

# Was ist Unicode?

Beitrag zum Symposium  
„Teuthonista goes Unicode“  
am 4. Juni 2010 in Passau

Karl Pentzlin

# Was ist Unicode?

Unicode gibt jedem Textzeichen eine weltweit eindeutige Nummer (Zeichencode).

**Unicode**

= **universal unique  
character encoding.**

# Was ist Unicode?

Unicode gibt jedem Textzeichen eine weltweit eindeutige Nummer.

A — 65	A — 913	Ÿ — 5792	ᵝ — 7440
B — 66	B — 914	Ŋ — 5794	ᵞ — 7441
C — 67	Γ — 915	ƀ — 5798	ᵠ — 7442
D — 68	Δ — 916	Ƒ — 5800	ᵡ — 7443

**Umgekehrt identifiziert diese Nummer das Zeichen eindeutig, gleich in welchem Zusammenhang.**

## Zentraler Begriff:

- **Plain text** (dt. etwa: »elementarer Text«)

*Text, dessen Bedeutung von Darstellungsweise und Positionierung seiner Zeichen unabhängig ist.*

*Im Gegensatz zu: **Notation.***

### Elementarer Text:

Gallia est omnis divisa in partes tres, quarum unam incolunt Belgae, aliam Aquitani, tertiam qui ipsorum lingua Celtae, nostra Galli appellantur.

### Elementarer Text:

ains 'stɔɪtn zɪç 'nɔɛtvɪnt ʊn  
'zɔnə, vɛə fɔn im 'baɪdn vol də  
'stɛəkəʁə vɛʁə, als ain 'vandərəə,  
dəə in ain 'vaəm 'mantl gə'hɪlt  
vəə, dəs 'vegəs da'heəkɑ:m.

### Notation:

$$\sum_{j=1}^3 \frac{d}{dx_j} \frac{\partial \mathcal{L}}{\partial \frac{\partial \psi^*}{\partial x_j}}$$

Musical notation for Horn in F and Violin 1. The Horn part is in F major, 2/4 time, and the Violin part is in F major, 2/4 time. The Horn part plays a sequence of notes: G4, A4, B4, C5, D5, E5, F5. The Violin part plays a sequence of notes: G4, A4, B4, C5, D5, E5, F5.

## *Zentrale Begriffe:*

- **Character** (dt. hier: »**Textzeichen**«)

*"Kleinste bedeutungstragende Einheit" von Texten:*

- Buchstaben
- Diakritische Zeichen
- Ziffern
- Satzzeichen
- Leerzeichen
- ...

- **Glyph** (dt. »Glyphe«)

*Konkrete Darstellung eines Textzeichens.*

*Beispiel: Glyphen des lateinischen Buchstabens a:*

a a a a a a a a

## Elementarer Text

- ist also vollständig gegeben durch die einfache Aneinanderreihung seiner Textzeichen.
  - kann so: **ohne spezielle Hilfsmittel,**
    - Vorbereitungen und Kenntnisse**
      - weltweit gelesen werden, z.B. im Internet
      - in ein anderes Dokument kopiert werden
      - in Dateien und Datenbanken gespeichert, zurückgelesen und gesucht werden
      - sicher langzeitarchiviert werden
- wenn die Textzeichen einheitlich codiert sind.**

## Elementarer Text:

- Sprachliche Äußerungen sind grundsätzlich sequenziell (im Unterschied z.B. zu Mathematik und Musik).
- Dies gilt auch für die Wiedergabe in Lautschrift.
- Lautschriften sind deshalb grundsätzlich für Unicode-Codierung geeignet.
- Mit UPA (Uralic Phonetic Alphabet) ist bereits eine Lautschrift in Unicode aufgenommen, die ähnlich wie Teuthonista:
  - nicht direkt IPA entspricht,
  - eine Tradition bis zum Anfang des 20. Jhdts. hat,
  - noch heute in ihrem speziellen Bereich verwendet wird,
  - nicht vollständig formal normiert ist.
- Teuthonista kann also einer Aufnahme in Unicode entgegen sehen, sodass die Dialektforschung unseres Gebietes von den Unicode-Vorteilen profitieren kann.

## **Kurzer Abriss der Zeichencodierungs-Geschichte**

**ASCII** (ab 1963): *Nummernbereich bis 127.*

Nur Grundbuchstaben (A-Z, a-z), Ziffern, wichtigste Satzzeichen.

**ISO 8859, DOS- und Windows-Codepages:**

*Nummernbereich bis 255.*

Unterschiedliche Nutzung der höheren Nummern  
128-255 für verschiedene Anwendungsbereiche  
(Westeuropa, Osteuropa, ...)

Für Lautschrift usw. "Bastellösungen"

(von Anwendergruppen ohne internationale Normung  
festgelegte Codes, meist an die Verwendung  
bestimmter Schriftartdateien gebunden).

**Unicode** (in Entwicklung seit 1991):

*Nummernbereich bis 1.114.109 .*

## **In Unicode enthalten sind z.B.:**

- Alle verbreiteten Schriften (z.B. Latein, Arabisch, Chinesisch)
- Fast alle Sonderbuchstaben in diesen Schriften (z.B. in Latein: deutsches "ß", Sami "Ń/ņ", Sioux "ᑎ/ᑏ")
- Fast alle sonstigen "lebenden" Schriften (z.B. Balinesisch)
- Zahlreiche "tote" Schriften (z.B. Runen, sumerische Keilschrift)
- Lautschrift (IPA, Americanist Usage, Uralic Phonetic Alphabet)
- Ziffern und Satzzeichen
- Symbole , die in elementarem Text vorkommen (z.B. Metriksymbole, Pfeile, Währungszeichen, Smileys)
- "Bauelemente" für Notationen (z.B. Musiknotenköpfe, mathematische Buchstabenformen, Schachfiguren)

## **Unicode ist laufend in Entwicklung:**

- Die aktuelle Version 5.2 enthält 107.296 Zeichen
- Version 6.0 mit weiteren 1.864 Zeichen auf dem Weg

**Fehlende Zeichen werden also laufend ergänzt.**

-----

*Für Teuthonista stellen sich nunmehr die Fragen:*

- **Wer entscheidet über die Aufnahme von Zeichen?**
- **Wie und wo beantragt man diese Aufnahme?**
- **Wie verläuft der Entscheidungsprozess, und welche Mitwirkung ist für einen Erfolg nötig?**

*Wer entscheidet über die Aufnahme von Zeichen?*

## **Zwei Standards, zwei Organisationen:**

- Das "**Unicode Consortium**" ist Herausgeber des "**Unicode Standard**".
- **ISO** (International Organization for Standardization) ist Herausgeber des "International Standard **ISO/IEC 10646** — Information technology — Universal Multiple-Octet Coded Character Set (UCS)"
- Beide Organisationen koordinieren ihre Standards.
- Die codierten Zeichen und ihre Codes sind stets gleich.
- Technische Details sind in unterschiedlicher Detailliertheit geregelt, i.d.R. im "Unicode Standard" ausführlicher.

## Unicode-Consortium:

- Mitgliedschaft von Institutionen und Einzelpersonen möglich
- Full and Institutional Members (stimmberechtigt):  
Adobe, Apple, Denic, Google, IBM, Microsoft, Oracle, SAP, Sybase, Yahoo, Gov. of India, Univ. Berkeley (CA), Univ. Santa Cruz (CA)  
*Jahresbeitrag: 10.000 \$ - 15.000 \$*  
Neu-Mitglieder sind laut Internet-Seite jederzeit willkommen.
- Neben anderen stimmrechtslosen Stufen für Institutionen sind Individual Memberships (ohne Stimmrecht) für 75 \$ jährlich möglich.
- Mitglieder (bei Institutionen die festbenannten Repräsentanten) dürfen an allen Meetings teilnehmen und haben vollen Zugriff auf Information (z.B. interne Mailinglisten).
- Meetings sind 4x im Jahr für je 5 Tage, fast immer an der Westküste der USA. Die Repräsentanten der US-Firmen sind regelmäßige Teilnehmer (*die deutschen Mitglieder Denic und SAP sind regelmäßig nicht vertreten*).
- Die anstehenden Punkte werden dort technisch eingehend diskutiert. Die Entscheidungen fallen fast alle im Konsens.

## **ISO/IEC JTC1/SC2/WG2:**

ISO / International Electrotechnical Commission

Joint Technical Committee 1: Information Technology Standards

Subcommittee 2: Coded Character Sets

Workgroup 2: Universal Coded Character Set

- Mitglieder sind Staaten. Jeder Mitgliedsstaat hat eine Stimme.
- Die Staaten werden vertreten durch ihre nationalen Normungsinstitute (für Deutschland: DIN; zuständig: DIN NA 043-01-29-01 AK "Codierte Zeichensätze").
- Meetings: 2x jährlich für 1 Woche an wechselnden Orten in der Welt (zuletzt: Dublin, Tokio, San Jose (CA); Juni 2011: Helsinki)
- Jeder Staat darf beliebig viele Vertreter entsenden (einen davon stimmberechtigt). Außerdem werden Experten eingeladen.
- Die Vertretung der USA (ANSI) ist praktisch personalidentisch mit den amerikanischen Unicode-Teilnehmern.
- Entscheidungen fallen auch hier fast ausschließlich im Konsens.

# Aufnahme neuer Zeichen in Unicode

- "Von alleine" passiert nichts.  
Die Gremien suchen nicht selbst nach nicht codierten Zeichen.
- Es muss ein Codierungsvorschlag erstellt werden.

L2/10-161

Universal Multiple-Octet Coded Character Set  
International Organization for Standardization  
Organisation Internationale de Normalisation  
Международная организация по стандартизации

**Doc Type:** Working Group Document  
**Title:** Proposal to encode two missing modifier letters for Extended IPA  
**Source:** Karl Pentzlin  
**Status:** Individual Contribution  
**Action:** For consideration by JTC1/SC2/WG2 and UTC  
**Date:** 2010-04-30

**Additions for Extended IPA**

H	U+A7F8	MODIFIER LETTER CAPITAL H WITH STROKE · faucalized → 0126 latin capital letter h with stroke ≈ <sup> 0126
œ	U+A7F9	MODIFIER LETTER SMALL LIGATURE OE · labialized: open-rounded → 0153 latin small ligature oe ≈ <sup> 0153

**Properties:**  
A7F8;MODIFIER LETTER CAPITAL H WITH STROKE;Lm;0;L;;;;N;;;;

## Vorschläge zur Codierung neuer Zeichen sind einreichbar

- *von jedermann beim Unicode-Consortium,*
- *von den staatlichen Organisationen (wie DIN) bei ISO.*
- Eine Einreichung reicht.  
Das empfangende Gremium prüft in der Regel zuerst.
- Die Wahl des Empfängers richtet sich u.a. danach:
  - Bei welchem Gremium kann man persönlich vertreten sein?
  - *Wenn man die voraussichtlich teilnehmenden Personen kennt:*  
Wem traut man am ehesten das Verständnis seiner Problematik zu?

## Formale Anforderungen an einen Codierungsvorschlag

- Die Zeichen müssen nach Unicode-Regeln eindeutig benannt werden.
- Verschiedene "properties" müssen eindeutig benannt werden (z.B. Kombinationsverhalten bei Diakritika).
- Die Zeichen müssen genau beschrieben sein; z.B. muss klargestellt sein, dass es Textzeichen mit klarer Semantik (also nicht nur "glyph variants") sind.
- "the Unicode Standard does not encode idiosyncratic, personal, novel, or private-use characters"  
— *für Teuthonista eher kein Problem.*
- Ein Formblatt mit speziellen Angaben ist beizufügen.

## Inhaltliche Anforderungen an einen Codierungsvorschlag

- Die Zeichen müssen durch Beispiele belegt werden.
- Dies können historische und aktuelle Anwendungen sein, einschließlich aktuelle Zitate von historischen Anwendungen.
- Sofern relevant, ist die historische Entwicklung von Zeichen zu diskutieren, speziell wenn für die Zeichen-Identität relevant ("glyph vs. character").
- Zeichenlisten und vorhandene standardähnliche Dokumente sind hilfreich.
- *Ein Codierungsvorschlag ist keine wissenschaftliche Publikation, sondern ein zielgerichtetes Aktionspapier.*

## Der Weg des Codierungsvorschlag nach der Einreichung

- ISO-Meetings (zukünftig vsl. alle 8 Monate):
- **Neue Vorschläge: Diskussion und PDAM –**  
die eingegangenen Vorschläge werden diskutiert, und die für gut befundenen Vorschläge zu einem PDAM (Proposed Draft Amendment for ISO/IEC 10646) zusammengestellt.
- **Kommentarphase, dann auf nächstem Meeting FPDAM –**  
die nationalen Organisationen (DIN usw.), sowie das Unicode-Konsortium (UTC = Unicode Technical Committee) , prüfen die Vorschläge und geben Kommentare ab (Änderungs- oder Ablehnungswünsche). Speziell das UTC diskutiert und prüft die Vorschläge eingehend. Das nächste ISO-Meeting diskutiert entscheidet über die Berücksichtigung der Kommentare und erstellt ein FPDAM (Final Proposed Draft Amendment for ISO/IEC 10646).
- **Kommentarphase, dann auf nächstem Meeting FDAM –**  
weiterer Kommentierungszyklus: letzte Möglichkeit für technische Änderungen. Ein FDAM (Final Draft Amendment) kann nur noch redaktionell geändert werden.  
**Durch neue ISO-Richtlinien ist ab 2011 vsl. ein zusätzlicher Zyklus (DAM) erforderlich.**
- **Dann auf nächstem Meeting Verabschiedung.**  
I.d.R. nach 2 verabschiedeten FDAMs wird eine synchronisierte Unicode-Version herausgegeben.
- **Aktive Mitarbeit und Meeting-Teilnahme bis zum FDAM (d.h. an 3 ISO-Meetings und ggf. UTC-Meetings) wichtig!**

Vielen Dank.