১টি বর্ণ, ইডিওগ্রাফ ও চিহ্নের সংগ্রহশালাকে সাপোর্ট দিচ্ছে এবং এনকোডেড কোড পয়েন্ট প্রদান করছে।

সাধারণত ইউনিকোডের প্রথম ৬৪ কিলোবাইট কোড পয়েন্টে অধিকাংশ ভাষার ব্যবহৃত হয় এমন বহুল প্রচলিত বর্ণগুলোকে রাখা হয়েছে। এই ৬৪ কিলোবাইট কোড স্পেসকে বলা হয় Basic Multilingual Plane (BMP)। বর্তমানে এই BMP অংশে প্রায় ৬,৭০০ ফাঁকা কোড পয়েন্ট রয়েছে ভবিষ্যতে ব্যবহারের জন্য। অন্যান্য অংশেও প্রায় ৮ লাখ ৭০ হাজার ফাঁকা কোড পয়েন্ট রয়েছে। তাছাড়া ভবিষ্যতের সংস্করণগুলোতে অন্তর্ভুক্তির জন্য প্রচুর বর্ণ বিবেচনামূলক রয়েছে।

এছাড়া ব্যক্তিগতভাবে ব্যবহারের জন্যও ইউনিকোড স্ট্যান্ডার্ডে বেশকিছু সংরক্ষিত কোড পয়েন্ট রয়েছে। এই সংরক্ষিত কোড পয়েন্টগুলো বিভিন্ন ভেভর ও সাধারণ ব্যবহারকারীরা তাদের নিজস্ব বর্ণ কিংবা প্রতীকের জন্য ব্যবহার করতে পারেন। তাছাড়া বিশেষায়িত ফন্টের জন্যও এসব সংরক্ষিত কোড পয়েন্ট রয়েছে, পাশাপাশি অন্যান্য অংশেও ১,৩১,০৬৮ টি সংরক্ষিত কোড পয়েন্ট রয়েছে যদি BMP অংশের সংরক্ষিত কোড পয়েন্ট কম বলে মনে হয়।

ইউনিকোডের এনকোডিং ফরম্যাট

ইউনিকোডে ক্যারেক্টার এনকোডিং স্ট্যান্ডার্ড শুধু প্রতিটি বর্ণের আইডেন্টিটি ও সংখ্যামান বা কোড পয়েন্টই নির্ধারণ করে না, বরং এই মানটি বিট হিসেবে কিভাবে প্রকাশিত হবে তাও নির্ধারণ করে ইউনিকোড স্ট্যান্ডার্ড। ইউনিকোড স্ট্যান্ডার্ড তিন ধরনের এনকোডিং ব্যবহার করে একই ডাটা কখনো byte, কখনো word আবার কখনো double word হিসেবে সংরক্ষণ করে। অর্থাৎ প্রতি কোড ৮, ১৬ কিংবা ৩২ বিট ব্যবহার করে সংরক্ষণ করা হয়। তবে তিন ধরনের এনকোডিংই একই বর্ণভাণ্ডার ব্যবহার করে বলে কার্যকরভাবে কোনোরকম ডাটা না হারিয়েই এক এনকোডিং থেকে অন্য এনকোডিং-এ রূপান্তর করা যায়। এই ফরম্যাটটিকে বলা হয় Unicode Transformation Format (UTF)।

UTF-8 মূলত HTML ও ঐ ধরনের প্রোটোকলেই বেশি ব্যবহার হয়। এটি সব ধরনের ইউনিকোড বর্ণকে যে-কোনো দৈর্ঘ্যের byte ডাটায় রূপান্তর করতে পারে। আর এই ফরম্যাটের আরেকটি বড় সুবিধা হলো এই যে, এতে যেসব বর্ণ এসকি ক্যারেক্টার সেটে আছে তাদের কোড পয়েন্ট ও byte মান একই। ফলে ইউনিকোড ক্যারেক্টার যদি UTF-8 এ রূপান্তর করা হয় তবে এসকি'র জন্য তৈরি সফটওয়্যার কোনো ধরনের ঝামেলা ছাড়াই ইউনিকোড-এ কাজ করবে।

UTF-16 কম মেমোরি স্পেসের জন্য খুবই কার্যকরী। কেননা, এটি বেশ কমপ্যাক্ট এবং বহুল ব্যবহৃত বর্ণগুলো এতে একটি ১৬ বিটের কোড ব্যবহার করা হয়।

UTF-32 সেখানেই বেশি ব্যবহৃত হয় মেমোরি যেখানে কোনো সমস্যাই নয়। এতে একটি নির্দিষ্ট আকারের একক কোড পয়েন্টের মাধ্যমে সব বর্ণ ব্যবহার করা যায়। আর প্রতিটি ইউনিকোড ক্যারেক্টার এমনিতেই একটি একক ১৬ বিটের কোড ব্যবহার করে UTF-32 ফরম্যাটে সংরক্ষিত হয়।

আর সবগুলো ফরম্যাটই সর্বোচ্চ ৪ বাইট (বা ৩২ বিট)-এর ডাটা ব্যবহার করতে পারে একটি বর্ণের জন্য।

টেক্সটের উপাদান নির্ধারণ

প্রতিটি ভাষার লিখিতরূপই কতগুলো লেখনী চিহ্ন দ্বারা সূচিত হয়— যার ব্যবহারে শব্দ ও বাক্য তৈরি হয়। এইসব লেখনী চিহ্নের মধ্যে থাকতে পারে বর্ণ ('ক', 'a'), অথবা জাপানি হিরাগানার মতো অক্ষর (syllables) কিংবা চীনাাদের মতো সরাসরি শব্দ জ্ঞাপক প্রতীক (ideograph)।

তাই টেক্সটের উপাদান টেক্সটে কিভাবে ব্যবহৃত হবে, তার উপর নির্ভর করে নির্ধারণ করা হয়। যেমন— ঐতিহাসিক স্প্যানিশ ভাষায় 'll'-কে একটি একক শব্দ হিসেবে বিন্যস্ত করা হয় অথচ লেখার সময় সেটি দুটি 'l' দ্বারা লেখা হয়। অথবা বাংলায় 'কেকা' শব্দটি ডিকশনারির 'ক' অনুচ্ছেদে সংরক্ষণ করা হয়— যদিও 'কেকা' লিখতে 'c' আগে ব্যবহৃত হয়।

এ কারণে কোনটি টেক্সটের উপাদান আর কোনটি নয়— তা বিভিন্ন সময়ে আলাদাভাবে বুঝে নিতে হয় আমাদের। আর এই বিভ্রান্তি এড়াতে ইউনিকোড স্ট্যান্ডার্ডে 'কোড এলিমেন্ট' (যাকে ক্ষেত্রবিশেষে বর্ণ বলা চলে)-কে টেক্সটের উপাদান নির্ধারণ করা হয়। কম্পিউটারে টেক্সট প্রেসেটিং করতে কোড এলিমেন্টই হলো সবচে' মৌলিক ও গুরুত্বপূর্ণ অংশ। অন্যদিক থেকে ভিত্তি করলে কোড এলিমেন্টই হলো বহুল ব্যবহৃত টেক্সট এলিমেন্ট। স্প্যানিশ 'll'-এর ক্ষেত্রে ইউনিকোড স্ট্যান্ডার্ড প্রতিটি 'l'-কে আলাদা আলাদা কোড এলিমেন্ট দিচ্ছে। আর দুটি 'l'-কে একটি ধরে বিন্যাসের দায়িত্ব, যে সফটওয়্যার টেক্সট প্রেসেটিং করবে তার উপরই ছেড়ে দিয়েছে। আবার বাংলায় 'কেকা' লিখতে উচ্চারণ ভিত্তিক বা ফোনেটিক ব্যবহারকেই গুরুত্ব দিয়ে টাইপ করার জন্য ক-কে-ক টাইপ করতে বলছে এবং সেভাবেই কোড পয়েন্ট সংরক্ষণ করছে। কিন্তু দেখানোর সময় ঠিকই 'কেকা' দেখাচ্ছে। ফলে শব্দ বিন্যাসের ক্ষেত্রে বাংলা ব্যাকরণ সিদ্ধ হচ্ছে।

টেক্সট প্রেসেটিং

কম্পিউটার টেক্সটকে দুটি স্তরে ব্যবহার করে প্রেসেটিং এবং এনকোডিং। যেমন— ধরা যাক, একজন টাইপিষ্ট কী-বোর্ডের সাহায্যে টেক্সট টাইপ করছে। এক্ষেত্রে কম্পিউটারের সিস্টেম সফটওয়্যার প্রতিটি কী-প্রেসের জন্য একটি করে কোড পাচ্ছে। ধরি, টাইপিষ্টটি 'T' টাইপ করল। তাহলে 'T'-এর জন্য U+0054 কোডটি যাচ্ছে কম্পিউটারের কাছে। ওয়ার্ড প্রেসের সফটওয়্যারটি সেটি মেমোরিতে সংরক্ষণ করছে এবং ক্রমে দেখানোর জন্য ডিসপ্লে সফটওয়্যারের কাছে কোডটি পাঠিয়ে দিল। এভাবেই ডিসপ্লে সফটওয়্যার ঐ কোডটির জন্য নির্দিষ্ট বর্ণটি প্রদর্শন করবে। এক্ষেত্রে ডিসপ্লে সফটওয়্যারটি হতে পারে একটি উইন্ডো ম্যানেজার কিংবা ওয়ার্ড প্রেসের নিজেই। ডিসপ্লে সফটওয়্যার সাধারণত কোডটির জন্য একটি ইমেজ খোঁজ করে ফন্টের মধ্যে এবং সেই ইমেজটি মনিটরে প্রদর্শন করে। এভাবেই প্রতি মুহুর্তে টেক্সট প্রেসেটিংয়ের মাধ্যমে কম্পিউটারে টেক্সট প্রেস হতে থাকে।

ইউনিকোড স্ট্যান্ডার্ড সরাসরি টেক্সটের ব্যাকরণগত নীতি ও এনকোডিংকে উদ্দেশ্য করে কাজ করে। এটি টেক্সটের অন্য কোনো বিষয়ের সাথে সম্পৃক্ত নয়। যেমন— ওয়ার্ড প্রেসের টাইপিষ্টের ইনপুট সাথে সাথে চেক করে এবং ভুল বানান থাকলে তাতে লাল রং দিয়ে টেড খেলানো আন্ডারলাইন দিয়ে দেয়। অথবা এটি নির্দিষ্ট স্থান পর একটি লাইন ব্রেক দিয়ে

দেয়, যাতে বোঝা যায় প্রিন্টে কেমন আসবে। ইউনিকোড স্ট্যান্ডার্ডের একটি নীতি হলো যে, এর মধ্যে এ ধরনের কোনো প্রসেস রাখা হবে না। আপাতত, এর নীতির মধ্যে শুধু ক্যারেক্টার এনকোডিং ও ডিকোডিং টিমতো হচ্ছে কি-না সেটিই চেক করা হবে।

বর্ণের সঠিক রূপ নির্ধারণ এবং গ্লিফ রেন্ডারিং

একটি কোড পয়েন্টকে সঠিকভাবে চিহ্নিত করা এবং স্ক্রিন কিংবা প্রিন্টারে সঠিকভাবে তা রেন্ডারের মাধ্যমে প্রকাশ করা ইউনিকোড স্ট্যান্ডার্ডের টেক্সট প্রসেসিংয়ের একটি গুরুত্বপূর্ণ অংশ। ইউনিকোডের কোড পয়েন্টের মাধ্যমে চিহ্নিত করা একটি বর্ণ হচ্ছে একটি এককীয় উপাদান, যেমন—'ল্যাটিন' বর্ণ বড় হাতের 'A' কিংবা বাংলা অংক ৫। এই চিহ্ন যখন স্ক্রিনে প্রদর্শিত হবে বা কাগজে প্রিন্ট হবে তখন তাকে বলা হয় গ্লিফ। মূলত বর্ণের ভিজুয়াল রিপ্রেজেন্টেশনকেই বলা হয় গ্লিফ।

আর ইউনিকোড স্ট্যান্ডার্ড এই গ্লিফ ইমেজ ডিফাইন করে না। স্ট্যান্ডার্ড শুধুমাত্র বলে দেয় কিভাবে বর্ণটি অনুদিত হবে। গ্লিফ কিভাবে রেন্ডারিং হবে তার সাথে ইউনিকোড স্ট্যান্ডার্ডের কোনোই সম্পর্ক নেই। বরং বর্ণগুলো স্ক্রিনে কিভাবে প্রদর্শিত হবে তার জন্য দায়ী হবে সফটওয়্যার কিংবা কম্পিউটারের হার্ডওয়্যারের রেন্ডারিং ইঞ্জিন। এমনকি ইউনিকোড স্ট্যান্ডার্ড স্ক্রিনে বর্ণটির আকার-আকৃতি ও স্টাইল কেমন হবে সে সম্পর্কেও কিছু বলে না।

ক্যারেক্টার সিকোয়েন্স

ইউনিকোড স্ট্যান্ডার্ডের আরেকটি গুরুত্বপূর্ণ টেক্সট এলিমেন্ট হলো— এক বা একাধিক বর্ণের সিকোয়েন্সে তৈরি Combining Character Sequence। এতে সাধারণত একটি মূল বর্ণ এবং এক বা একাধিক কন্স্ট্রাক্টিভ চিহ্ন থাকে— যা রেন্ডারের সময় মূল বর্ণের উপরে নিচে বা পাশে সঠিক জায়গায় বসে যায়। যেমন—'a'-এর সাথে একটি কন্স্ট্রাক্টিভ চিহ্ন \wedge ব্যবহার করে রেজারিংয়ের পর তৈরি হয় 'á'। কিংবা 'ক', '্', 'ক্' টাইপের ফলে তৈরি হয় 'ক্'। আর এ ধরনের ক্যারেক্টার সিকোয়েন্স বিভিন্ন ভাষার জন্য বিভিন্ন রকমের হয়ে থাকে।

কন্স্ট্রাক্টিভ ক্যারেক্টার সিকোয়েন্সের ক্ষেত্রে ইউনিকোড স্ট্যান্ডার্ড বলে দেয় কোন বর্ণের পর কোন বর্ণ দিয়ে কোন বর্ণ তৈরি করতে হবে। তবে এক্ষেত্রে বিন্যাসের সুবিধা রক্ষার্থে মূল বর্ণটিই প্রথমে দিতে হয় এবং এরপর কন্স্ট্রাক্টিভ চিহ্ন বা বর্ণ দিতে হয়। তবে ইউনিকোড স্ট্যান্ডার্ডে কন্স্ট্রাক্টিভ চিহ্ন বা বর্ণের আগে স্পেস দিলে ক্যারেক্টার সিকোয়েন্স নষ্ট হয়। তাই কাঙ্ক্ষিত বর্ণ পাওয়া যায় না। ক্যারেক্টার সিকোয়েন্সের ক্ষেত্রে এমতো বর্ণ বা চিহ্ন কন্স্ট্রাক্টিভ ক্যারেক্টার হিসেবে কাজ করতে পারে যার সাথে টাইপোগ্রাফিতে আদৌ কোনো সম্পর্ক নেই। যেমন— 'ক', '্', 'ক্' যখন ক্ তৈরিতে ব্যবহৃত হবে তখন '্' টাইপোগ্রাফিতে আদৌ ব্যবহৃত হবে না। তাই ক্যারেক্টার সিকোয়েন্সের ক্ষেত্রে দৃশ্যত কোন কোন বর্ণ দেখা যাবে তা নয়, বরং কোন বর্ণের পর কোন বর্ণ ইনপুট দিতে হবে সেটাই গুরুত্বপূর্ণ। ইনপুট সঠিক ধারাবাহিকতায় দেয়া হলে ইউনিকোড স্ট্যান্ডার্ডই বলে দেবে মূল বর্ণের কোন পরিবর্তিত বা বর্ধিত রূপ প্রদর্শন করতে হবে।

অনেক ক্যারেক্টার সিকোয়েন্সের ফলেই পূর্ব নির্ধারিত যৌগিক একক বর্ণ দেখা যেতে পারে (সেটি যুক্তাক্ষরও হতে পারে, আবার পরিবর্তিত রূপও হতে পারে)। যেমন— ল্যাটিন বর্ণ 'ü', দেখানোর জন্য সরাসরি তার কোড পয়েন্ট U+00FC ইনপুট দেয়া যেতে পারে কিংবা ক্যারেক্টার সিকোয়েন্সে মূল বর্ণ হিসেবে 'u' (U+0075) এবং কন্স্ট্রাক্টিভ চিহ্ন '¨' (U+0308) ইনপুট দেয়া যেতে পারে। এ ধরনের দ্বৈত ক্ষেত্র তৈরি করা হয়েছে মূলত অন্যান্য প্রচলিত ক্যারেক্টার সেট Latin-1 বা অন্যান্যদের সাথে কম্প্যাটিবিলিটি রক্ষার জন্য— যাদের মধ্যে ü কিংবা ñ-এর মতো পূর্বনির্ধারিত যৌগিক বর্ণ রয়েছে।

তাছাড়া ইউনিকোড স্ট্যান্ডার্ডে পূর্বনির্ধারিত যৌগিক একক বর্ণ বিশ্লেষণ কিংবা ধারাবাহিকতা রক্ষার জন্য ভেঙে ফেলাও সম্ভব। যেমন— বিন্যাসের সময় ü বর্ণটি u-এর গ্রুপেই পড়বে এবং সে সময় তার কোড সংরক্ষিত হবে ü হিসেবে। তাছাড়া বিভিন্ন ধরনের টেক্সট প্রসেসিংয়ের জন্য যৌগিক বর্ণ বা যুক্তাক্ষরের চেয়ে তার মূল বর্ণটি থাকলে প্রসেসিং সহজ ও দ্রুততর হয়। বিশেষ করে শব্দ বিন্যাসের সময় এটি একটি অত্যন্ত প্রয়োজনীয় বৈশিষ্ট্য হিসেবেই প্রতিপন্ন হয়।

ইউনিকোড স্ট্যান্ডার্ডের মূলনীতি

ইউনিকোড স্ট্যান্ডার্ডটি প্রণয়নের পেছনে জড়িত ছিল কম্পিউটার পেশাজীবী, ভাষাবিদ ও বিদ্বানদের একটি দল। আর এই স্ট্যান্ডার্ডটি প্রণয়নের প্রথম থেকেই মূল লক্ষ ছিল বিশ্বব্যাপী একটি স্ট্যান্ডার্ড প্রণয়ন করা, যা যে-কোনো জায়গায় টেক্সট এনকোডিংয়ের জন্য ব্যবহার করা সম্ভব। আর তাই ইউনিকোড স্ট্যান্ডার্ডের মূলনীতি হলো—

- সার্বজনীন বর্ণমালার এক সংগ্রহশালা প্রণয়ন;
- উপযুক্ত সংগ্রহশালা হিসেবে এটিকে প্রতিষ্ঠিত করা ;
- বর্ণ সংরক্ষণ করা, গ্লিফ নয়;
- উপযুক্তভাবে প্রয়োগ করা এবং শব্দ সম্পর্কে সঠিক জ্ঞান সংরক্ষণ করা;
- শুধুমাত্র টেক্সটকেই সাপোর্ট দেয়া;
- লজিক্যালভাবে বিন্যাস করা;
- সঠিকভাবে সমন্বয়সাধন করা;
- ডাইনামিক কম্পোজিশন করা;
- সমতুল্য ক্রম রক্ষা করা;
- রূপান্তরযোগ্য রাখা।

আর এই নীতিমালার কারণেই এটি বিভিন্ন আন্তর্জাতিক, জাতীয় ও করপোরেট ক্যারেক্টার সেটের সাথে একত্রে কাজ করতে পারে। যেমন—

ইউনিকোড ক্যারেক্টার সেটের প্রথম ২৫৬ টি বর্ণ বহুল ব্যবহৃত Latin-1 ক্যারেক্টার সেট থেকেই নেয়া হয়েছে।

বিভিন্ন ভাষার লিপির মধ্যে একই বর্ণ থাকলে তা একাধিকবার যাতে এনকোডিং করা না হয়, সে বিষয়টিও ইউনিকোডের ক্যারেক্টার সেটে লক্ষ রাখা হয়েছে। আর এর ফলেই এই স্ট্যান্ডার্ডে এক বর্ণ দুইবার ব্যবহৃত হয় নি। যেসব বর্ণ দেখতে একই রকম অথচ একাধিক ভাষার লিপিতে রয়েছে তাদেরকে একটি মাত্র লিপিভুক্তই করা হয়েছে। সে কারণে দেবনাগরী লিপিতে দাঁড়ি সদৃশ ডাঙা থাকার ফলে পৃথকভাবে বাংলা লিপিতে দাঁড়ি সংযুক্ত করা হয় নি। তাছাড়া আসাম অঞ্চলের বাংলা বর্ণমালায় এমন কিছু বর্ণ রয়েছে— যা আমাদের বর্ণমালায় নেই। কিন্তু ইউনিকোড সার্বজনীন বলে এখানে কোনো বর্ণকে বাদ দেয়া হয় নি, বরং সকল বর্ণকেই সংযুক্ত করা হয়েছে।

তাছাড়া বাংলা, ইংরেজি প্রভৃতি ভাষা বাম থেকে ডান দিকে লেখা হয়। আবার, আরবি লেখা হয় ডান থেকে বাঁয়ে। এই উভয়মুখী দিকে লেখার সাপোর্ট দেয়ার জন্য ইউনিকোডে রয়েছে একটি এলগোরিদম। ইউনিকোড স্ট্যান্ডার্ডে যে-কোনো মুখে লেখার বর্ণই রয়েছে। এবং দুটো মিলিয়ে লিখলেও তা সাপোর্ট দিতে সক্ষম এই স্ট্যান্ডার্ড। এসব লিপির জন্যই এটি মেমোরিতে লজিক্যালভাবেই ইউনিকোড টেক্সট সংরক্ষণ করে— তা কী-বোর্ডে যা-ই টাইপ করা হোক না কেন!

ক্যারেক্টারের কোড নির্ধারণ

ইউনিকোডের প্রতিটি কোড পয়েন্টের জন্যই একটি সংখ্যা বাচক মান দেয়া হয়। এই প্রতিটি সংখ্যাকে বলা হয় কোড পয়েন্ট। আর যখন টেক্সটে এটি সংরক্ষণ করা হয় তখন একটি হেক্সাডেসিমেল সংখ্যা হিসেবে এটি সংরক্ষণ করা হয় এবং শুরুতে U দেয়া হয়। যেমন—যদি U+0041 কোড পয়েন্টটি আসলে হেক্সাডেসিমেল সংখ্যা 0041, যার দশমিক মান হলো ৬৫। আর ইউনিকোড স্ট্যান্ডার্ডে A-এর মান হলো ৬৫।

তাছাড়া প্রতিটি বর্ণের একটি একক নামও রয়েছে এবং শুধু সেই নামেই তার পরিচয়। যেমন— U+0041 কোডটির বর্ণটির নাম হলো—'LATIN CAPITAL LETTER A'। আবার U+0A1B কোডটির বর্ণটির নাম হলো—'GURMUKHI LETTER CHA'। ইউনিকোড স্ট্যান্ডার্ডের এই নামগুলো ISO/IEC 10646-এর নামের ছবছ অনুসরণ।

ইউনিকোড স্ট্যান্ডার্ডে প্রতিটি কোড ব্লকে এক গ্রুপ বর্ণ একত্রে থাকে—যারা একই লিপির অন্তর্ভুক্ত। আর লিপি বা Script হলো একটি ভাষার বর্ণমালার লিখিত রূপ। এই স্ট্যান্ডার্ডে যেখানেই সম্ভব, মূল বর্ণমালার ক্রম সংরক্ষণের চেষ্টা করা হয়। যখন কোনো লিপির বর্ণমালা স্থানীয়ভাবেই যুগের পর যুগ একই ক্রমে বিন্যস্ত থাকে, বিশেষত বর্ণনুক্রমে, সেক্ষেত্রে ইউনিকোড স্ট্যান্ডার্ডও সেই ক্রমেই সাজানোর চেষ্টা করে। তবে এই কোড ব্লকগুলোর আকার বিভিন্ন লিপিতে বিভিন্ন। যেমন—Cyrillic লিপির কোড ব্লক ২৫৬-র বেশি কখনোই হয় না। কিন্তু চীনা-জাপান-কোরিয়ান (CJK) কোড ব্লকের মধ্যে একই কোড হাজার কোড পয়েন্ট রয়েছে।

কোড এলিমেন্টগুলো পুরো কোড পয়েন্টের সীমার মধ্যে লজিক্যালি গ্রুপ করা থাকে, যাকে বলা হয় কোড স্পেস। এই কোডিং শুরু হয়েছে U+0000 থেকে এবং প্রথমেই রয়েছে এসকি বর্ণমালা। এরপর তাতে রয়েছে গ্রিক, সিরিলিক, হিব্রু, আরবি, ভারতীয় সহ অন্যান্য অসংখ্য লিপিমাল। এরপর এতে রয়েছে বিভিন্ন প্রতীক ও বিরামচিহ্ন। এরপর কোড স্পেসে রয়েছে হিরাগানা, কাটাকানা ও হোপোমোফো'র মতো এশীয় কিছু লিপি। হান সম্প্রদায়ভুক্ত একত্রীভূত ইডিগ্রাফ রয়েছে আধুনিক হানগুলি লিপির ঠিক পরেই। সারোগেট কোড পয়েন্টের সীমা সংরক্ষিত রয়েছে UTF-16-এর সাথে ব্যবহারের জন্য। BMP-র শেষের দিকে একটি অংশ ব্যক্তিগত ব্যবহারের জন্য সংরক্ষিত রয়েছে। আর এই সংরক্ষিত অংশের পরেই রয়েছে কম্প্যাটিবিলিটি বর্ণের অংশ। এই কম্প্যাটিবিলিটি বর্ণগুলো হলো ক্যারেক্টার ডেরিভেট যা শুধুমাত্র পূর্ববর্তী স্ট্যান্ডার্ড ও পুরনো ব্যবহারী ক্যারেক্টার সেট থেকে ট্রান্সকোডের জন্য ব্যবহৃত হয়।

ইউনিকোড স্ট্যান্ডার্ডের কোড পয়েন্টের একটি বড় অংশ জুড়ে রয়েছে BMP ও দুই সেট সংরক্ষিত অংশ। এই কোড পয়েন্টগুলোর কোনো সার্বজনীন অর্থ না থাকলেও, এক শ্রেণীর লোকের নিজস্ব সুবিধার কথা বিবেচনা করে এগুলো উন্মুক্ত রাখা হয়েছে। যেমন— যদি কখনো কোরিগ্রাফাররা চিন্তা করেন যে, তারা বিভিন্ন নাচের মুদ্রার জন্য একসেট বর্ণমালা তৈরি করতে চান, তবে তারা তা এই সংরক্ষিত অংশ করে তা ইউনিকোড স্ট্যান্ডার্ডের মধ্যেই ব্যবহার করতে পারবেন। আবার এক সেট পেইজ লে-আউট লিপি তৈরি করে তা কোনো প্রোগ্রামে ইউনিকোডের মাধ্যমে সরাসরি ব্যবহার করতে চাইলেও এই সংরক্ষিত অংশ ব্যবহার করতে পারেন। তেমনি, কেউ যদি বাংলার বিভিন্ন যুক্তাক্ষর সরাসরি ইউনিকোডে রাখতে চান এবং এই সংরক্ষিত প্রাইভেট অংশ ব্যবহার করেন— তবে কিন্তু তা ইউনিকোড নীতিমালার লংঘন হবে। কেননা, ইউনিকোড বাংলার জন্যও রয়েছে পূর্ণাঙ্গ ক্যারেক্টার সেট এবং

যুক্তাক্ষর তৈরির জন্যও রয়েছে সুস্পষ্ট নীতিমালা। আসলে এই সংরক্ষিত অংশ কোনো জাতি বা অঞ্চলের ভাষার জন্য নয়, বরং কোনো বিশেষ পেশাজীবী শ্রেণীর জন্য আজীবন সংরক্ষণ করে রাখবে ইউনিকোড কনসোর্টিয়াম।

ইউনিকোড স্ট্যান্ডার্ড যা মেনে নেয়

ইউনিকোড স্ট্যান্ডার্ড যা মেনে নেয়, সে ব্যাপারে সংশয় এড়ানোর জন্য একটি নীতি মেনে চলে এবং এনকোডিং আর্কিটেকচারেও তার প্রতিফলন ঘটায়। তবে ইউনিকোড স্ট্যান্ডার্ড মেনে নিবে তখনই, যখন বর্ণ নিচের এই শর্তগুলো পূরণ করবে—

- সাধারণ ভাষারের অন্তর্ভুক্ত কোনো বর্ণ হলে;
- ইউনিকোডের তিনটি এনকোডিং ফরম্যাটের কোনোটি অনুযায়ী বর্ণ এনকোডিং করা হলে;
- ইউনিকোডের রীতি অনুযায়ী বর্ণ ব্যবহৃত হলে;
- অনির্ধারিত কোড ব্যবহৃত না হলে; এবং
- অপরিচিত বর্ণ নষ্ট না হলে।

ইউনিকোড স্ট্যান্ডার্ডের প্রয়োগ তখনই সিদ্ধ হবে, যতক্ষণ এটি এনকোডিংয়ের জন্য byte, word ও double word সিকোয়েন্সের নিয়ম মেনে চলবে এবং রূপান্তরিত বর্ণ ইউনিকোড স্পেসিফিকেশন অনুযায়ী হবে।

স্থায়িত্ব

ইউনিকোড স্ট্যান্ডার্ড পরিপূর্ণ হতে এখনো অনেক দেরী। এবং ইউনিকোডে অন্তর্ভুক্তির অপেক্ষায় এখনো অনেক লিপির অপেক্ষমাণ। আশা করা যায়, পরবর্তী সংস্করণগুলোতেই এই লিপিগুলো অন্তর্ভুক্ত হবে। আর অন্তর্ভুক্তির এই বিষয়টি শুধুই 'সংযোজন' প্রক্রিয়া। অন্যভাবে বলা যায় যে, ইউনিকোড স্ট্যান্ডার্ডে শুধু নতুন বর্ণ সংযুক্ত করা হয় বা যায়। কোনো বর্ণ বাদ দেয়া কিংবা নতুন করে এনকোডিং করা যায় না। আর এভাবেই ইউনিকোড স্ট্যান্ডার্ডের স্থায়িত্ব নিশ্চিত করা হয়। কেননা, একবার একটি বর্ণ এই স্ট্যান্ডার্ডে যেভাবে ইন্টারপ্রেট করা হয়, ভবিষ্যতের সংস্করণগুলোতেও সেভাবেই ইন্টারপ্রেট করা হবে এটা নিশ্চিত করার মাধ্যমেই এই স্ট্যান্ডার্ডটিকে স্টেবল রাখা হয়।

ইউনিকোড এবং ISO/IEC 10646

ইউনিকোড স্ট্যান্ডার্ডের সাথে আন্তর্জাতিক স্ট্যান্ডার্ড ISO/IEC 10646 অত্যন্ত সম্পর্কিত। ISO/IEC 10646 ইউনিভার্সেল ক্যারেক্টার সেট (UCS) নামেই বেশি পরিচিত। আর দুই কমিটির মধ্যে ঘনিষ্ঠ সম্পর্ক থাকার ফলে দুটি স্ট্যান্ডার্ডই একইভাবে বেড়ে উঠেছে এবং একটি শব্দ ভাঙার ও এনকোডিং ব্যবহার করছে।

ইউনিকোড স্ট্যান্ডার্ড ৩.০ ও ISO/IEC 10646-1:2000 হুবহু 'কোড ফর কোড' এক। আর একই সাম্যতা এই দুই স্ট্যান্ডার্ডের সব এনকোডেড বর্ণের জন্যই প্রযোজ্য, এমনকি পূর্ব এশিয়ার হান সম্প্রদায়ভুক্ত আইডিওগ্রাফিক বর্ণও এর মধ্যে পড়ে।

ইউনিকোড টেকনিক্যাল রিপোর্ট

ইউনিকোড স্ট্যান্ডার্ড সংক্রান্ত সব ধরনের তথ্য পাওয়ার জন্য তাদের ওয়েবসাইট একটা খুব ভালো উৎস। আর ইউনিকোড স্ট্যান্ডার্ডের টেকনিক্যাল সমস্ত তথ্যই পাওয়া যাবে ইউনিকোড টেকনিক্যাল রিপোর্টে। এই রিপোর্টে পাওয়া যাবে—

- ইউনিকোড টেস্ট তুলনা ও সংরক্ষণের জন্য নরমালাইজিং প্রক্রিয়া;
- সংরক্ষণের জন্য ইউনিকোড টেস্ট কমপ্রেস পদ্ধতি;
- শব্দ বিন্যাস পদ্ধতি;
- টেস্টে লাইন ব্রেকিং পদ্ধতি;
- বিভিন্ন ধরনের কেস (লোয়ারকেস, আপার কেস, টাইটেল কেস) রূপান্তর প্রক্রিয়া;
- CRLF নিয়ন্ত্রণ;
- নিয়মিত কমান্ড তৈরির পদ্ধতি সহ আরো অনেক টেকনিক্যাল বিষয়।

ইউনিকোডের টেকনিক্যাল রিপোর্ট বেশ কয়েক ধরনের হয়ে থাকে। তার মধ্যে A Unicode Technical Report (UTR)-এর মধ্যে থাকে নানান ধরনের তথ্য বা সংজ্ঞা, অথবা উভয়টি। প্রতিটি ইউনিকোড স্ট্যান্ডার্ডের মূল ভাঙ্গনের জন্য একটি UTR থাকবেই। তবে ক্ষেত্রবিশেষে এবং প্রায় প্রতিক্ষেত্রেই UTR দুইভাগে বিভক্ত হয়— A Unicode Standard Annex (UAX) ও A Unicode Technical Standard (UTS)। UAX ইউনিকোড স্ট্যান্ডার্ডেরই একটি সমন্বিত অংশ। কিন্তু এটি আলাদাভাবে প্রকাশিত হয়। ইউনিকোড স্ট্যান্ডার্ডের ভাঙ্গন নম্বরই UAX-এরও ভাঙ্গন নম্বর হয়ে থাকে। UTS হলো একটি স্বাধীন সংজ্ঞাবাচক ডকুমেন্টেশন। UTS-এ সাধারণত ইউনিকোডের স্বীকৃত স্ট্যান্ডার্ডগুলো প্রকাশিত হয়, তাই মূল স্ট্যান্ডার্ডের ভাঙ্গন নম্বরটাই হয় UTS-এর ভাঙ্গন নম্বর। অন্যদিকে UAX-এ বিভিন্ন প্রস্তাবিত স্ট্যান্ডার্ড ও স্বীকৃতির অপেক্ষায় আছে এমন তথ্য থাকে। তাছাড়া স্ট্যান্ডার্ড তৈরির বিভিন্ন পর্যায়ে ইউনিকোড টেকনিক্যাল কমিটি বিভিন্ন ড্রাফট বা প্রস্তাবিত ড্রাফট রিপোর্টও সাধারণের জন্য উন্মুক্ত করে রাখে। তবে এই ড্রাফটগুলো সাধারণত ইউনিকোড কনসোর্টিয়ামের অনুমোদনের পূর্বেই প্রদর্শিত হয়। এরকম দু'ধরনের রিপোর্ট হলো— A Draft Unicode Technical Report (DUTR) এবং A Proposed Draft Unicode Technical Report (PDUTR)। DUTR হলো অনুমোদনের আগে কোনো UTR, যাতে UTR-এর সব বৈশিষ্ট্যই রয়েছে, এবং তখন শুধু অনুমোদনেরই অপেক্ষা। আর PDUTR হলো কোনো UTR তৈরির উদ্দেশ্যে কাজ শুরু হওয়ার একেবারে প্রারম্ভিক অবস্থার রিপোর্ট। DUTR ও PDUTR-এর সবই হলো অস্থায়ী। অনুমোদনের পূর্বে এগুলোতে একাধিক পরিবর্তন ও পরিমার্জনের সম্ভাবনা থাকে।

ইউনিকোড স্ট্যান্ডার্ডে কোনো বর্ণ খুঁজতে হলে

ইউনিকোড স্ট্যান্ডার্ডে কোনো নির্দিষ্ট বর্ণ খুঁজতে হলে তা খুঁজতে হবে অনলাইন কোড চার্টে। ইউনিকোডে প্রতিটি বর্ণের জন্য একটি কোড পয়েন্ট রয়েছে যেটি একটি হেক্সাডেসিমেল সংখ্যা। তবে ইউনিকোডের যে জায়গায় বর্ণটি পাওয়া যেতে পারে বলে কেউ ধারণা করে সেখানেই যে বর্ণটি পাওয়া যাবে— এমনটা ভাবা ঠিক নয়। ইউনিকোডে সব বর্ণই কোনো একটি ব্লকে গ্রুপ আকারে থাকে। আর কোনো ভাষার বর্ণমালাকে বিভিন্নভাবে শ্রেণীবিন্যাস করা যায়। তাছাড়া যতি চিহ্নসহ ভাষার অন্যান্য চিহ্নগুলো লিপিতে বিভিন্নভাবে ব্যবহার করা যায়। এমনকি লিপিতে অনেক বর্ণই কিভাবে লেখা হবে— তাই আদৌ সংজ্ঞায়িত নয়। আবার, কোনো একটি ভাষার লিপিতে এমন কিছু বর্ণ থাকতে পারে যা আরো একাধিক ভাষাতে লিপিতে বিদ্যমান। যেমন— 0-9 অংকগুলো অসংখ্য ভাষা লিপিতেই রয়েছে। একইভাবে দেবনাগরী লিপির 'ডাঙা' সমস্ত ভারতীয় লিপিতেই হয় 'ডাঙা', নয়তো 'দাঁড়ি' হিসেবে রয়েছে। এ কারণেই কোনো বর্ণ খুঁজে পেতে হলে তাকে একাধিক স্থানে খুঁজে দেখা উচিত। আর সেজন্য ইউনিকোডের 'ক্যারেক্টার ইন্ডেক্স' হলো আদর্শ জায়গা। কিন্তু এই 'ক্যারেক্টার ইন্ডেক্স' শুধুমাত্র মুদ্রিত আকারে পাওয়া যায়। বিকল্প হিসেবে অনলাইনে ইউনিকোডের NameList-এ টেস্টসার্চ করে খুঁজে দেখা যেতে পারে। তাছাড়া বিভিন্ন টেকনিক্যাল রিপোর্টে বর্ণগুলো বিভিন্নভাবে সজ্জিত আছে। তবে এই সাজানোর ফলে বর্ণ খোঁজার কাজটি বেশ সহজ হয়ে গেছে।

আবার, বিভিন্ন ভাষায় লিপিতে একই বর্ণ একেক সময় একেকভাবে লেখা হয়। কিন্তু ইউনিকোড স্ট্যান্ডার্ডের বর্ণটি কেবলমাত্র একটিভাবেই সংরক্ষিত হয়। তাই অন্য কোনো রূপে খোঁজা হলে—তা নাও পাওয়া যেতে পারে। যেমন— আরবি বর্ণ 'হা' ৪ ভাবে লেখা হয়। কিন্তু ইউনিকোড সেটি একভাবেই সংরক্ষিত আছে। তাছাড়া অনেক বহুরূপী বর্ণই কোড পয়েন্টের সিকোয়েন্সে প্রকাশিত হয় তাই সেগুলোকে সরাসরি চার্টে পাওয়া না-ও যেতে পারে। একইভাবে ভারতীয় লিপিগুলোর যুক্তাক্ষরের ভগ্নাংশগুলোও আলাদাভাবে কোড চার্টে নেই। যেহেতু তারা একটি ব্যঞ্জনবর্ণ + হলন্ত + অন্য ব্যঞ্জনবর্ণের সমন্বয়ে তৈরি হয়, অর্থাৎ কোড পয়েন্টের সিকোয়েন্সে প্রকাশিত হয়।

আবার, অল্প কিছু বর্ণ আছে যা একাধিক অর্থে ব্যবহৃত হয়, যেমন— A কোনো ব্যবহৃত হয় Capital letter A with ring আবার কখনো Angstrom চিহ্ন হিসেবে। এধরনের ক্ষেত্রে একই বর্ণের দুটি কোড পয়েন্ট হবার পেছনের মূল কারণ শুধুমাত্র যে পুরনো ক্যারেক্টার সেট থেকে তাদের নেয়া হয়েছে— সেই ক্যারেক্টার সেটের সাথে সামঞ্জস্য রাখা করা। তাছাড়াও ক্ষেত্র বিশেষে ইউনিকোড ইন্টারনেটের জন্য একটি বিশেষ ফরম্যাট ব্যবহার করে, যা Form C নামে পরিচিত। এই ফরম্যাটে সব সময়ই একটি কোড পয়েন্ট বা কোড পয়েন্টের সিকোয়েন্স ব্যবহার করা হয় আর কিছুই না। আবার, অল্প কিছু ক্ষেত্রে ইউনিকোড গ্রিফকেও আলাদা বর্ণ হিসেবে ধরে নেয়। যেমন— ‘’ চিহ্নটি Cyrillic বর্ণে হালকা উচ্চারণ চিহ্নিত করে বলে একে একটি বর্ণ হিসেবে ব্যবহার করা হয়, আবার প্রতীক হিসেবে অন্যান্য ক্ষেত্রেও এটি ব্যবহার করা যায়। অবশ্য এগুলো খুব কমই হয় এবং এদেরকে ব্যতিক্রম হিসেবেই ধরে নেয়া হয়।

ইউনিকোড এবং বাংলা

ইউনিকোড স্ট্যান্ডার্ড সম্পর্কে বিস্তারিত জানার পর একটা বিষয় হয়তো স্পষ্ট হয়ে গেছে যে, বাংলা নিয়ে যেসব সন্দেহ ও বিতর্ক রয়েছে তা মূলত অমূলক। তারপরও ইউনিকোডে বাংলার অবস্থান আরো স্পষ্ট করে দেয়া প্রয়োজন।

ইউনিকোডে কাছাকাছি বা সমগোত্রীয় ভাষার লিপিসমূহকে সবসময় একই শ্রেণীভুক্ত বা পরিবারভুক্ত বিবেচনা করা হয়। এবং এক শ্রেণীভুক্ত সবার জন্যই একটি সাধারণ ও একক রূপরেখা প্রণীত ও ব্যবহৃত হয়। আমরা এতক্ষণে একটি বিষয় জেনে গেছি যে ভাষা আর লিপি এক নয়। কার্যত, ভাষার লিখিত রূপই লিপি। আবার একটি বিষয়ও আমাদের কাছে স্পষ্ট হওয়া উচিত যে, অনেক ভাষার বর্ণমালাতে একই বর্ণ থাকলেও ইউনিকোড তাকে একটিমাত্রই কোড পয়েন্ট দেবে। ইউনিকোড স্ট্যান্ডার্ড ২.০-এ বাংলা কে ইউনিকোডভুক্ত করা হয়।

ছক-১.১-এ ইউনিকোড বাংলা লিপির অবস্থান দেখানো হলো।

দেবনাগরী, বাংলা, গুরুমুখী, গুজরাটি, ওড়িয়া, তামিল, তেলেগু, কানাডা, মালায়ালম, সিনহালা, থাই, লাও, তিব্বতী, বার্মিজ ও খেয়ার সহ মোট ১৫টি লিপি ইউনিকোডে একই শ্রেণীভুক্ত। একই শ্রেণীভুক্ত হবার কারণ হচ্ছে, ইউনিকোডের পর্যবেক্ষণ অনুযায়ী এই সব ক'টি লিপিরই পূর্বপুরুষ হচ্ছে ব্রাহ্মী লিপি। অর্থাৎ ব্রাহ্মী লিপি থেকেই এসব ভাষার উৎপত্তি। আর ব্রাহ্মীর উত্তরসূরি হবার কারণে এসব লিপির গঠনতন্ত্র থেকে শুরু করে প্রয়োগরীতিও প্রায় একই রকম। ইউনিকোডের এই শ্রেণীবিন্যাসগত রূপরেখাটির কারণে অনেকেই মনে করেন, ভারতীয় পরিবারের অংশ হিসেবেই বুঝি বাংলা ইউনিকোডে অবস্থান করছে। কিন্তু এটা আসলে একটা ভুল ধারণা। কারণ, এই শ্রেণীতে থাই ও লাও ভাষাকে দেখলেই এটা স্পষ্ট হয়ে যায়। থাই ও লাও ভারতীয় লিপি না হয়েও ব্রাহ্মী লিপি থেকে উৎপন্ন বলে এই শ্রেণীভুক্ত।

একই সাথে, এ কথাও ঠিক যে, ইউনিকোড লিপি শ্রেণীবিন্যাসের ক্ষেত্রে লিপির গঠনগত দিকটিকেই প্রাধান্য দেয়। কে কার অন্তর্ভুক্ত হলো তা এখানে একেবারেই গৌণ। তাছাড়া ইউনিকোড কখনো ভাষা নিয়ে কাজ করে না, তাদের একমাত্র উদ্দেশ্য হলো ভাষার লিখিত রূপ তথা লিপিকে

গ্রহণযোগ্যভাবে (ব্যাকরণ সম্মতভাবে) কোড পয়েন্ট দেয়া। ইউনিকোডের মূলনীতি থেকে আমরা এটাও জানি যে, কোনো ভাষার বর্ণমালাকে ইউনিকোড দু'ভাগে ভাগ করে— মূল লিপি (World Script) ও পরিবর্তিত লিপি বা গ্লিফ। যাবতীয় স্বাধীন বর্ণকে (স্বরবর্ণ, ব্যঞ্জনবর্ণ, মাত্রা (কার বা ফলা) ও অন্যান্য মৌলিক চিহ্ন সমূহ) মূল লিপির অন্তর্ভুক্ত বিবেচনা করা হয়। আর যে সমস্ত বর্ণ ও মাত্রা অন্য বর্ণের সংস্পর্শে বা সংযুক্তির ফলে উৎপন্ন হয় কিংবা পরিবর্তিত রূপে ব্যবহৃত হয়, সেগুলো গ্লিফ হিসেবে বিবেচিত হয়। ইউনিকোড শুধুমাত্র মূল লিপিকেই কোড পয়েন্ট দেয় এবং প্রত্যেকের একটি সনাক্তকারী পৃথক নাম থাকে। কিন্তু গ্লিফের জন্য একটি রূপরেখা প্রণীত হয় মাত্র— কোনো কোড পয়েন্ট বরাদ্দ দেয়া হয় না। মূল লিপির আরো একটি লক্ষণীয় বৈশিষ্ট্য হলো যে, সেটি কোনো একক বর্ণ ও মাত্রার একাধিক বা পরিবর্তিত উপস্থিতি কিংবা নামকরণ গ্রহণ করে না। তাছাড়া ইউনিকোডের নীতি অনুযায়ী ইউনিকোড স্ট্যান্ডার্ড মূলত একটি সংরক্ষণ ব্যবস্থা। বর্ণের আকার-আকৃতি নিয়ে এটি কাজ করে না। সুতরাং যে-কোনো ভাষার লিপিকেই ইউনিকোড প্রস্তাবনার ক্ষেত্রে এসব বিষয় অত্যন্ত সতর্কতার সাথে বিবেচনা রাখতে হবে। আর ইউনিকোডের এসব নীতিমালার কারণেই বিএসটিআই কর্তৃক উদ্ভাবিত BSD-1520 :1995 কোডিং প্রস্তাবনা ইউনিকোড কনসোর্টিয়াম কর্তৃক ত্রুটিপূর্ণ বিবেচিত হওয়ায় প্রত্যাখ্যাত হয়। তবে BSD-1520 : 2000 (First Revision) প্রায় অনেকটাই ইউনিকোডের আদলে তৈরি। এবং এর ০.৫ অনুচ্ছেদে স্পষ্ট করে উল্লেখ করা হয়েছে যে, এটি ISO/IEC 10646-1991-এর ভিত্তিতে প্রণয়ন করা হয়েছে। সেই একই অনুচ্ছেদে তাদেরকে দ্বন্দ্ববাদও দেয়া হয়েছে। BSD-1520 : 2000-এ ইউনিকোডের ছব্ব প্রতিলিপি হিসেবে সব বর্ণ বসানো হলো 'ঐ' (লি) বর্ণটি তাতে নেই এবং বাড়তি হিসেবে রয়েছে 'ৎ' (খও ত) ও '।' (দাঁড়ি)।

বাংলা ব্যাকরণের পুনরাবৃত্তি

বাংলা ব্যাকরণে যে-কোনো বর্ণের সংক্ষিপ্ত রূপকেই বলা হয় 'মাত্রা'। আর স্বরবর্ণের সংক্ষিপ্তরূপকে বলা হয় 'কার' ও ব্যঞ্জনবর্ণের সংক্ষিপ্তরূপকে বলা হয় 'ফলা'। সেই হিসেবে আ, ই, ঈ, উ, ঊ, ঋ, ঌ, ঍, ঐ, ও, ঔ প্রতিটিরই সংক্ষিপ্ত রূপ রয়েছে— যাদেরকে আমরা উচ্চারণ করি মূল বর্ণের পরে কার যুক্ত করে। আর এসব স্বরমাত্রা বা কার-এর লিখিত রূপ হলো—
 আ, ই, ঈ, উ, ঊ, ঋ, ঌ, ঍, ঐ, ও, ঔ

এই প্রতিটি কারই একটিমাত্র বর্ণের সংক্ষিপ্ত রূপ। তাই 'ঐ' বা ও-কার কখনই একটি 'ঐ' (এ-কার) ও 'ঐ' (আ-কার)-এর সমন্বয়ে গঠিত হয় না। যদিও একটি শিশুকে লিখতে শেখানোর সময় এ-কার ও আ-কার-এর সমন্বয়ে ও-কার লিখতে শেখানো হয়, কিন্তু একজন প্রাপ্তবয়স্ক ব্যক্তি যিনি সামান্যতম ব্যাকরণ জানেন, তিনি বলবেন না, ও-কার মানে এ-কার ও আ-কারের সমন্বয়। বরং ব্যাকরণিক দৃষ্টিতে ও-কার (ঐ) একটি স্বরবর্ণ ও 'ঐ'-এ সংক্ষিপ্ত রূপ। তাই যদি আবহমান কাল ধরে 'ঐ' ও 'ঐ'-এর সমন্বয়ে 'ঐ' লিখিত বলে ইউনিকোডেও এমনটা দাবি করেন— তবে তা যেমন ইউনিকোডের নীতিমালা অনুযায়ী অযৌক্তিক, তেমনি বাংলা ব্যাকরণের দৃষ্টিতেও ভুল। একই কথা ঔ-কারের ক্ষেত্রেও প্রযোজ্য।

তেমনি বাংলা বর্ণ 'ক্ষ', 'ৎ' কিংবা ব-ফলার ক্ষেত্রেও একই কথা প্রযোজ্য। 'ক্ষ' হলো 'ক', 'খ', 'ব'-এর মিলিত একটি যৌগিক বর্ণ। তাই ইউনিকোড নীতিমালায় এটি আলাদা কোড পয়েন্ট পাবার উপযুক্ত নয়। একইভাবে 'ৎ' (উচ্চারণ-খও ত)-ই বলে দেয় যে এটি 'ত'-এর খও রূপ। এবং আদিত 'ত' দ্রুত লিখতে গিয়েই এই বর্ণের উৎপত্তি। অতএব, 'ৎ'-ও কোড পয়েন্টের দাবিদার হতে পারে না। আবার 'ব-ফলা' বা 'মাত্রাহীন ব'-এর ক্ষেত্রেও সেই একই কথাই প্রযোজ্য যে 'ফলা' হলো ব্যঞ্জনবর্ণের সংক্ষিপ্ত রূপ। এবং সংক্ষিপ্ত রূপ বা গ্লিফ আলাদা কোড পয়েন্ট পেতে পারে না। একইভাবে, র-ফলা, য-ফলা, রেফ ' ' ইত্যাদিও ইউনিকোডে আলাদা কোড পয়েন্ট পায় নি, কিন্তু সঠিকভাবে বর্ণের সিকোয়েন্সে ইনপুট দিলে এরা স্ক্রিনে প্রদর্শিত হবে।

বাংলার প্রতি অবিচার (!)

কেউ যদি চাইনিজ বা জাপানিজ ভাষার উদাহরণ টেনে বলেন যে, ইউনিকোডে এসব ভাষার যুক্তাক্ষর অন্তর্ভুক্ত হলে বাংলার প্রতি কেন অবিচার! তাহলে একটি বিষয় স্পষ্ট হওয়া জরুরি যে, হান লিপিভুক্ত ভাষায় বর্ণের পরিবর্তে সরাসরি শব্দ লেখা হয় বলে এক একটি শব্দই এখানে এক একটি লিপি দ্বারা প্রকাশ করা হয়। তদনুসারে ইউনিকোড এদের সবগুলোকে পৃথক কোড পয়েন্ট বরাদ্দ না দিয়ে সেসব লিপির বিশেষজ্ঞ দ্বারা সেগুলোর জন্য বিশেষভাবে কোড পয়েন্ট বরাদ্দের ব্যবস্থা গ্রহণ করেছে। যা এখানে আলোচনা অবাস্তব।

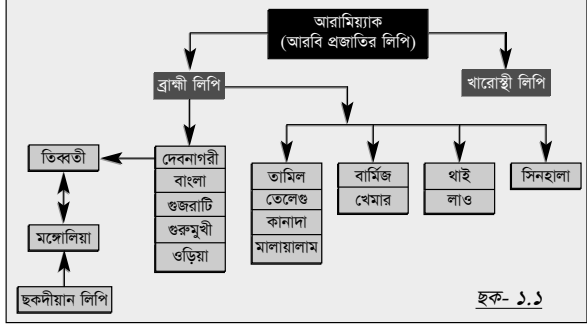
কিন্তু কেন যুক্তাক্ষর রাখা উচিত নয় ?

অনেকেরই মনে সন্দেহ থাকতে পারে কেন আমরা যুক্তাক্ষরের জন্য কোড পয়েন্ট বরাদ্দ নিতে পারি না। কিংবা নিলেই বা কী ক্ষতি হতো ? একটি উদাহরণের মাধ্যমে ব্যাখ্যা করলেই বিষয়টি স্পষ্ট হবে। ধরি, ৬টি শব্দ কমল, কাজ, ক্লাস, ক্রান্তি, পাচন, প্রাণ রয়েছে। এদেরকে দু'ভাবে কম্পিউটারে ইনপুট দেয়ার ব্যবস্থা করা হলো। প্রথমটিতে যুক্তাক্ষরগুলোকে কোড পয়েন্ট প্রদানের মাধ্যমে একবারে ইনপুট দেয়া হচ্ছে, আর দ্বিতীয়টিতে ইউনিকোড মেনে শুধুমাত্র মূল বর্ণ বা মূল বর্ণের ক্যারেক্টার সিকোয়েন্সের মাধ্যমে ইনপুট দেয়া হচ্ছে।

প্রথম পদ্ধতিতে যুক্তাক্ষর 'ক্র', 'ক্র' কিংবা 'স্ত' একটি মাত্র কোড পয়েন্টের মাধ্যমে ইনপুট দেয়ার ফলে 'ক্র' কিংবা 'ক্র'-এর সাথে 'ক'-এর কোনো সম্পর্কই থাকল না। কেননা, কোড পয়েন্ট কিংবা সাংখ্য মানের বিবেচনায় প্রতিটিই একক। ফলে শব্দবিন্যাসের সময় কোনটি আগে আসবে 'ক', 'ক্র' নাকি 'ক্র'! হ্যাঁ, এটি নির্ধারণ করা সহজ হবে না। কেননা, সেক্ষেত্রে আমাদের প্রায় আড়াই হাজার বর্ণ (বহুল মতামতের মধ্যে সবচেয়ে বড় সংখ্যাটি মেনে নিয়ে)-কে বিবেচনা করতে হবে। আর শব্দ বিন্যাসের জন্য যে Collation Table প্রয়োজন হবে— তা হবে বিশাল বড়, ফলে সময় যেনো বেশি লাগবে তেমনি কার্যকারিতাও কমে যাবে অনেকখানি।

বাংলা শব্দবিন্যাসের ক্ষেত্রে সমস্যা ও ইউনিকোডের সমাধান

বাংলায় 'মাত্রা' ব্যবহার বাংলা লিপির সবচেয়ে বিময়কর দিক। এবং শুধু এই কারণেই কম্পিউটারে বাংলার প্রয়োগ ঘটানো এক দুরূহ কাজ হয়ে দাঁড়িয়েছে। কেননা, বাংলা মাত্রার



	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
00	Basic Latin								Latin 1 Supplement							
01	Latin Extended-A								Latin Extended-B							
02	Latin Extended-B				IPA Extensions				Spacing Modifiers							
03	Combining Diacritics												Greek			
04	Cyrillic															
05	Cyrillic Sup.				Armenian								Hebrew			
06	Arabic															
07	Syriac		Mandac?		???		Thaana				Juko?					
08	(Avestan and Pahlavi)				(Phoenician)		(Aramaic)		(Tifinagh)		(Samaritan?)					
09	Devanagari								Bengali							
0A	Gurmukhi								Gujarati							
0B	Oriya								Tamil							
0C	Telugu								Kannada							
0D	Malayalam								Sinhala							
0E	Thai								Lao							

ইউনিকোড কোড ব্লকে বাংলার অবস্থান

কোনোটি ডানে, কোনোটি বামে, কোনোটি নিচে, আবার কোনোটি বা ডানে, বামে দু'দিকেই বসে। আর ফলশ্রুতিতে বাংলা শব্দবিন্যাস হয়ে দাঁড়িয়েছে আরো জটিল। আর এ ক্ষেত্রে ইউনিকোডের ফোনেটিক নীতি সত্যিকার অর্থেই একটি কার্যকরী নীতি।

ইউনিকোড টেকনিক্যাল কমিটির তাত্ত্বিকদের মতে, বাংলার যাবতীয় মাত্রাকেই যদি মূল বর্ণের ডানে বসানো হয় তবে সর্ট করা আর কোনো সমস্যাই হবে না। শুনে অনেকেই অবাক হলেও, এ বিষয়টি খুবই সস্তব। কেননা, কম্পিউটার স্ক্রিনে যা দেখায় এবং মেমোরিতে ডাটা যেভাবে সংরক্ষণ করে এ দুটি পদ্ধতি এক নয়। তাই মেমোরিতে সংরক্ষণের ক্ষেত্রে মূল বর্ণের পরে মাত্রা বসালেই সর্টিংয়ের আর কোনো সমস্যা হবে না। আর ইউনিকোডের বদৌলতে মাত্রা স্ক্রিনে কোথায় দেখানো হবে— সেটি কোনো সমস্যাই নয়।

যদি ফোনেটিক নীতিতে ডাটা সংরক্ষণ না করা হয় এবং যেভাবে শব্দ আমরা লিখি কিংবা স্ক্রিনে দেখি তাহলে যে সমস্যা হতে পারে তা একটি উদাহরণের সাহায্যে ব্যাখ্যা করা হচ্ছে।

চিত্র-১.২-এ স্পষ্ট বোঝা যাচ্ছে যে, তিন ক্ষেত্রেই প্রথম বর্ণ 'ক' হলেও শুধু লেখন পদ্ধতিতে ইনপুট নেবার ফলে দুটি ক্ষেত্রে 'ে' পরিণত হচ্ছে প্রথম বর্ণে— যা ডাটা সর্টিংয়ে একটি বড় বাধা। কিন্তু ফোনেটিক পদ্ধতিতে ডাটা সংরক্ষণ করলে বাংলায় ডাটা সর্টিং সমস্যা দূর হবে সহজেই।

আবার, উচ্চারণের দিক থেকেও ফোনেটিক উচ্চারণ বহুল প্রচলিত একটি পদ্ধতি। ফলে ইউনিকোডের এই সমাধানের ফলে সত্যিকার অর্থেই ডাটাবেজ থেকে শুরু করে সব ক্ষেত্রেই বাংলা ডাটা সর্টিং করা সস্তব হবে। আর এক্ষেত্রে ডাটা ইনপুট, সংরক্ষণ ও প্রদর্শন হবে চিত্র-১.৩ অনুযায়ী।

শব্দ	লেখন পদ্ধতিতে সংরক্ষণ	প্রথম বর্ণ	ফোনেটিক পদ্ধতিতে সংরক্ষণ	প্রথম বর্ণ
কাজ	কাজ	ক	কাজ	ক
কেন	কেন	ে	কেন	ক
কোন	কোন	ে	কোন	ক

চিত্র-১.২

ইনপুট		সংরক্ষণ ও প্রদর্শন
ক U+0995	ি U+09BF	কি U+0995 U+09BF
ক U+0995	ৌ U+09CB	কৌ U+0995 U+09CB

চিত্র-১.৩

ফোনেটিক পদ্ধতিতে সংরক্ষণের জন্য ফোনেটিক ইনপুট কতটা জরুরি ?

অনেকেই মনে করেন, ফোনেটিক পদ্ধতিতে ডাটা সংরক্ষণের জন্য ফোনেটিক পদ্ধতিতে ইনপুট নেয়াটা বাধ্যতামূলক। কিন্তু এই ধারণাটা পুরোপুরি সত্যি নয়। ইনপুট যেভাবেই নেয়া হোক না কেন, তা যদি কী-বোর্ড ড্রাইভারের মাধ্যমে ইউনিকোড সমর্থিত পদ্ধতিতে সংরক্ষণ করা হয়, তবেই সেটি ইউনিকোড সমর্থিত হবে। এমনকি 'ে', 'ক', 'ী' ইনপুট নিয়েও তা যদি 'ক', 'ে', 'ী' -এর কোড পয়েন্ট হিসেবে সংরক্ষণ করা হয়, তবেও সেটি ইউনিকোড সমর্থিত হবে। কিন্তু সেজন্য যে ধরনের কী-বোর্ড ড্রাইভারের প্রয়োজন হবে, তা উইন্ডোজে যতটা সহজে ব্যবহার করা যাবে, লিনাক্স কিংবা অন্যান্য অপারেটিং সিস্টেমে তত সহজে করা যাবে না। তাই ফোনেটিক ইনপুটই এক্ষেত্রে আদর্শ।

বাংলা যতি চিহ্ন 'দাঁড়ি' নেই ইউনিকোড স্ট্যান্ডার্ডে !

অনেক বাংলা বর্ণ নেই ইউনিকোড স্ট্যান্ডার্ডে। সেগুলো না থাকার ব্যাখ্যা অনেকের কাছে গ্রহণযোগ্য হলেও যতি চিহ্ন 'দাঁড়ি' না থাকার ব্যাখ্যা অনেকেই মেনে নিতে পারে না।

ইউনিকোড স্ট্যান্ডার্ডের নীতি অনুযায়ী একই বর্ণ যদি একাধিক ভাষার লিপিতে থাকে তবে তাকে আলাদাভাবে দুইবার এনকোডিং করা হবে না। এই নীতি অনুযায়ী ব্রাহ্মী লিপির অন্তর্ভুক্ত অপর একটি ভাষা দেবনাগরী (অনেকেই একে 'হিন্দি' বলে থাকেন)-র যতি চিহ্ন 'डा' (।) আর বাংলা ভাষার যতি-চিহ্ন 'দাঁড়ি' (।)-এর লিখিতরূপ একই হওয়ায় বাংলা 'দাঁড়ি'র জন্য পৃথক কোড পয়েন্ট বরাদ্দ দেয়া হয় নি। কেননা, ইউনিকোড স্ট্যান্ডার্ড সবার জন্য উন্মুক্ত এবং কেউ যদি দেবনাগরী ডাঙা ব্যবহার করতে চায়, তাতে কোনো বাধা নেই। কারণ ডাঙা ও 'দাঁড়ি' দুটোই ব্রাহ্মী লিপি থেকে উৎপন্ন, দুটো দেখতেই এক রকম এবং দুটোর কাজই এক।

ইউনিকোডের জন্যই ওপেন টাইপ প্রযুক্তি

ইউনিকোড কিভাবে একভাবে ইনপুট নিয়ে অন্যভাবে প্রদর্শন করবে— এ প্রশ্নটাও মনে জাগা খুবই স্বাভাবিক। কিন্তু ইউনিকোড এই কাজটি করে ওপেন টাইপ প্রযুক্তির সাহায্যে। মাত্রার প্রয়োগ থেকে শুরু করে আমাদের যাবতীয় সংযুক্ত বর্ণ তৈরিতে প্রধান ভূমিকা পালন করে এই ওপেন টাইপ প্রযুক্তি। এই প্রযুক্তির কল্যাণেই 'ক+ি' ইনপুট দিয়ে 'কি' দেখা সস্তব হয়।

কোন বাংলা কী-বোর্ড আদর্শ ?

ইউনিকোডের সাথে কী-বোর্ডের কোনো বাড়তি সম্পর্ক নেই। যে কী-বোর্ডই তার কী-প্রেসে ইউনিকোডের বাংলা বর্ণের কোড জেনারেট করবে, সেটিই হতে পারে ইউনিকোড সমর্থিত বা ইউনিকোড ভিত্তিক বাংলা কী-বোর্ড।

কেমন হওয়া উচিত ইউনিকোড সমর্থিত বাংলা ফন্ট ?

একটি আদর্শ ইউনিকোড সমর্থিত ওপেন টাইপ ফন্ট বাংলা বা ইংরেজির জন্য নির্দিষ্ট হবে এমনটা ভাবা ঠিক নয়। কেননা, ইউনিকোডের মূল নীতিই হলো আন্তর্জাতিকীকরণ। বরং এমন যদি একটি ফন্ট হতো যার প্রতিটি কোড পয়েন্টে ইউনিকোডের সঠিক বর্ণটির টাইপ্রোগ্রাফি আছে এবং সমস্ত গ্রিফ সঠিকভাবে আছে— তবে সেটিই হতো আদর্শ ইউনিকোড সমর্থিত ফন্ট। অর্থাৎ যে একক ফন্টে শুধু কী-বোর্ড পরিবর্তন করেই ইউনিকোড সাপোর্টকৃত সব ভাষা লেখা যায় সেটিই হলো আদর্শ ইউনিকোড সমর্থিত ফন্ট। বাংলার ক্ষেত্রেও এটি একইভাবে প্রযোজ্য। কিন্তু ফন্ট একটি বাণিজ্যিক পণ্য, তাই সব ফন্টে সব ভাষার লিপি সমন্বিত করা সস্তব নয়। তাছাড়া সব বর্ণ দিয়ে যদি ফন্ট তৈরি করা হয়, তবে ফন্ট ফাইলটির সাইজ কয়েক মেগাবাইটে পরিণত হবে এবং ফন্টটির পোর্টেবিলিটি অনেক কমে যাবে। তারপরও ভালো বাংলা ফন্টগুলো এমন হওয়া উচিত, যাতে বাংলাভাষাভাষীরা বাংলা ছাড়াও ইংরেজি সহ অন্যান্য বহুল প্রচলিত ভাষা লিখতে পারে। আর ফন্টের টাইপ যে ওপেন টাইপ ফন্ট হতে হবে— সেটি এখন আর উল্লেখের অপেক্ষা রাখে না।

Bengali, না Bangla

বাংলাভাষাভাষী মাত্রই Bangla বলতেই বেশি পছন্দ করেন, তা সে পশ্চিমবঙ্গবাসীই হোক আর বাংলাদেশীই হোক। কিন্তু ব্রিটিশ শাসনামলের দুইশত বছর ধরেই ইংরেজরা বাংলাকে Bengali লিখে গেছে তাদের দলিল দস্তাবেজে। আর সেসব তথ্যের কল্যাণে আজ বাংলাভাষী ছাড়া অন্য ভাষাভাষীরা Bangla নয়, Bengali-ই জানে বাংলাকে। আর যেহেতু ইউনিকোডের জন্ম আন্তর্জাতিকীকরণে, তাই সারাবিশ্বের লোক যে নামে চেনে বাংলাকে, ইউনিকোড সেই Bengali নামেরই পক্ষপাতী। অবশ্য ইউনিকোডে অন্তর্ভুক্তির পূর্বে যদি এই সংশোধনী দেয়া হতো— তাহলেও হয়তো Bengali না হয়ে Bangla-ই হতো আমাদের ভাষার নাম। কিন্তু, যেহেতু ইউনিকোড অন্তর্ভুক্তির পর আর কোনো পরিবর্তন করা যায় না, তাই Bengali পরিবর্তন করে তাকে Bangla করা আর সস্তব নয়। তবে ইউনিকোডে Bangla বললে তা Bengali-ই ধরে নেয়। ফলে কেউ Bengali ব্যবহার করতে না চাইলেও সমস্যা নেই।

	098	099	09A	09B	09C	09D	09E	09F
0		ঐ	ঊ	ঋ	ঌ		ঙ	ঞ
1		ঔ	ঘ		ঙ		ঞ	ট
2		ং	ঢ	ঢ়	ণ	ণ়	ত	ত্
3		ং	ঙ	ণ		ত	ত্	ত্
4			ঙ	ত				ত্
5		অ	ক	খ				খ
6		আ	খ	দ	শ		০	৩
7		ই	গ	ধ	ষ	ে	ৌ	।
8		ঐ	ঘ	ন	স	ৈ	২	৮
9		উ	ঙ		হ		৩	০
A		ঊ	চ	প			৪	৩
B		ঋ	ছ	ফ	ৌ		৫	
C		ঌ	জ	ব	ৌ	ড	৬	
D			ঝ	ভ	া	ঢ	৭	
E			ঞ	য	া		ঢ	
F		এ	ট	ষ	ি	য়	৯	

ইউনিকোডে বাংলার কোড ব্লক

ইউনিকোড কনসোর্টিয়াম

ইউনিকোড ক্যারেক্টারগুলোর আচরণ ও পারস্পরিক সম্পর্ক নির্ধারণের জন্য যে সংস্থাটি কাজ করে— সেটি হলো ইউনিকোড কনসোর্টিয়াম। তাছাড়া যারা ইউনিকোড ক্যারেক্টার সেটের প্রয়োগ ঘটাবে, তাদেরকে বিভিন্ন ধরনের টেকনিক্যাল তথ্য প্রদানের দায়িত্বও এই কনসোর্টিয়ামই পালন করে থাকে। আর ইউনিকোড ক্যারেক্টার সেট প্রণয়ন যে তাদেরই কাজ সেটি নিশ্চয়ই আলাদা করে বলার আর প্রয়োজন পড়ে না।

এই কনসোর্টিয়াম ১৯৯১ সালে প্রতিষ্ঠার পর বেশ কয়েক বছর সাদামাটাভাবে তাদের কার্যক্রম পরিচালনা করে। এই সময়ই তারা প্রথমবারের মতো ইউনিকোড স্ট্যান্ডার্ড প্রণয়ন এবং স্ট্যান্ডার্ডটি অন্যদের সামনে তুলে ধরার কাজটি করে। পাশাপাশি ইউনিকোডের প্রয়োগে ও মান নিয়ন্ত্রণেও তারা বেশ উল্লেখ্য সহযোগিতা করে। এই কনসোর্টিয়াম সফটওয়্যার ইন্ডাস্ট্রির বিভিন্ন কোম্পানি ও আন্তর্জাতিক স্ট্যান্ডার্ড প্রণয়নে নিবেদিত গবেষকদের একটি একক প্র্যাটফর্মে নিয়ে আসে। আর এই একত্রিত হওয়ার ফলস্বরূপ জন্ম নেয় 'ইউনিকোড স্ট্যান্ডার্ড'— যা সফটওয়্যারের আন্তর্জাতিকীকরণের প্রথম ধাপ হিসেবে কাজ করে। আর ইউনিকোডের সাথে ISO/IEC 10646 স্ট্যান্ডার্ডটি সম্পূর্ণভাবেই জড়িত। এমনকি ইউনিকোড কনসোর্টিয়াম আইএসও'র ক্যারেক্টার সেটকে পরিবর্তন ও পরিমার্জনের কাজটিও করে থাকে। এই কনসোর্টিয়াম স্ট্যান্ডার্ডের লিয়ার্স হিসেবে ISO/IEC/JTC 1/SC2/WG2-এর সাথে কাজ করে।

বর্তমানে ইউনিকোড কনসোর্টিয়ামের সদস্য হিসেবে শীর্ষস্থানীয় কম্পিউটার পণ্য নির্মাতা, সফটওয়্যার কোম্পানি, ডাটাবেজ নির্মাতা, গবেষণাকেন্দ্র, আন্তর্জাতিক সংস্থা ও বিভিন্ন ধরনের সংগঠনসহ রয়েছে অনেক অগ্রহী ব্যক্তিবর্গ। কনসোর্টিয়ামের ডিরেক্টরসহ অন্যান্য কর্মকর্তা নির্বাচিত হয় বিভিন্ন ধরনের প্রতিষ্ঠান থেকে যারা কম্পিউটিং এপ্লিকেশন ও টেক্সট এনকোডিংয়ে বিপুল পরিমাণে কাজ করেছেন। তাছাড়া কনসোর্টিয়ামের বিভিন্ন ধরনের নীতিমালা রয়েছে— যার মধ্যে পেটেন্ট এবং স্ট্যান্ডার্ডের গ্রহণযোগ্যতার বিষয়গুলোও রয়েছে। ইউনিকোড স্ট্যান্ডার্ড প্রণয়ন, রক্ষণাবেক্ষণ ও মান নিয়ন্ত্রণের জন্য অর্থাৎ ইউনিকোড টেকনিক্যাল কার্যক্রম পরিচালনার জন্য রয়েছে 'ইউনিকোড টেকনিক্যাল কমিটি'। তাছাড়া এই কনসোর্টিয়াম মাঝে মাঝে ইউনিকোড স্ট্যান্ডার্ডে একক বা ব্যক্তিগত পর্যায়ের বিশেষ অবদানের জন্য বিভিন্ন ব্যক্তিকে 'বুলডগ পুরস্কার' প্রদান করে থাকে।

ইউনিকোড কনসোর্টিয়ামের মূল লক্ষ্য

ইউনিকোড কনসোর্টিয়াম কতগুলো সুনির্দিষ্ট লক্ষ্যকে সামনে রেখে কাজ করে। তারমধ্যে কয়েকটি হলো—

১. বর্তমান ও প্রচলিত স্ক্রিপ্টের ভিত্তিতে বর্ণমালার ভাঙার গড়ে তুলে এনকোডিং করা।
২. বর্ণমালার ভাঙারকে সুসজ্জিত ও সুবিদ্যমান করা, বিশেষ করে বিভিন্ন ধরনের প্রতীকের জন্য বিন্যাস সহজ করা।
৩. ডাটার ধারাবাহিকতা রক্ষা ও ডাটাকে সহজে পরিবহনযোগ্য করার জন্য বিভিন্ন ধরনের নিয়ম-নীতি প্রণয়ন করা।
৪. প্রয়োগের জন্য বিভিন্ন মডেল তৈরি করা।
৫. ইউনিকোড স্ট্যান্ডার্ড ও বিভিন্ন আঞ্চলিক মডেলের মধ্যে সম্পর্ক গড়ে তোলা।
৬. অন্যান্য স্ট্যান্ডার্ড করার প্রজেক্ট প্রোটোকল, ডাটা ফরম্যাট ও আদান-প্রদানে ইউনিকোডকে গ্রহণযোগ্য করার জন্য কাজ করা।

ইউনিকোড কনসোর্টিয়ামের নীতিমালা

ক্যারেক্টার এনকোডিং স্ট্যান্ডার্ড পলিসি, পেটেন্ট, জেনারেল প্রাইভেসি পলিসি এবং ট্রেডমার্ক ও লোগো পলিসি শীর্ষক ৪টি ভিন্ন ভিন্ন নীতিমালা রয়েছে ইউনিকোড কনসোর্টিয়ামের। এখানে সংক্ষেপে প্রতিটি সম্পর্কে আলোচনা করা হলো—

১. ক্যারেক্টার এনকোডিং স্ট্যান্ডার্ড পলিসি

অন্যান্য স্ট্যান্ডার্ডের মতোই ইউনিকোড স্ট্যান্ডার্ডও নিয়মিত বৃদ্ধি পাচ্ছে। নিত্যানতন বর্ণ সংযুক্ত হচ্ছে এতে। বিশেষ করে ব্যবহারের প্রয়োজনে ও সুবিধার্থে বিভিন্ন টেকনিক্যাল চিহ্ন থেকে শুরু করে প্রাচীন বর্ণমালাও এতে সংযুক্ত হয়েছে। তাছাড়া বিভিন্ন বর্ণের বৈশিষ্ট্যও পরিবর্তিত ও পরিমার্জিত হয়েছে বিভিন্ন সময়। বিশেষ করে প্রয়োগের সুবিধার্থেই এই পরিবর্তন। তবে, এই পরিবর্তন ইচ্ছেমতো করা সম্ভব নয়। ইউনিকোড স্ট্যান্ডার্ডের যে-কোনো পরিবর্তন বা পরিমার্জন কখন সম্ভব হবে তা স্পষ্ট বলে দেয় এই স্ট্যান্ডার্ড পলিসি। এর প্রধান প্রধান ধারাগুলো হলো—

১. এনকোডিং : একবার একটি বর্ণ এনকোড করা হয়ে গেলে তার অবস্থান পরিবর্তন কিংবা মুছে ফেলা সম্ভব নয়।
২. নাম : একবার একটি বর্ণ এনকোড করা হয়ে গেলে এর নাম পরিবর্তন করা যাবে না।
৩. নরমলাইজেশন : একবার একটি বর্ণ এনকোড করা হয়ে গেলে এর ক্যানোনিক্যাল কনস্ট্রাক্ট ক্লাস ও ডিকম্পোজিশন ম্যাপিং পরিবর্তন করা যাবে না। যদি না তা ডিস্ট্যাংগুইশ নরমলাইজেশন হয়।
৪. আইডেন্টিটি : একবার একটি বর্ণ এনকোড করা হয়ে গেলে তার বৈশিষ্ট্য পরিবর্তন করা যাবে, যদি তাতে ঐ বর্ণের মূল আইডেন্টিটি পরিবর্তন না হয়।
৫. শ্রোপার্টি ভ্যালু : ইউনিকোড ক্যারেক্টার ডাটাবেইজের কোনো নির্দিষ্ট শ্রোপার্টি ভ্যালু পরিবর্তন করা যাবে না।

২. পেটেন্ট পলিসি

ইউনিকোড কনসোর্টিয়ামের পেটেন্ট পলিসিটি এনসি'র পেটেন্ট পলিসির উপর ভিত্তি করেই তৈরি করা হয়েছে। যে পরিবর্তনগুলো এতে করা হয়েছে, তা মূলত পলিসিটিকে ইউনিকোড কনসোর্টিয়াম ইউনিকোড স্ট্যান্ডার্ড ও ইউনিকোড টেকনিক্যাল রিপোর্টের উপযুক্ত করার জন্যই করা হয়েছে। এমনকি এনসিতে যেখানে 1.2.11 অনুচ্ছেদে পেটেন্ট পলিসি বর্ণনা করা হয়েছে ইউনিকোডেও সেই একই নম্বরে এই পলিসি রাখা হয়েছে সহজবোধ্য ও পার্থক্য চিহ্নিত করার জন্য।

১.২.১১ একটি স্ট্যান্ডার্ড প্রস্তাব করে ড্রাফট করার ব্যাপারে ইউনিকোডের কোনো বিধি নিষেধ নেই। তবে যদি কোনো প্রস্তাবে পেটেন্টের ঘোষণা আসে তবে ১.২.১১.১ থেকে ১.২.১১.৪ ধারা অনুযায়ী তা পালিত হবে।

১.২.১১.১— পেটেন্টধারীর কাছ থেকে লিখিত বক্তব্য আসতে হবে।

১.২.১১.২— সেই বক্তব্যের যাবতীয় তথ্য কনসোর্টিয়ামের কাছে সংরক্ষণের জন্য সরবরাহ করতে হবে।

১.২.১১.৩— পেটেন্টধারীর কাছ থেকে একটি ঘোষণা স্ট্যান্ডার্ডের সাথে সংযুক্ত থাকতে হবে।

১.২.১১.৪— চিহ্নিত পেটেন্টের জন্য কোনো লাইসেন্স প্রয়োজন হলে তার দায়ভার ইউনিকোড কনসোর্টিয়াম বহন করবে না।

৩. জেনারেল প্রাইভেসি পলিসি

ইউনিকোড কনসোর্টিয়াম একটি সাধারণ গোপনীয়তা নীতিমালা মেনে চলে। তার মধ্যে রয়েছে—

১. ইউনিকোডের ওয়েবসাইট ব্যবহারকারীদের তারা কোনোভাবে চিহ্নিত করার চেষ্টা করে না। কিংবা কুকি ব্যবহার করে না। ওয়েবসাইটে প্রবেশ সংক্রান্ত যে তথ্য সংগ্রহ করা হয়, তা শুধু সার্ভার এডমিনিস্ট্রেশন ও সাইটের কনটেন্ট পরিবর্তন ও পরিমার্জনের ক্ষেত্রে ব্যবহৃত হয়।
২. ই-মেইল কিংবা অন্যভাবে যেসব তথ্য তারা সংগ্রহ করে তা শুধুমাত্র স্ট্যান্ডার্ডের উন্নয়নে ও অন্যান্য ডকুমেন্ট তৈরিতে ব্যবহৃত হয়।
৩. ইউনিকোড টেকনিক্যাল কমিটির কাছে অংশগ্রহণকারীকে চিহ্নিত করতে সক্ষম এমন ডাটা থাকতে পারে।
৪. পাবলিক মেইলিং লিষ্টের সংগ্রহশালা সাধারণের দেখার জন্যই ব্যবহৃত হবে এবং তাতে নাম ও ই-মেইলের মতো সাধারণ তথ্যই থাকবে।
৫. সদস্য প্রতিষ্ঠানের প্রতিনিধিকে চিহ্নিত করা হয় মূলত সদস্যদের জন্য সংরক্ষিত তথ্য প্রদানের ক্ষেত্রে নিশ্চয়তার জন্যই। কিন্তু সেসব তথ্য কোনো সদস্য যদি বাণিজ্যিকভাবে ব্যবহার করে তবে তা ইউনিকোড কনসোর্টিয়ামের নীতিবিরুদ্ধ হবে।
৬. আন্তর্জাতিক ইউনিকোড সম্মেলন এই নীতির আওতাভুক্ত নয়।

৪. ট্রেডমার্ক ও লোগো পলিসি

ইউনিকোডের লোগো ও ট্রেডমার্ক সঠিকভাবে ব্যবহারের জন্য ইউনিকোড কনসোর্টিয়াম একটি নীতিমালা প্রণয়ন করেছে। আর তারমধ্যে রয়েছে ইউনিকোডের লোগো ব্যবহারের ক্ষেত্রে ইউনিকোড কনসোর্টিয়ামের লিখিত অনুমতি নেবার বিষয়টি। তবে অবাণিজ্যিক ক্ষেত্রে এই লোগো ব্যবহারের জন্য কোনো ফি দিতে না হলেও বাণিজ্যিক ক্ষেত্রে ২ বছর মেয়াদী লাইসেন্স সংগ্রহ করতে হতে পারে। এবং সব ক্ষেত্রেই লোগোর লিঙ্ক হিসেবে <http://www.unicode.org> সাইটটি রাখতে হবে।

১. ট্রেডমার্কের গুরুত্ব : Unicode® দ্বারা ইউনিকোড কনসোর্টিয়াম ও Unicode™ দ্বারা ইউনিকোড স্ট্যান্ডার্ড বোঝানো হয়।

২. পণ্যের নাম : পণ্যের নামের সাথে ইউনিকোডের নাম সম্পৃক্ত করা যাবে না।

৩. প্যাকেজিং ও বিজ্ঞাপন : কোনো পণ্যের ইউনিকোড ভার্শন সংযুক্ত হলে বা ইউনিকোড সপোর্ট যুক্ত হলে তাকে বলতে হবে XYZ for Unicode™ Standard কিংবা XYZ for Unicode™। কিন্তু কোনো অবস্থাতেই XYZ Unicode কিংবা Unicode XYZ বা XYZ/Unicode বলা যাবে না।
৪. যথাযথ নাম কিংবা প্রতীক ব্যবহার : কখনই শুধু Unicode লেখা যাবে না। সব সময়ই Unicode® কিংবা Unicode™ লিখতে হবে যাতে বোঝা যায় কাকে বোঝানো হচ্ছে।
৫. ট্রেডমার্কের যথাযথ ব্যবহার— Unicode™ কিংবা Unicode® ব্যবহার করলে ফুটনোট অবশ্যই
Unicode™— "Unicode is a trademark of Unicode, Inc."
Unicode®— "Unicode is a registered trademark of unicode . Inc"
The Unicode Logo "Unicode and Unicode Logo are trademarks of Unicode, Inc."
৬. শব্দ সংক্ষেপ না করা— ইউনিকোড কোনোভাবেই সংক্ষিপ্ত করা যাবে না।
৭. ইউনিকোড লোগো— ইউনিকোড লোগো ইউনিকোড কনসোর্টিয়ামের নিজস্ব সম্পত্তি এবং কোনো অবস্থাতেই ইউনিকোড কনসোর্টিয়ামের কাছ থেকে লাইসেন্স সংগ্রহ ব্যতীত তা ব্যবহার করা যাবে না।

ইউনিকোড কনসোর্টিয়ামের সদস্য পদ

ইউনিকোড কনসোর্টিয়ামের সদস্য হতে তেমন কোনো যোগ্যতার প্রয়োজন নেই। যে কেউই এর সদস্য হতে পারবে। কেননা, ইউনিকোড কনসোর্টিয়াম চলেই তার সদস্যদের দ্বারা— যারা করপোরেট লেভেল থেকে হতে পারে, গবেষক হতে পারে, শিক্ষাপ্রতিষ্ঠান হতে পারে, শিল্প খাতের কোনো সংগঠন বা সংস্থা হতে পারে— এমনকি ব্যক্তিবিশেষও হতে পারে। এই কনসোর্টিয়ামের সদস্য হবার মূল মন্ত্র হলো ইউনিকোড স্ট্যান্ডার্ডের প্রয়োগ, পরিবর্ধন, রক্ষণাবেক্ষণ সর্বোপরি এর প্রচারে সহযোগিতা করা। তবে এই কনসোর্টিয়াম সব সময়ই চায় যে আইটি সংশ্লিষ্ট ব্যক্তিবর্গই এর সদস্যপদ গ্রহণ করুক— যারা ইউনিকোড কনসোর্টিয়ামের লক্ষ্য বাস্তবায়নে সহযোগিতাসহ বড় ধরনের ভূমিকা পালনে ব্রত হবে। ইউনিকোড কনসোর্টিয়ামের সদস্যদের সবসময়ই সুযোগ থাকে ইউনিকোড স্ট্যান্ডার্ড প্রণয়নে নিজেদের মতামত উপস্থাপনের। আর আইটি ব্যক্তিবর্গের পাশাপাশি সাধারণ ব্যবহারকারীরাও ইচ্ছা করলে এই কনসোর্টিয়ামের সদস্যপদ পেতে পারে।

ইউনিকোড কনসোর্টিয়ামে মোট ৫ ধরনের সদস্যপদ রয়েছে। পূর্ণ সদস্য, এসোসিয়েট সদস্য, বিশেষজ্ঞ সদস্য, একক সদস্য ও লিয়াজো সদস্য।

পূর্ণ সদস্য

পূর্ণ সদস্য কেবলমাত্র একটি ব্যবসায়িক প্রতিষ্ঠান কিংবা সংগঠন হতে পারে এবং তারা ইউনিকোড স্ট্যান্ডার্ড প্রণয়নে ভূমিকা রাখতে পারে। কেননা, যে-কোনো সভায় ভোট প্রদানের ক্ষমতা পায় শুধু এই ধরনের সদস্যরা। আর এভাবেই পূর্ণ সদস্যদের প্রত্যক্ষ সহযোগিতায় ইউনিকোড স্ট্যান্ডার্ডের পরিবর্ধন ও পরিমার্জন সাধিত হয়ে আসছে। পূর্ণ সদস্যদের বার্ষিক ফি বছরে ১২ হাজার মার্কিন ডলার। বর্তমানে ভারত ও পাকিস্তান সরকারসহ ১০টি নামী দামী প্রতিষ্ঠান পূর্ণ সদস্য হিসেবে কার্যক্রম চালিয়ে যাচ্ছে।

এসোসিয়েট সদস্য

এসোসিয়েট সদস্য আমন্ত্রণের ভিত্তিতে শুধুমাত্র সেইসব প্রতিষ্ঠান ও সংগঠনের জন্য প্রয়োজন যারা কেবল প্রকাশ্যে ইউনিকোড কনসোর্টিয়ামের সঙ্গে থাকতে চাইলেও ইউনিকোড স্ট্যান্ডার্ড প্রণয়নের বিষয়ে সরাসরি অবদান রাখতে চায় না। এসোসিয়েট সদস্যপদ প্রাপ্ত প্রতিষ্ঠানের বা সংগঠনের দু'জন ইউনিকোড স্ট্যান্ডার্ডের যাবতীয় তথ্য অবধি আনাগোনা করতে পারেন। এ ধরনের সদস্যদের বার্ষিক ফি লাভজনক প্রতিষ্ঠানের জন্য বছরে ২ হাজার মার্কিন ডলার ও অলাভজনক সংগঠনের জন্য বছরে ১২ শত মার্কিন ডলার। বর্তমানে প্রায় ২৮টি প্রতিষ্ঠান এ ধরনের সদস্য হিসেবে রয়েছে।

বিশেষজ্ঞ সদস্য

এটি শুধুমাত্র ব্যক্তিগত পর্যায়ে যারা ইউনিকোড স্ট্যান্ডার্ড কাজ করতে অগ্রহী সেইসব বিশেষজ্ঞদের জন্য প্রযোজ্য। এরা কনসোর্টিয়ামের বিভিন্ন টেকনিক্যাল কর্মকাণ্ডে যোগদান সহ সব ধরনের তথ্য অবধি প্রবেশ করতে পারে। তবে তাদের ভোটাধিকার ক্ষমতা না থাকলেও তারা ইউনিকোড স্ট্যান্ডার্ড ডেভেলপমেন্টে সরাসরি অংশগ্রহণ করতে পারবে। এ ধরনের সদস্যদের বার্ষিক ফি ৬০০ মার্কিন ডলার।

একক সদস্য

যে-কোনো সাধারণ ব্যক্তি বা ব্যবহারকারীদের যদি ইউনিকোডের ব্যাপারে অগ্রহী হন এবং কনসোর্টিয়ামের সদস্য হতে চান— তারা এই ধরনের সদস্য হতে পারবেন। ফলে যে-কোনো পূর্ণাঙ্গ তথ্য তারা পেতে পারলেও যেসব বিষয়ে তখনো কাজ চলছে— সেসব তথ্য তারা পাবেন না। একক সদস্যদের বার্ষিক ফি মাত্র ১২০ মার্কিন ডলার।

লিয়াজো সদস্য

এটি শুধুমাত্র আমন্ত্রণের ভিত্তিতে বিনা সদস্য ফিতে শুধুমাত্র সেইসব সংগঠনের জন্য প্রযোজ্য— যারা কনসোর্টিয়ামের টেকনিক্যাল বিষয়গুলোতে সক্রিয় ভূমিকা রাখতে ইচ্ছুক। সাধারণ লিয়াজো সদস্যপদ ভুক্ত প্রতিষ্ঠানের মাত্র একজন প্রতিনিধি যাবতীয় কর্মকাণ্ডে অংশগ্রহণ করতে পারে। এবং এই প্রতিনিধি তার প্রতিষ্ঠান ও ইউনিকোড কনসোর্টিয়ামের মধ্যে সেতু হিসেবে কাজ করেন।

সদস্যপদের জন্য আবেদন

ইউনিকোড কনসোর্টিয়ামের সদস্যপদ প্রাপ্তির জন্য অনলাইনে তাদের নির্ধারিত ফরম পূরণের মাধ্যমে আবেদন করতে হয়। আবেদনপত্রে সদস্যপদের ধরন, আবেদনকারীর নাম, প্রতিষ্ঠান, ঠিকানা, শহর-পোস্ট কোড, দেশ, ফোন নম্বর, ফ্যাক্স ও ই-মেইল উল্লেখ করতে হয়। পাশাপাশি ফি প্রদানের জন্য Unicode, Inc বরাবর মার্কিন ডলারের চেক, অথবা ক্রেডিট কার্ডের মাধ্যমে প্রদানের ক্ষেত্রে ক্রেডিট কার্ডের ধরন (AMEX, JCB, MasterCard, Visa), ক্রেডিট কার্ডের নম্বর, ক্রেডিট কার্ডধারীর নাম ও মেয়াদ উত্তীর্ণের তারিখ উল্লেখ করতে হবে। আর ফ্যাক্স কিংবা পত্র মারফত আবেদনের ক্ষেত্রে অবশ্যই স্বাক্ষর দিতে হবে।

যোগাযোগের মাধ্যম

শুধু ইন্টারনেটের ওয়েবসাইট নয়— ইউনিকোড কনসোর্টিয়ামের সাথে যোগাযোগের জন্য একাধিক মাধ্যম রয়েছে।

ওয়েবসাইট : <http://www.unicode.org/> অথবা <ftp://ftp.unicode.org/>

চিঠি লেখার ঠিকানা :

The Unicode Consortium
P.O. Box 391476
Mountain View, CA 94039-1476
U.S.A

ফোন নম্বর : +1-650-693-3921

ফ্যাক্স : +1-650-693-3010

কুরিয়ার করার ঠিকানা :

The Unicode Consortium
1065 L'Avenida Street
Microsoft Building 5
Mountain View, CA 94043
U.S.A

ইউনিকোড কনসোর্টিয়ামের সম্মেলন

ইউনিকোড নিয়ে যারা কাজ করতে চান, ইউনিকোড কনসোর্টিয়ামের বিভিন্ন সম্মেলন তাদের জন্য খুবই কার্যকরী ও ফলপ্রসূ একটি মাধ্যম। কেননা, এখানে তারা যেমন তাদের নিজস্ব মতামত তুলে ধরতে পারবেন, তেমন অন্যান্য ইউনিকোড বিশেষজ্ঞদের সাথেও তারা আলোচনার একটি সুযোগ পাবেন। তাছাড়া অংশগ্রহণকারীরা ইউনিকোডের কাজের

সর্বশেষ তথ্য জানারও একটি সুযোগ পাবেন। Internationalization & Unicode Conferences নামে পরিচিত এই সম্মেলনগুলো বিভিন্ন সময়ে বিশ্বের বিভিন্ন দেশে অনুষ্ঠিত হয়। তাই এ সম্পর্কে সর্বশেষ তথ্য জানতে ইউনিকোডের ওয়েবসাইটই আর্দশ স্থান।

আর এই সম্মেলন সরাসরি ইউনিকোড কনসোর্টিয়ামের আয়োজন করে না, বরং এটি আয়োজন করে থাকে 'গ্লোবাল মিটিং সার্ভিসেস' নামক একটি আন্তর্জাতিক ইভেন্ট ম্যানেজমেন্ট কোম্পানি।

ইউনিকোড পাবলিক ই-মেইল লিস্ট

যাদের পক্ষে ইউনিকোড কনসোর্টিয়ামের সদস্য হওয়া সম্ভব নয়, অথচ ইউনিকোডের সর্বশেষ তথ্য পেতে চান তারা ইউনিকোডের পাবলিক ই-মেইল লিস্টের সদস্য হতে পারেন। এতে করে ইউনিকোডের সমস্ত অফিসিয়াল নিউজ পাওয়া যাবে এবং ইউনিকোড সম্পর্কে অন্যান্য সদস্যদের মন্তব্য জানা যাবে ই-মেইলের মাধ্যমে। এই অফিসিয়াল মেইল লিস্টের সদস্য হতে চাইলে ecartis@unicode.org-এ একটি ফাঁকা ই-মেইল পাঠাতে হবে যার Subject-এ লিখতে হবে Subscribe Unicode। আর Unsubscribe করতে চাইলে যেতে হবে ইউনিকোডের ওয়েবসাইটে। এই মেইল লিস্টে নিজের মতামত পাঠাতে চাইলে ই-মেইল করতে হবে unicode@unicode.com ঠিকানা।

তাছাড়া ইয়াহু গ্রুপসেও একটি ব্যক্তিগত পর্যায়ের মেইল গ্রুপ রয়েছে Unicode বিষয়ক— যাতে ইউনিকোডের অফিসিয়াল মেইল লিস্টের আগের সমস্ত তথ্য HTML ফরম্যাটে সংরক্ষিত আছে। এই সাইটটি হলো—[http://groups.yahoo.com/group/ unicode/](http://groups.yahoo.com/group/unicode/)

বহুল প্রচলিত কিছু প্রশ্ন

ইউনিকোড নিয়ে আমাদের অনেকের মনেই অনেক রকম প্রশ্ন রয়েছে। আবার বিভিন্ন ধরনের কলাম ও প্রতিবেদন পড়ে অনেকের ধারণাই সম্পূর্ণ ভ্রান্ত। এই ভ্রান্ত ধারণা শুধু যে আমাদের দেশেই— তা নয়, বরং বিশ্বব্যাপী এসব ভুল ধারণা মানুষের মনে রয়ে গেছে। তারই কিছু এখানে আলোচনা করা হলো।

ইউনিকোড সফটওয়্যার কিংবা ইউনিকোড ফন্ট কোথায় পাওয়া যাবে ?

—ইউনিকোড স্ট্যান্ডার্ড যেমন কোনো প্রোগ্রাম নয়, তেমনি এটি কোনো ফন্টও নয়। এটি শুধুমাত্র এসকি'র মতো একটি ক্যারেক্টার এনকোডিং সিস্টেম। অবশ্য ইউনিকোড ব্যবহার করে সফটওয়্যার তৈরি ইতোমধ্যেই যেমন শুরু হয়েছে, তেমনি ইউনিকোড সমর্থিত ফন্টও রয়েছে প্রচুর। বিশেষত একাধিক ভাষা নিয়ে কাজ করতে চাইলে ইউনিকোডের কোনোই বিকল্প নেই।

ইউনিকোড কোন কোন ভাষা সাপোর্ট করে ?

— এটি খুবই জটিল একটি প্রশ্ন। ল্যাটিন বর্ণমালা ভিত্তিক একাধিক ভাষার পাশাপাশি বর্তমানে এটি গ্রিক, সিরিলিক, আর্মেনিয়ান, হিব্রু, আরবি, সিরিয়াক, থান্না, দেবনাগরী, বাংলা, গুরুমুখী, ওড়িয়া, তামিল, তেলুগু, কর্ণাটক, মালয়লাম, সিংহলী, খাইলাও, তিব্বতী, মায়ানমার, জিওর্জিয়ান, হানগুল, ইথিওপিক, চেরোকী, কানাডিয়ান-এবংওরিজিন্যাল সিলেবিকস, ওগহাম, রুনিক, খেমের, মঙ্গোলীয়, হান (জাপানি, চীনা, কোরিয় ইউইওফাফ), হিরাগানা, কাটাকানা, বোপোমোফো এবং ইয়াই লিপি সাপোর্ট করে।

ইউনিকোড লিপি না ভাষা কোনটি এনকোডিং করে ?

— ইউনিকোড কখনোই ভাষা নিয়ে কাজ করে না, বরং ভাষার লিখিত রূপ (লিপি) নিয়ে কাজ করে। তাই এতে বর্ণ এনকোডিং করার সময় লিপি থেকেই বর্ণ নেয়া হয়। যেমন— ল্যাটিন বর্ণ একাধিক ভাষায় ব্যবহৃত হয়। তাই ইউনিকোড শুধুমাত্র বর্ণটিকে লিপি অনুযায়ী এনকোডিং করেছে। একইভাবে সিরিলিক, আরবী, ইথিওপিক, দেবনাগরী ইত্যাদি লিপিও একাধিক ভাষায় ব্যবহৃত হয় বলে সেভাবেই এনকোডিং করা হয়েছে।

আমাদের বর্ণমালা ভুল ক্রমে এনকোডিং করে সাজানো হয়েছে বলে বিন্যাস বা সটিং ঠিকমতো হচ্ছে না! কী করা যাবে ?

—এটি প্রায় অসম্ভব একটি দাবী। আর এখানে একটি ভুল বোঝাবুঝিও রয়েছে। ভাষাগতভাবে শব্দ বিন্যাস কখনোই কোড পয়েন্টের ভিত্তিতে করা হয় না। এভাবে করা হলে ইংরেজির ক্ষেত্রেও ভুল সটিং হতে পারে। ভাষাগতভাবে সটিংয়ের নিয়ম হলো— বর্ণ বা বর্ণক্রমের জন্য একাধিক স্তরের মান নির্ধারণ করা এবং প্রতিটি স্তরে সেই মান তুলনা করে সটিং করা। এটি করার জন্য একাধিক এলগরিদমও রয়েছে। তার মধ্যে Unicode Collation Algorithm (UCA) ইউনিকোড বর্ণমালার সাথে খুব ভালো কাজ করে— কেননা, এর একটি ডিফল্ট মানের তালিকা রয়েছে এবং প্রয়োজন বোধে স্থানীয় নিয়ম প্রয়োগের জন্য টেইলারিং মেকানিজম প্রয়োগেরও সুযোগ রয়েছে।

কোন ধরনের ফন্ট ইউনিকোড সাপোর্ট করে ?

— যেসব ফন্ট ইউনিকোড বর্ণ থেকে গ্লিফ তৈরির পূর্ণাঙ্গ ম্যাপ CMAP রয়েছে সেসব ফন্টই ইউনিকোড সাপোর্ট করে। নতুন Type1 ও TrueType ফন্টে এই CMAP থাকার ফলে, তারাও ইউনিকোড সাপোর্ট দিতে পারে। তবে Open Type ফন্ট, এপলের Apple Advanced Typography ফন্ট কিংবা গ্রাফাইট টেবিল সংবলিত Graphite TrueType ফন্টও ইউনিকোড সাপোর্ট দেয় সবচেয়ে ভালো।

ইউনিকোডের ক্ষেত্রে ল্যান্ডমার্ক ট্যাগিং কি অত্যাৱশ্যক ?

— যেহেতু ইউনিকোডের পূর্ণাঙ্গ টেবিলে একটি ভাষার বর্ণের জন্য একটিই কোড পয়েন্ট থাকে তাই অধিকাংশ ক্ষেত্রেই ল্যান্ডমার্ক ট্যাগ অত্যাৱশ্যক নয়। আবার, কিছু কিছু ক্ষেত্রে ল্যান্ডমার্ক ট্যাগ বেশ গুরুত্বপূর্ণ।

Digraph ও Ligature-এর মধ্যে পার্থক্য কী ?

— Digraph ও Ligature দু'টোই দুটো গ্লিফের সমন্বয়ে তৈরি হয়। Digraph-এ গ্লিফ দুটো আলাদাই থাকে, তবে খুব কাছাকাছি স্থাপিত হয়। কিন্তু Ligature-এর ক্ষেত্রে গ্লিফ দুটো ভেঙে একত্রিত করে একটি গ্লিফ হিসেবে দেখানো হয়।

আমাদের কী ইউনিকোড ব্যবহার করতেই হবে ? আমরা তো নিজস্ব এনকোডিং সিস্টেম ব্যবহার করতে পারি।

— ইউনিকোড হলো একটি ইন্ডাস্ট্রি স্ট্যান্ডার্ড— যা সব প্রসিটিফর্মের, সব স্ট্যান্ডার্ডে ব্যবহৃত হচ্ছে এবং বিশ্বের প্রায় সমস্ত নামি-দামি কোম্পানি ব্যবহার করছে। ফলে শুধু বাংলা কিংবা বাংলা-ইংরেজি মিলিয়ে কাজ করতে এবং সেই কাজ সর্বত্র গ্রহণযোগ্য করে তুলতে ইউনিকোডই সর্বোত্তম ও গ্রহণযোগ্য এনকোডিং সিস্টেম।

বিভিন্ন প্রসিটিফর্ম বা বিভিন্ন এপ্লিকেশনে যেহেতু ইউনিকোড সাপোর্ট নেই। তাহলে এই মুহূর্তে আমাদের ইউনিকোড সাপোর্ট কি খুবই প্রয়োজন ?

— উইন্ডোজসহ অনেক সফটওয়্যারই বর্তমানে ইউনিকোড সাপোর্ট করে। যেগুলোতে এখনো ইউনিকোড সাপোর্ট নেই, সেগুলোও ২-১ বছরের মধ্যে ইউনিকোড সাপোর্ট দেবে। অতএব, আমাদেরও দেরি না করে এখনই ইউনিকোড ব্যবহার করা উচিত।

ইউনিকোড সাপোর্ট করে এমন অনেক প্রোগ্রামই বাংলা গ্লিফ ঠিকমতো দেখায় না ?

— এ সমস্যাটা হয় মূলত বিভিন্ন আরামিয়াক যৌগিক লিপির ক্ষেত্রে। যৌগিক লিপি সাপোর্ট দেয়ার জন্য অপারেটিং সিস্টেম কিংবা এপ্লিকেশনে অতিরিক্ত স্তরের প্রয়োজন হয়— যা অনেক সফটওয়্যারেই নেই। তবে সেসব সফটওয়্যারের পরবর্তী ভার্সনে সেই সমস্যার সমাধান দেয়া হয়েছে বা হচ্ছে। অবশ্য সাময়িকভাবে প্রাইভেট ব্লকে যুক্তাক্ষরের জন্য সাপোর্ট দিয়ে কাজ চালানো যেতে পারে— যা ইউনিকোড সমর্থিত নয়।

ইউনিকোডে আমরা আমাদের '৫' সহ অন্যান্য কিছু বর্ণ অন্তর্ভুক্ত করতে চাই। এজন্য কি আমাদের পূর্ণ সদস্য হতেই হবে ?

— ইউনিকোডে বর্ণ অন্তর্ভুক্তির জন্য ইউনিকোড কনসোর্টিয়ামের সদস্য হবার কোনো প্রয়োজন নেই। এই কনসোর্টিয়াম মূলত ইন্ডাস্ট্রি স্ট্যান্ডার্ড নিয়ে কাজ করে, যার মধ্যে ভারত ও পাকিস্তান সরকার ছাড়া আর কোনো দেশ বা সরকার নেই। অথচ, বর্তমান সময়ের প্রচলিত অধিকাংশ ভাষার লিপিই রয়েছে ইউনিকোডে।

ইউনিকোডের বাংলা বিষয়ক সব কাজে ভারতের প্রাধান্যের কারণ কী ?

— ভারত সরকার ইউনিকোড কনসোর্টিয়ামের একজন পূর্ণ সদস্য এবং UTC সহ বিভিন্ন প্রকল্পে সরাসরি জড়িত। অন্যভাবে বলতে গেলে, ভারত সরকারের উদ্যোগেই বাংলা ইউনিকোড স্ট্যান্ডার্ডভুক্ত হয়েছে। তবে ইউনিকোডে কোনো বর্ণ অন্তর্ভুক্ত করতে হলে যে কনসোর্টিয়ামের পূর্ণ সদস্য হতেই হবে, এ ধারণাটি সত্যি নয়।

ইউনিকোডে যে বাংলা ব্লক রয়েছে, সেটা তো বাংলাদেশী বাংলা নয়। আমরা কি বাংলাদেশের বাংলার জন্য আলাদা কোড ব্লক চাইতে পারি না ?

— ইউনিকোড কখনোই ভাষা নিয়ে কাজ করে না। কাজ করে লিপি নিয়ে। সেজন্যই একাধিক দেশের ভাষা ইংরেজি বা আরবী হওয়া সত্ত্বেও এই বর্ণগুলো ইউনিকোডে একবারই রয়েছে। একইভাবে দুই বাংলার ভাষা আলাদা হলেও বর্ণমালা একই বলে, আলাদা কোনো কোড পয়েন্ট আমরা পেতে পারি না।

ইউনিকোডে বাংলার ব্লকে বেশ অনেক ফাঁকা জায়গা আছে। সেখানে কিছু যুক্তাক্ষর রাখতে অসুবিধা কি ?

— 'কিন্তু কেন যুক্তাক্ষর রাখা উচিত নয় ?' অংশে বিষয়টি ব্যাখ্যা করা হয়েছে।

ইউনিকোড যেভাবে বদলে দেবে আন্তর্জাতিক কম্পিউটিংয়ের রূপরেখা

ইউনিকোড ভুক্ত হবার ফলে এবং অন্যান্য ভাষার কী-বোর্ডের মতো বাংলার স্ট্যান্ডার্ড কী-বোর্ড প্রণীত হলে বাংলা কম্পিউটিং হবে ইংরেজির মতোই। তখন, আলাদা করে কী-বোর্ড কিংবা ফন্ট পরিবর্তন করার প্রয়োজন হবে না, বরং অপারেটিং সিস্টেমের

ল্যান্ডুয়েজ সেটিং কিংবা স্রেফ দেশ বদলে দিলেই অর্থাৎ LOCAL চলকটির মান বদলে দিলেই অপারেটিং সিস্টেম ভাষা হিসেবে বাংলা, কী-বোর্ড হিসেবে স্ট্যান্ডার্ড বাংলা কী-বোর্ড এবং ফন্ট হিসেবে বাংলার জন্য বিশেষায়িত ফন্টগুলোই কেবল ব্যবহার হবে। অর্থাৎ পরিপূর্ণভাবেই উপভোগ করা যাবে বাংলায় কম্পিউটিং। আর বিশ্বের সব ভাষাতেই যাতে কম্পিউটিং উপভোগ করা যায় সে লক্ষ্যেই কাজ শুরু করেছিল ইউনিকোড এবং এখনো সেভাবেই কাজ করে যাচ্ছে।

শেষ কথা

কম্পিউটারে পরিপূর্ণভাবে বাংলা ব্যবহার করতে চাইলে ইউনিকোডই যে একমাত্র সমাধান, এ বিষয়ে এখন আর কোনো বিতর্ক নেই। অহেতুক হাজার হাজার কী-বোর্ড লেআউট প্রণয়নের হাস্যকর প্রচেষ্টা বাদ দিয়ে সঠিকভাবে ইউনিকোড স্ট্যান্ডার্ড মেনে চলে এমন কী-বোর্ড প্রণয়ন করা উচিত যেটি হবে বাস্তবসম্মত এবং প্রযুক্তিগতভাবে গ্রহণযোগ্য।

অত্যন্ত আনন্দের ব্যাপার এটি যে, ইউনিকোড কনসোর্টিয়ামের পূর্ণ সদস্য হবার বিষয়ে আমাদের কর্তব্যাক্ষিরা ঐকমত্যে সক্ষম হয়েছেন। কিন্তু শুধু সদস্য হওয়াটাই সব নয়। কেননা, ইউনিকোড বাংলা লিপির বর্তমান অবস্থান বা বাংলা ক্যারেক্টার সেট কোনোটাই অসম্পূর্ণ তো নয়ই, বরং এক কথায় 'পারফেক্ট'। অতএব, ১২ হাজার মার্কিন ডলার খরচ করে কনসোর্টিয়ামের পূর্ণ সদস্য হয়ে নতুন বাংলা ক্যারেক্টার সেট প্রস্তাব করে অহেতুক টাকা ও সময় দুটোই অপচয় করার কোনো মানে হয় না। কেননা, ইউনিকোড স্ট্যান্ডার্ড থেকে কোনো কিছু মুছে ফেলা বা পরিবর্তন করা যায় না— তাও আবার এমন কিছু যা সঠিক।

তারচে' যদি বাংলা এপ্লিকেশন সফটওয়্যার থেকে শুরু করে ইউনিকোড ভিত্তিক বাংলা কম্পিউটিং সঠিকভাবে চালু করতে তারা উদ্যোগী হন তবে সেটাই হবে সময়োচিত ও সঠিক পদক্ষেপ। কেননা, একটি জাতিকে শিক্ষিত করে তুলতে চাইলে মাতৃভাষাতেই সেটি করতে হবে। তাই মাতৃভাষার মাহাত্ম্য যত দ্রুত আমরা উপলব্ধি করতে পারব ততই আমাদের মঙ্গল। মঙ্গল দেশ তথা সমগ্র জাতির।

কৃতজ্ঞতা স্বীকার :

রাহাত আইয়ুব, শামসুদ্দোহা রঞ্জু

Glossary

Accent Mark - A mark placed above, below, or to the side of a character to alter its phonetic value.

ASCII - Acronym for American Standard Code for Information Interchange, a 7-bit code that is the U.S. national variant of ISO/IEC 646. Formally, the U.S. standard ANSI X3.4.

Base Character - A character that does not graphically combine with preceding characters, and that is neither a control nor a format character.

Block / Character Block - A grouping of related characters within the Unicode encoding space. A block may contain unassigned positions, which are reserved.

Character - (1) The smallest component of written language that has semantic value; refers to the abstract meaning and/or shape, rather than a specific shape, though in code tables some form of visual representation is essential for the reader's understanding. (2) Synonym for abstract character. (3) The basic unit of encoding for the Unicode character encoding. (4) The English name for the ideographic written elements of Chinese origin.

Character Class - A set of characters sharing a particular set of properties.

Character Encoding Form - Mapping from a character set definition to the actual code units used to represent the data.

Character Properties - A set of property names and property values associated with individual characters.

Character Set - A collection of elements used to represent textual information.

Coded Character Representation - An ordered sequence of one or more code units that is associated with an abstract character in a given character repertoire.

Coded Character Sequence - An ordered sequence of coded character representations.

Coded Character Set - A character set in which each character is assigned a numeric code point. Frequently abbreviated as character set, charset, or code set.

Code Page - A coded character set, often referring to a coded character set used by a personal computer, for example, PC code page 437, the default coded character set used by the U.S. English version of the DOS operating system.

Code Point - (1) A numerical index (or position) in an encoding table used for encoding characters. (2) Synonym for Unicode scalar value.

Code Position - Synonym for code point. Used in ISO character encoding standards.

Codespace - A range of numerical values available for encoding characters.

Code Unit - The minimal bit combination that can represent a unit of encoded text for processing or interchange.

Code Value - Synonym for code unit.

Collation / Alphabetic Sorting - The process of ordering units of textual information. Collation is usually specific to a particular language. Also known as alphabetizing or alphabetic sorting. In Unicode Technical Report #10, Unicode Collation Algorithm, defines a complete, unambiguous, specified ordering for all characters in the Unicode Standard.

Combining Character - A character that graphically combines with a preceding base character. The combining character is said to apply to that base character.

Composite Character / Decomposable Character - A character that is equivalent to a sequence of one or more other characters. It may also be known as a precomposed character or a composite character.

Decomposition - The process of separating or analyzing a text element into component units. These component units may not have any functional status, but may be simply formal units, that is, abstract shapes.

Diacritic - (1) A mark applied or attached to a symbol to create a new symbol that represents a modified or new value. (2) A mark applied to a symbol irrespective of whether it changes the value of that symbol. In the latter case, the diacritic usually represents an independent value (for example, an accent, tone, or some other linguistic information). Also called diacritical mark or diacritical.

Digraph - A pair of signs or symbols (two graphs), which together represent a single sound or a single linguistic unit. The English writing system employs many digraphs (for example, th, ch, sh, qu, and so on). The same two symbols may not always be interpreted as a digraph (for example, cathode versus cathouse). When three signs are so combined, they are called a trigraph. More than three are usually called an n-graph.

Font - A collection of glyphs used for the visual depiction of character data. A font is often associated with a set of parameters (for example, size, posture, weight, and serifness), which, when set to particular values, generate a collection of imitable glyphs.

Glyph - (1) An abstract form that represents one or more glyph images. (2) A synonym for glyph image. In displaying Unicode character data, one or more glyphs may be selected to depict a particular character. These glyphs are selected by a rendering engine during composition and layout processing.

Glyph Code - A numeric code that refers to a glyph. Usually, the glyphs contained in a font are referenced by their glyph code. Glyph codes may be local to a particular font; that is, a different font containing the same glyphs may use different codes.

Glyph Image - The actual, concrete image of a glyph representation having been rasterized or otherwise imaged onto some display surface.

Halant - A synonym for the virama character. It literally means killer, referring to its function of killing the inherent vowel of a consonant letter.

Independent Vowel - In Indic scripts, certain vowels are depicted using independent letter symbols that stand on their own. This is often true when a word starts with a vowel or a word consists only of a vowel.

Letter - (1) An element of an alphabet. In a broad sense, includes elements of syllabaries and ideographs. (2) Informative property of characters that are used to write words.

Ligature - A glyph representing a combination of two or more characters. In the Latin script, there are only a few in modern use, such as the ligatures between f and i or f and l. Other scripts make use of many ligatures, depending on the font and style.

Mathematical Property - Informative property of characters that are used as operators in mathematical formulae.

Matra - A dependent vowel in an Indic script. It is the name for vowel letters that follow consonant letters in logical order. A matra often has a completely different letter form from that for the same phonological vowel used as an independent letter.

Normalization - A process of removing alternate representations of equivalent sequences from textual data, to convert the data into a form which can be binary-compared for equivalence. In the Unicode Standard, normalization refers specifically to processing to ensure that canonically equivalent (and/or compatibility-equivalent) strings have unique representations.

Plain Text - Computer-encoded text that consists only of a sequence of code points from a given standard, with no other formatting or structural information. Plain text interchange is commonly used between computer systems that do not share higher-level protocols.

Private Use - Unicode scalar values (code points) from U+E000 to U+F8FF, U+F0000 to U+FFFFD, and U+100000 to U+10FFFF are available for private use. Refers to code points of the standard whose interpretation is not specified by the standard and whose use may be determined by private agreement among cooperating users.

Script - A collection of symbols used to represent textual information in one or more writing systems.

Transformation Format - A mapping from a coded character sequence to a unique sequence of code units (typically bytes).

Unicode Character Database - A collection of files providing normative and informative Unicode character properties and mappings.

Unicode Signature - An implicit marker to identify a file as containing Unicode text in a particular encoding form. An initial byte order mark (BOM) may be used as a Unicode signature.

Virama - The name of a symbol used with Indic scripts to indicate a dead consonant. Also called halant.

Annexure II

The primary scripts currently supported by Unicode 4.0 are:

Arabic	Devanagari	Hebrew	Mongolian	Tagalog
Armenian	Ethiopic	Hiragana	Myanmar	Tagbanwa
Bengali	Georgian	Kannada	Ogham	Tai Le
Bopomofo	Gothic	Katakana	Old Italic (Etruscan)	Tamil
Buhid	Greek	Khmer	Osmanya	Telugu
Canadian Syllabics	Gujarati	Lao	Oriya	Thaana
Cherokee	Gurmukhi	Latin	Runic	Thai
Cypriot	Han	Limbu	Shavian	Tibetan
Cyrillic	Hangul	Linear B	Sinhala	Ugaritic
Deseret	Hanun o	Malayalam	Syriac	Yi

Not Supported Yet:

Balinese	Kanglei	Pahawh Hmong	Tai Lu
Javanese	Moso (Naxi)	Rong (Lepcha)	Tai Mau
Manipuri (Meithei,	N ko (Mandekan)	Siloti Nagri (Syloti Nagri)	Tifinagh

Archaic and Obsolete Scripts:

Ahom	Glagolitic	Mandaic	Pyu	Cuneiform
Aramaic	Hieroglyphic Egyptian	Mangyan	Old Persian Cuneiform	Tangut (Xi Xia)
Balti	Hieroglyphic Hittite	Mayan	Phoenician	Tircul
Batak	Kaithi	Meroitic (Kush)	Satavahana	Turkmenistan Brahmic (Khotanese)
Brahmi	Kawi	Modi	Siddham	
Buginese	Khamti	Numidian	South Arabian	
Chola	Kharoshthi	Pahlavi and Avestan	Sumero-Akkadian	
Cypro-Minoan	Lahnda	Phags-pa		

Bengali Code Block

@@	0980	Bengali	09FF	09C7	BENGALI VOWEL SIGN E
@		Based on ISCI 1988			* stands to the left of the consonant
@		Various signs		09C8	BENGALI VOWEL SIGN AI
0981		BENGALI SIGN CANDRABINDU			* stands to the left of the consonant
0982		BENGALI SIGN ANUSVARA		09CB	BENGALI VOWEL SIGN O
0983		BENGALI SIGN VISARGA			* pieces on both sides of the consonant
@		Independent vowels			: 09C7 09BE
0985		BENGALI LETTER A		09CC	BENGALI VOWEL SIGN AU
0986		BENGALI LETTER AA			* pieces on both sides of the consonant
0987		BENGALI LETTER I			: 09C7 09D7
0988		BENGALI LETTER II		@	Various signs
0989		BENGALI LETTER U		09CD	BENGALI SIGN VIRAMA
098A		BENGALI LETTER UU			= halant
098B		BENGALI LETTER VOCALIC R		09D7	BENGALI AU LENGTH MARK
098C		BENGALI LETTER VOCALIC L		@	Additional consonants
098F		BENGALI LETTER E		09DC	BENGALI LETTER RRA
0990		BENGALI LETTER AI			: 09A1 09BC
0993		BENGALI LETTER O		09DD	BENGALI LETTER RHA
0994		BENGALI LETTER AU			: 09A2 09BC
@		Consonants		09DF	BENGALI LETTER YYA
0995		BENGALI LETTER KA			: 09AF 09BC
0996		BENGALI LETTER KHA		@	Generic additions
0997		BENGALI LETTER GA		09E0	BENGALI LETTER VOCALIC RR
0998		BENGALI LETTER GHA		09E1	BENGALI LETTER VOCALIC LL
0999		BENGALI LETTER NGA		09E2	BENGALI VOWEL SIGN VOCALIC L
099A		BENGALI LETTER CA		09E3	BENGALI VOWEL SIGN VOCALIC LL
099B		BENGALI LETTER CHA		@	Digits
099C		BENGALI LETTER JA		09E6	BENGALI DIGIT ZERO
099D		BENGALI LETTER JHA		09E7	BENGALI DIGIT ONE
099E		BENGALI LETTER NYA		09E8	BENGALI DIGIT TWO
099F		BENGALI LETTER TTA		09E9	BENGALI DIGIT THREE
09A0		BENGALI LETTER TTHA		09EA	BENGALI DIGIT FOUR
09A1		BENGALI LETTER DDA		09EB	BENGALI DIGIT FIVE
09A2		BENGALI LETTER DDHA		09EC	BENGALI DIGIT SIX
09A3		BENGALI LETTER NNA		09ED	BENGALI DIGIT SEVEN
09A4		BENGALI LETTER TA		09EE	BENGALI DIGIT EIGHT
09A5		BENGALI LETTER THA		09EF	BENGALI DIGIT NINE
09A6		BENGALI LETTER DA		@	Bengali-specific additions
09A7		BENGALI LETTER DHA		09F0	BENGALI LETTER RA WITH MIDDLE
09A8		BENGALI LETTER NA			DIAGONAL (Assamese)
09AA		BENGALI LETTER PA			* Assamese
09AB		BENGALI LETTER PHA		09F1	BENGALI LETTER RA WITH LOWER
09AC		BENGALI LETTER BA			DIAGONAL (Assamese)
		= Bengali va, wa			= BENGALI LETTER VA
09AD		BENGALI LETTER BHA			WITH LOWER DIAGONAL
09AE		BENGALI LETTER MA			* Assamese
09AF		BENGALI LETTER YA		09F2	BENGALI RUPEE MARK
09B0		BENGALI LETTER RA		09F3	BENGALI RUPEE SIGN
09B2		BENGALI LETTER LA		09F4	BENGALI CURRENCY
09B6		BENGALI LETTER SHA			NUMERATOR ONE
09B7		BENGALI LETTER SSA			* not in current usage
09B8		BENGALI LETTER SA		09F5	BENGALI CURRENCY
09B9		BENGALI LETTER HA			NUMERATOR TWO
@		Various signs			* not in current usage
09BC		BENGALI SIGN NUKTA		09F6	BENGALI CURRENCY
		* for extending the alphabet to new			NUMERATOR THREE
		letters			* not in current usage
@		Dependent vowel signs		09F7	BENGALI CURRENCY
09BE		BENGALI VOWEL SIGN AA			NUMERATOR FOUR
09BF		BENGALI VOWEL SIGN I		09F8	BENGALI CURRENCY
		* stands to the left of the consonant			NUMERATOR ONE LESS
09C0		BENGALI VOWEL SIGN II			THAN THE DENOMINATOR
09C1		BENGALI VOWEL SIGN U		09F9	BENGALI CURRENCY
09C2		BENGALI VOWEL SIGN UU			DENOMINATOR SIXTEEN
09C3		BENGALI VOWEL SIGN VOCALIC R		09FA	BENGALI ISSHAR
09C4		BENGALI VOWEL SIGN VOCALIC RR			