

Unicode and the Storage of Data

Jost Gippert

Berlin, 25.10.2007

Outline

- Background of the Unicode Standard
- Questions of Data Storage and Retrieval
- Strategies and Recommendations
- Summary

Two sorts of text encoding:

- Encoding of CHARACTERS
- Encoding of STRUCTURAL ELEMENTS OF TEXTS (“Formatting”)
- UNICODE is about the former, not the latter

Background of UNICODE

- Encoding of characters - a history of chaos

Encoding from "stone age" to UNICODE: Rigveda

```
R700123011AGNI!M+ NA!RO DI:!D)ITIB)IR ARA!N\YOR HA!STACYUTI: JANAYANTA PRAS=ASTA
R700123012!M / DU:RED9!S=AM+ G9HA!PATIM AT)ARYU!M
R700123021TA!M AGNI!M A!STE VA!SAVO NY 9&N\VAN SUPRATICA!KS\AM A!VASE KU!TAS= CI
R700123022T / DAKS\A:!YYO YO! DA!MA A:!SA NI!TYAH-
R700123031PRE!DD)O AGNE DI:DIHI PURO! NO! 'JASRAYA: SU:RMYA:& YAVIS\T\A / TVA:!
R700123032M+ S=A!S=VANTA U!PA YANTI VA:!JA:H-
R700123041PRA! TE AGNA!YO 'GNI!B)YO VA!RAM+ NI!H- SUVI:!RA:SAH- S=OS=UCANTA DYUM
R700123042A!NTAH- / YA!TRA: NA!RAH- SAMA:!SATE SUJA:TA:!H-
R700123051DA:! NO AGNE D)IYA:' RAYI!M+ SUVI:!RAM+ SVAPATYA!M+ SAHASYA PRAS=ASTA!
R700123052M / NA! YA!M+ YA:!VA: TA!RATI YA:TUMA:!VA:N
R700124061U!PA YA!M E!TI YUVATI!H- SUDA!KS\AM+ DOS\A:! VA!STOR HAVI!S\MATI: G)9T
R700124062A:!CI: / U!PA SVAI!NAM ARA!MATIR VASU:YU!H-
R700124071VI!S=VA: AGNE! 'PA DAHA:!RA:TI:R YE!B)IS TA!POB)IR A!DAHO JA!RU:T)AM /
R700124072 PRA! NISVARA!M+ CA:TAYASVA:!MI:VA:M
R700124081A:! YA!S TE AGNA ID)ATE! A!NI:KAM+ VA!SIS\T\A S=U!KRA DI:!DIVAH- PA:!
R700124082VAKA / UTO! NA EBI)! STAVA!T)AIR IHA! SYA:H-
```

|bMaṇḍala_7 |c1 (517)

|1 agnīm náro dīdhitibhir arāṇyor hástacyutī janayanta praśastām / dūredṛśam grhāpatim atharyúm
|2 tám agnīm áste vásavo ny ṛṇvan supratícaḥśam ávase kútaś cit / dakṣāyyo yó dámaś ása nítyaḥ
|3 prēddho agne dīdhi puró nó 'jasrayā sūrmyā yaviṣṭha / tvām śásvanta úpa yanti vājāḥ
|4 prá te agnáyo 'gnībhyo váram nīḥ suvīrāsaḥ śósucanta dyumántaḥ / yātrā náraḥ samásate sujātāḥ
|5 dá no agne dhiyā rayīm suvīram svapatyām sahasya praśastām / ná yām yāvā tárati yātumāvān
|6 úpa yām éti yuvatīḥ sudákṣam doṣā vástor haviṣmatī ghṛtācī / úpa svainam arámatir vasūyúḥ
|7 víśvā agné 'pa dahārātīr yébhis tápobhir ádaho járūtham / prá nisvarām cātayasvāmīvām
|8 ā yás te agna idhaté ánīkam vásiṣṭha śúkra dīdivaḥ pávaka / utó na ebhí staváthair ihá syāḥ

Encoding from “stone age” to UNICODE: Greek “Beta-Code”

nç nñ né Å ¢ @@@@{1\$20*Q*E*O*G*O*N*I*A\$}1 ü@*MOUSA/WN *(ELIKWNI
A/DWN A)RXW/MEQ' A)EI/DEIN, ÇAI(/ Q' *(ELIKW=NOS E)/XOUSIN O)/ROS ME/GA T
E ZA/QEO/N TE, ÇKAI/ TE PERI\ KRH/NHN I)OEIDE/A PO/SS' A(PALOI=SIN ÇO)RXE
U=NTAI KAI\ BWMO\N E)RISQENE/OS *KRONI/WNOS: ÇKAI/ TE LOESSA/MENAI TE/REN
A XRO/A *PERMHSSOI=O ÇH)' */IPPOU KRH/NHS H)' *)OLMEIOU= ZAQE/OIO ÇA)KRO
TA/TW| *(ELIKW=NI XOROU\S E)NEPOIH/SANTO, ÇKALOU\S I(MERO/ENTAS, E)PERRW/
SANTO DE\ POSSI/N. ÇE)/NQEN A)PORNU/MENAI KEKALUMME/NAI H)E/RI POLLW=| ÇE

|bHesiod

|cTheog. |p0 ΘEOΓONIA

[p1 Μουσάων Ἐλικωνιάδων ἀρχώμεθ' αἰεΐδειν,

|p2 αἱ θ' Ἐλικῶνος ἔχουσιν ὄρος μέγα τε ζάθεόν τε,

|p3 καί τε περὶ κρήνην ἰοειδέα πόσσ' ἀπαλοῖσιν

|p4 ὀρχεῦνται καὶ βωμὸν ἐρισθενέος Κρονίωνος·

|p5 καί τε λοεσσάμεναι τέρενα χροά Περμησοῖο

|p6 ἡ' Ἰππου κρήνης ἡ' Ὀλμειοῦ ζαθέοιο

|p7 ἀκροτάτῳ Ἑλικῶνι χοροὺς ἐνεποιήσαντο,

|ρ8 καλοὺς ἡμερόεντας, ἐπερρώσαντο δὲ ποσσίν.

|p9 ἔνθεν ἀπορνύμεναι κεκαλυμμένοι ἥερι πολλῷ

Encoding from “stone age” to UNICODE: Avesta

35,2

Haptanghâiti.

35,6

1. Երբ 2. անգամ 3. անգամ 4. անգամ 5. Բարեկամ 6. Բարեկամ 7. Բարեկամ 8. Բարեկամ 9. Բարեկամ 10. Բարեկամ

1. אלהינו. 2. אלהינו. 3. אלהינו. 4. אלהינו. 5. אלהינו. 6. אלהינו. 7. אלהינו.

၁။ အထွေထွေအားဖြင့် မြန်မာ့အလင်း
 ၂။ အထွေထွေအားဖြင့် မြန်မာ့အလင်း
 ၃။ အထွေထွေအားဖြင့် မြန်မာ့အလင်း
 ၄။ အထွေထွေအားဖြင့် မြန်မာ့အလင်း
 ၅။ အထွေထွေအားဖြင့် မြန်မာ့အလင်း
 ၆။ အထွေထွေအားဖြင့် မြန်မာ့အလင်း
 ၇။ အထွေထွေအားဖြင့် မြန်မာ့အလင်း
 ၈။ အထွေထွေအားဖြင့် မြန်မာ့အလင်း
 ၉။ အထွေထွေအားဖြင့် မြန်မာ့အလင်း
 ၁၀။ အထွေထွေအားဖြင့် မြန်မာ့အလင်း

[illegible]

3 2. 1. 3. 4. 5. 6. 7. 8. 9. 10. 11. 12. 13. 14. 15. 16. 17. 18. 19. 20. 21. 22. 23. 24. 25. 26. 27. 28. 29. 30. 31. 32. 33. 34. 35. 36. 37. 38. 39. 40. 41. 42. 43. 44. 45. 46. 47. 48. 49. 50. 51. 52. 53. 54. 55. 56. 57. 58. 59. 60. 61. 62. 63. 64. 65. 66. 67. 68. 69. 70. 71. 72. 73. 74. 75. 76. 77. 78. 79. 80. 81. 82. 83. 84. 85. 86. 87. 88. 89. 90. 91. 92. 93. 94. 95. 96. 97. 98. 99. 100. 101. 102. 103. 104. 105. 106. 107. 108. 109. 110. 111. 112. 113. 114. 115. 116. 117. 118. 119. 120. 121. 122. 123. 124. 125. 126. 127. 128. 129. 130. 131. 132. 133. 134. 135. 136. 137. 138. 139. 140. 141. 142. 143. 144. 145. 146. 147. 148. 149. 150. 151. 152. 153. 154. 155. 156. 157. 158. 159. 160. 161. 162. 163. 164. 165. 166. 167. 168. 169. 170. 171. 172. 173. 174. 175. 176. 177. 178. 179. 180. 181. 182. 183. 184. 185. 186. 187. 188. 189. 190. 191. 192. 193. 194. 195. 196. 197. 198. 199. 200. 201. 202. 203. 204. 205. 206. 207. 208. 209. 210. 211. 212. 213. 214. 215. 216. 217. 218. 219. 220. 221. 222. 223. 224. 225. 226. 227. 228. 229. 230. 231. 232. 233. 234. 235. 236. 237. 238. 239. 240. 241. 242. 243. 244. 245. 246. 247. 248. 249. 250. 251. 252. 253. 254. 255. 256. 257. 258. 259. 260. 261. 262. 263. 264. 265. 266. 267. 268. 269. 270. 271. 272. 273. 274. 275. 276. 277. 278. 279. 280. 281. 282. 283. 284. 285. 286. 287. 288. 289. 290. 291. 292. 293. 294. 295. 296. 297. 298. 299. 300. 301. 302. 303. 304. 305. 306. 307. 308. 309. 310. 311. 312. 313. 314. 315. 316. 317. 318. 319. 320. 321. 322. 323. 324. 325. 326. 327. 328. 329. 330. 331. 332. 333. 334. 335. 336. 337. 338. 339. 340. 341. 342. 343. 344. 345. 346. 347. 348. 349. 350. 351. 352. 353. 354. 355. 356. 357. 358. 359. 360. 361. 362. 363. 364. 365. 366. 367. 368. 369. 370. 371. 372. 373. 374. 375. 376. 377. 378. 379. 380. 381. 382. 383. 384. 385. 386. 387. 388. 389. 390. 391. 392. 393. 394. 395. 396. 397. 398. 399. 400. 401. 402. 403. 404. 405. 406. 407. 408. 409. 410. 411. 412. 413. 414. 415. 416. 417. 418. 419. 420. 421. 422. 423. 424. 425. 426. 427. 428. 429. 430. 431. 432. 433. 434. 435. 436. 437. 438. 439. 440. 441. 442. 443. 444. 445. 446. 447. 448. 449. 450. 451. 452. 453. 454. 455. 456. 457. 458. 459. 460. 461. 462. 463. 464. 465. 466. 467. 468. 469. 470. 471. 472. 473. 474. 475. 476. 477. 478. 479. 480. 481. 482. 483. 484. 485. 486. 487. 488. 489. 490. 491. 492. 493. 494. 495. 496. 497. 498. 499. 500. 501. 502. 503. 504. 505. 506. 507. 508. 509. 510. 511. 512. 513. 514. 515. 516. 517. 518. 519. 520. 521. 522. 523. 524. 525. 526. 527. 528. 529. 530. 531. 532. 533. 534. 535. 536. 537. 538. 539. 540. 541. 542. 543. 544. 545. 546. 547. 548. 549. 550. 551. 552. 553. 554. 555. 556. 557. 558. 559. 560. 561. 562. 563. 564. 565. 566. 567. 568. 569. 570. 571. 572. 573. 574. 575. 576. 577. 578. 579. 580. 581. 582. 583. 584. 585. 586. 587. 588. 589. 590. 591. 592. 593. 594. 595. 596. 597. 598. 599. 600. 601. 602. 603. 604. 605. 606. 607. 608. 609. 610. 611. 612. 613. 614. 615. 616. 617. 618. 619. 620. 621. 622. 623. 624. 625. 626. 627. 628. 629. 630. 631. 632. 633. 634. 635. 636. 637. 638. 639. 640. 641. 642. 643. 644. 645. 646. 647. 648. 649. 650. 651. 652. 653. 654. 655. 656. 657. 658. 659. 660. 661. 662. 663. 664. 665. 666. 667. 668. 669. 670. 671. 672. 673. 674. 675. 676. 677. 678. 679. 680. 681. 682. 683. 684. 685. 686. 687. 688. 689. 690. 691. 692. 693. 694. 695. 696. 697. 698. 699. 700. 701. 702. 703. 704. 705. 706. 707. 708. 709. 710. 711. 712. 713. 714. 715. 716. 717. 718. 719. 720. 721. 722. 723. 724. 725. 726. 727. 728. 729. 730. 731. 732. 733. 734. 735. 736. 737. 738. 739. 740. 741. 742. 743. 744. 745. 746. 747. 748. 749. 750. 751. 752. 753. 754. 755. 756. 757. 758. 759. 760. 761. 762. 763. 764. 765. 766. 767. 768. 769. 770. 771. 772. 773. 774. 775. 776. 777. 778. 779. 780. 781. 782. 783. 784. 785. 786. 787. 788. 789. 790. 791. 792. 793. 794. 795. 796. 797. 798. 799. 800. 801. 802. 803. 804. 805. 806. 807. 808. 809. 810. 811. 812. 813. 814. 815. 816. 817. 818. 819. 820. 821. 822. 823. 824. 825. 826. 827. 828. 829. 830. 831. 832. 833. 834. 835. 836. 837. 838. 839. 840.

[illegible]

Encoding from “stone age” to UNICODE: Avesta: 1985

```
Y 35 3
taT. aT. vairImaidI.1 ahurA.2
mazdA. a$A. srlrA. hiiaT.3 I.
mainimadicA.4 vaocOimAcA.5 vürüzimAcA.6
yA. hAtäm. ßiiaoöananäm.7 vahiStA.
XiiAT.8 ubOibiiA. ahubiiA.9

Y 35 4
lgauuOi. adAiS. tAiS. ßiiaoöanAiS.2
yAiS. vahiStAiS. fraEßiiAmahI.3
rAmAcA. vAstrümca. dazdiiAi. surunuuatascA.4
asurunuuatascA.5 x$aiaNtascA.6
ax$aiaNtascA.7

Y 35 5
lhux$aörOtümAi. bAT.2 x$aörüm. ahmaT.
hiiaT.3 aibl.4 dadümahicA.5 cIsmahicA.6
huuämahicA.7 hiiaT.8 mazdAi.
ahurAi. a$AicA. vahiStAi.: (hux$aörOtümAi.
.. vahiStAi. ...:)
```

Encoding from “stone age” to UNICODE: Avesta

Y 35 3

taT. aT. vairImaidI.1 ahurA.2
mazdA. ašA. srIrA. hiiaT.3 I.
mainimadicA.4 vaocOimAcA.5 vürüzimAcA.6
yA. hAtäm. ßiiaoÖananäm.7 vahiStA.
XiiAT.8 ubOibiiA. ahubiiA.9

Y 35 4

1gauuOi. adAiS. tAiS. ßiiaoÖanAiS.2
yAiS. vahiStAiS. fraEßiiAmahI.3
rAmAcA. vAstrümcA. dazdiiAi. surunuuatascA.4
asurunuuatascA.5 xšaiiaNtascA.6
axšaiiaNtascA.7

Y 35 5

1huxšaörOtümAi. bAT.2 xšaörüm. ahmaT.
hiiaT.3 aibl.4 dadümahicA.5 cİSmahicA.6
huuämahicA.7 hiiaT.8 mazdAi.
ahurAi. ašAicA. vahiStAi.: (huxšaörOtümAi.
.. vahiStAi. ...:)

Y 35 3

taṭ. aṭ. vairīmaidī.1 ahurā.2
mazdā. ašā. srīrā. hiiaṭ.3 ī.
mainimadicā.4 vaocōimācā.5 vərəzimācā.6
yā. hātām. šīiaoθananām.7 vahištā.
xīiaṭ.8 ubōibiiā. ahubiiā.9

Y 35 4

1gauuōi. adāiš. tāiš. šīiaoθanāiš.2
yāiš. vahištāiš. fraēšīiāmahī.3
rāmācā. vāstrēmācā. dazdiiāi. surunuuatascā.4
asurunuuatascā.5 xšaiiaṇtascā.6
axšaiiaṇtascā.7

Y 35 5

1huxšaθrōtēmāi. bāṭ.2 xšaθrēm. ahmaṭ.
hiiaṭ.3 aibī.4 dadəmahicā.5 cīšmahicā.6
huuəmahicā.7 hiiaṭ.8 mazdāi.
ahurāi. ašāicā. vahištāi.: (huxšaθrōtēmāi.
.. vahištāi. ...:)

Portation steps: from WP4 to WP5 (1990)

(1st 16-bit encoding system)

```
Y 35 3
taβ¹β. aβ¹β. vairβ¿βmaidβ¿β. 11 ahurβ²β. 12
mazdβ²β. aβ-ββ²β. srβ¿βrβ²β. hiiab¹β. 13 β¿β.
mainimadicβ²β. 14 vaocβ@βimβ²βcβ²β. 15 vβ¿βrβ¿βzimβ²βcβ²β. 16
yβ²β. hβ²βtβ¿βm. β-βiiaob²βananβ¿βm. 17 vahib¹βtβ²β.
βAbiiβ²β¹β. 18 ubβ@βibiiβ²β. ahubiiβ²β. 19

Y 35 4
11gauuβ@βi. adβ²βiβ²β. tβ²βiβ²β. β-βiiaob²βanβ²βiβ²β. 12
yβ²βiβ²β. vahib¹βtβ²βiβ²β. fraβ²ββ-βiiβ²βmahβ¿β. 13
rβ²βmβ²βcβ²β. vβ²βstrβ¿βmcβ²β. dazdiiβ²βi. surunuuatascβ²β. 14
asurunuuatascβ²β. 15 xβ-βaiiaβ-βtascβ²β. 16
axβ-βaiiaβ-βtascβ²β. 17

Y 35 5
11huxβ-βaβ²βrβ@βtβ¿βmβ²βi. hβ²ββ¹β. 12 xβ-βaβ²βrβ¿βm. ahmaβ¹β.
hiiab¹β. 13 aibβ¿β. 14 dadβ¿βmahicβ²β. 15 cβ¿ββ-βmahicβ²β. 16
huuβ¿βmahicβ²β. 17 hiiab¹β. 18 mazdβ²βi.
ahurβ²βi. aβ-ββ²βicβ²β. vahib¹βtβ²βi. : <huxβ-βaβ²βrβ@βtβ¿βmβ²βi.
.. vahib¹βtβ²βi. ....>
```

```
Y 35 3
EALCEP CY 01 LCAE ta L> 0 L. a L> 0 L. vair L'a L maid L'a L. H 1- E- ahur L 1 L. H 2- E-
EALCEP CY 02 LCAE mazd L 1 L. a L q L L 1 L. sr L'a L r L 1 L. hiiab L> 0 L. H 3- E- L'a L
EALCEP CY 03 LCAE mainimadic L 1 L. H 4- E- vaoc L'N L im L 1 L c L 1 L. H 5- E- v L B L r L B L zi
EALCEP CY 04 LCAE y L 1 L. h L 1 L t L 1 L m. L L iiaob L> 1 L anan L 1 L m. H 7- E- vahib L 1 L t L 1 L
EALCEP CY 05 LCAE L L iiaob L> 0 L. H 8- E- ub L'N L libii L 1 L. ahubii L 1 L. H 9- E-

Y 35 4
EALCEP CY 06 LCAE H 1- E- gauu L'N L i. ad L 1 L i L L 1 L. t L 1 L i L L 1 L. L L iiaob L> 1 L an L 1 L i L
EALCEP CY 07 LCAE y L 1 L i L L 1 L. vahib L 1 L t L 1 L i L L 1 L. fra L'o L L L L iiaob L> 1 L mah L'a L. H 3- E-
EALCEP CY 08 LCAE r L 1 L m L 1 L c L 1 L. v L 1 L str L B L mc L 1 L. dazdii L 1 L i. surunuuatasc L
EALCEP CY 09
LCAE asurunuuatasc L 1 L. H 5- E- x L q L L iiaob L 1 L tasc L 1 L. H 6- E-
EALCEP CY 10
LCAE ax L q L L iiaob L 1 L tasc L 1 L. H 7- E-

Y 35 5
EALCEP CY 11 LCAE H 1- E- hux L q L L a L> 1 L r L'N L t L B L m L 1 L i. b L 1 L L L> 0 L. H 2- E- x L q L L a L>
EALCEP CY 12 LCAE hiiab L> 0 L. H 3- E- aib L'a L. H 4- E- dad L B L mahic L 1 L. H 5- E- c L'a L
EALCEP CY 13 LCAE huub L 1 L mahic L 1 L. H 7- E- hiiab L> 0 L. H 8- E- mazd L 1 L i.
ahur L 1 L i. a L q L L L 1 L i L L 1 L. vahib L 1 L t L 1 L i. : <hux L q L L a L> 1 L r L'N L t L B L m L 1 L i.
.. vahib L 1 L t L 1 L i. ....>
```

Portation steps: from WP5 to WP9 (automatic, but failing)

Y 35 3

taṭ. aṭ. vairīmaidī.¹ ahurā.²
mazdā. aṣā. srīrā. hīiaṭ.³ ī.
mainimadicā.⁴ vaocōimācā.⁵ vərəzimācā.⁶
yā. hātəm. šīiaoθananəm.⁷ vahištā.
xīiaṭ.⁸ ubōibīiā. ahubīiā.⁹

Y 35 4

¹gauuōi. adāiš. tāiš. šīiaoθanāiš.²
yāiš. vahištāiš. fraēšīiāmahi.³
rāmācā. vāstrəmcā. dazdīiāi. surunuuatascā.⁴
asurunuuatascā.⁵ xšaiiaṇtascā.⁶
axšaiiaṇtascā.⁷

Y 35 5

¹huxšaθrōtəmāi. bāt.² xšaθrəm. ahmaṭ.
hīiaṭ.³ aibī.⁴ dadəmahicā.⁵ cīšmahicā.⁶
hūuəmahicā.⁷ hīiaṭ.⁸ mazdāi.
ahurāi. aṣāicā. vahištāi.: (huxšaθrōtəmāi.
.. vahištāi.)

Y 35 3

ta). a). vairīmaidī.¹ ahurā.²
mazdā. aqā. srīrā. hīia).³ ī.
mainimadicā.⁴ vaocōimācā.⁵ vBrBzimācā.⁶
yā. hātəm. (īiaoūananəm.⁷ vahištā.
īiā).⁸ ubōibīiā. ahubīiā.⁹

Y 35 4

¹gauuōi. adāiš. tāiš. (īiaoūanāiš.²
yāiš. vahištāiš. fraē(īiāmahi.³
rāmācā. vāstrBmcā. dazdīiāi. surunuuatascā.⁴
asurunuuatascā.⁵ xqaiialtascā.⁶
axqaiialtascā.⁷

Y 35 5

¹huxqaūrōtBmāi. bā).² xqaūrBm. ahma).
hīia).³ aibī.⁴ dadBmahicā.⁵ cīqmahicā.⁶
hūuəmahicā.⁷ hīia).⁸ mazdāi.
ahurāi. aqāicā. vahištāi.: (huxqaūrōtBmāi.
.. vahištāi.)

Portation steps: from WP5 to Word 2000 (automatic, but failing)

Y 35 3

taṭ. aṭ. vairīmaidī.¹ ahurā.²
mazdā. aṣā. srīrā. hīaṭ.³ ī.
mainimadicā.⁴ vaocōimācā.⁵ vərəzimācā.⁶
yā. hātəm. šīiaoθananəm.⁷ vahištā.
xīiāṭ.⁸ ubōibīā. ahubīā.⁹

Y 35 4

¹gauuōi. adāiš. tāiš. šīiaoθanāiš.²
yāiš. vahištāiš. fraēšīāmahi.³
rāmācā. vāstrəmcā. dazdīiāi. surunuuatascā.⁴
asurunuuatascā.⁵ xšaiiaṇtascā.⁶
axšaiiaṇtascā.⁷

Y 35 5

¹huxšaθrōtəmāi. bāt.² xšaθrəm. ahmaṭ.
hīaṭ.³ aibī.⁴ dadəmahicā.⁵ cīšmahicā.⁶
hūaṇmahicā.⁷ hīaṭ.⁸ mazdāi.
ahurāi. aṣāicā. vahištāi.: (huxšaθrōtəmāi.
.. vahištāi.)

Y 35 3

ta). a). vairīmaidī.¹ ahurā.²
mazdā. aqā. srīrā. hīia).³ ī.
mainimadicā.⁴ vaocōimācā.⁵ vBrBzimācā.⁶
yā. hātəm. (īiaoθananəm.⁷ vahištā.
īiā).⁸ ubōibīā. ahubīā.⁹

Y 35 4

¹gauuōi. adāiš. tāiš. (īiaoθanāiš.²
yāiš. vahištāiš. fraē(īāmahi.³
rāmācā. vāstrBmcā. dazdīiāi. surunuuatascā.⁴
asurunuuatascā.⁵ xqaiialtascā.⁶
axqaiialtascā.⁷

Y 35 5

¹huxqaθrōtBmāi. bā).² xqaθrBm. ahma).
hīia).³ aibī.⁴ dadBmahicā.⁵ cīqmahicā.⁶
hūaṇmahicā.⁷ hīia).⁸ mazdāi.
ahurāi. aqāicā. vahištāi.: (huxqaθrōtBmāi.
.. vahištāi.)

Portation steps: from WP5 to HTML (UTF-8) (special programming required)

```
Y 35 3
EALCEPEYCECLEAsta>L. a>L. vairlâmaïdla>L. H1-ahurL2-
EALCEPEYCEALEAmazdL. a>L. srâlrL. hïia>L. H3-â
EALCEPEYCEBLEAmainimadicL. H4-vaocNlimLcL. H5-vBzrBzï
EALCEPEYCECLEAyL. hLltLm. L<liiao>lananLm. H7-vahiLltL
EALCEPEYCEIDLEAeliLliL>L. H8-ubNlibiiL. ahubiiL. H9-

Y 35 4
EALCEPEYCHGLEA>H1-gauuNli. adLliL. tLliL. L<liiao>lanLliL
EALCEPEYCEHLEAyLliL. vahiLltLliL. fraLoL<liiL>mahla>L. H3-
EALCEPEYCEZILEAerLliLmLcL. vLlstrBmcL. dazdiiLli. surunuuatascL
EALCEPEYCELEAsurunuuatascL. H5-xLqLaiiaLltascL. H6-
EALCEPEYCELEAaxLqLaiiaLltascL. H7-

Y 35 5
EALCEPEYCEMLEA>H1-huxLqLa>HlrNltBmLli. bLliL>L. H2-xLqLa>
EALCEPEYCEMLEA>hïia>L. H3-aibla>L. H4-dadBmahicL. H5-cla>L
EALCEPEYCEVLEAhuuLmahicL. H7-hïia>L. H8-mazdLli.
ahurLli. aLliLlicL. vahiLltLli. vahiLltLli.: <huxLqLa>HlrNltBmLli.
.. vahiLltLli. ....>
```

```
Y 35 3 <BR>
tati°. atli°. vairA«maidA». <SUP><FONT SIZE=-3>1</FONT></SUP> ahurA. <SUP><FONT SIZE=-3>2</FONT></SUP><BR>
mazdA. aAjiEÄ. srA«rA. hïiatl. <SUP><FONT SIZE=-3>3</FONT></SUP> Ä. <BR>
mainimadicA. <SUP><FONT SIZE=-3>4</FONT></SUP> vaocÄlimÄcÄ. <SUP><FONT SIZE=-3>5</FONT></SUP> véTyrÉTzimÄcÄ. <SUP><FONT SIZE=-3>6</FONT></SUP></S
yÄ. hÄltÄ. ...m. Äjiliiao'ananÄ. ...m. <SUP><FONT SIZE=-3>7</FONT></SUP> vahiÄjtÄ. <BR>
xiliÄli°. <SUP><FONT SIZE=-3>8</FONT></SUP> ubÄlibiiÄ. ahubiiÄ. <SUP><FONT SIZE=-3>9</FONT></SUP><BR>
<BR>
Y 35 4 <BR>
<SUP><FONT SIZE=-3>1</FONT></SUP>gauuÄ. adÄliÄj. tÄliÄj. Äjiliiao'ananÄliÄj. <SUP><FONT SIZE=-3>2</FONT></SUP><BR>
yÄliÄj. vahiÄjtÄliÄj. fraÄÄjiliÄmahÄ. <SUP><FONT SIZE=-3>3</FONT></SUP><BR>
rÄmÄcÄ. vÄstrÉ™mcÄ. dazdiiÄ. surunuuatascÄ. <SUP><FONT SIZE=-3>4</FONT></SUP><BR>
asurunuuatascÄ. <SUP><FONT SIZE=-3>5</FONT></SUP> xÄjEaiiaä'ttascÄ. <SUP><FONT SIZE=-3>6</FONT></SUP><BR>
axÄjEaiiaä'ttascÄ. <SUP><FONT SIZE=-3>7</FONT></SUP><BR>
<BR>
Y 35 5 <BR>
<SUP><FONT SIZE=-3>1</FONT></SUP>huxÄjEal'rÄltÉ™mÄli. bÄli. <SUP><FONT SIZE=-3>2</FONT></SUP> xÄjEal'rÉ™m. ahmatl. <BR>
hïiatl°. <SUP><FONT SIZE=-3>3</FONT></SUP> aibÄ. <SUP><FONT SIZE=-3>4</FONT></SUP> dadÉ™mahicÄ. <SUP><FONT SIZE=-3>5</FONT></SUP> cÄÄjEämahic,
huuÄ. ...mahicÄ. <SUP><FONT SIZE=-3>7</FONT></SUP> hïiatl°. <SUP><FONT SIZE=-3>8</FONT></SUP> mazdÄ. <BR>
ahurÄ. aÄjEÄlicÄ. vahiÄjtÄli.: (huxÄjEal'rÄltÉ™mÄli. <BR>
.. vahiÄjtÄli. ....)<BR>
```

Portation steps: from WP5 to HTML (UTF-8) (special programming required)

Y 35 3

tatî°. atî°. vairÄ«maidÄ«. ^{1} ahurÄll. ^{2}

mazdÄll. aÄj]ËÄll. srÄ«rÄll. hiatî°. ^{3} Ä«.

mainimadicÄll. ^{4} vaocÄllimÄllcÄll. ^{5} vé¹mÉ¹mzimÄllcÄll. ^{6}</S

yÄll. hÄlltÄ...m. Äj]lliao¹ananÄ...m. ^{7} vahiÄj]tÄll.

x]lliaÄll]¹. ^{8} ubÄllibiiÄll. ahubiiÄll. ^{9}

Y 35 4

^{1}gauuÄll]i. adÄlliaÄj. tÄlliaÄj. Äj]lliao¹anÄlliaÄj. ^{2}

yÄlliaÄj. vahiÄj]tÄlliaÄj. fraÄ¹Äj]lliaÄllmahÄ«. ^{3}

rÄllmÄllcÄll. vÄllstrÉ¹mÄll. dazdiiÄll. surunuuatascÄll. ^{4}

asurunuuatascÄll. ^{5} xÄj]Ëaiiaá¹ttascÄll. ^{6}

axÄj]Ëaiiaá¹ttascÄll. ^{7}

Y 35 5

^{1}huxÄj]Ëa¹rÄllÉ¹mÄll. bÄll]¹. ^{2} xÄj]Ëa¹rÉ¹m. ahmatî°.

hiatî°. ^{3} aibÄ«. ^{4} dadÉ¹mahicÄll. ^{5} cÄ«Äj]Ëmahic

huuÄ...mahicÄll. ^{7} hiatî¹. ^{8} mazdÄll.

ahurÄll. aÄj]ËÄllcÄll. vahiÄj]tÄll. : (huxÄj]Ëa¹rÄllÉ¹mÄll.

.. vahiÄj]tÄll.)

Rendering:

UTF-8 = representation of UNICODE

(still requires special fonts, e.g. Arial Unicode MS)

Y 35 3

tatī°. atī°. vairÄ«maidÄ«. ^{1}
mazdÄ. aÄjġÄ. srÄ«rÄ. hiiatī°. ^{3}
mainimadicÄ. ^{4} vao
yÄ. hÄtÄ...m. Äjġġiaoġ'ananÄ...m. ^{7}
xġġiiÄtī°. ^{8} ubÄġlibiiÄ

Y 35 4

^{1}gauuÄ. adÄġiÄj. t
yÄġiÄj. vahiÄjtÄġiÄj. fraÄ"ÄjġġiiÄmahÄ«. ^{3}
rÄġmÄġcÄ. vÄġstrÉ™mcÄ. dazdiiÄ. surunuuatascÄ. ^{5}
asurunuuatascÄ. ^{5}
axÄjġġaiiaá'ġtascÄ. ^{7}

Y 35 5

^{1}huxÄjġġai'rÄġtÉ™r
hiiatī°. ^{3} aibÄ«. ^{3}
huuÄ...mahicÄ. ^{7} hi
ahurÄ. aÄjġÄġicÄ. vahiÄjtÄ. : (huxÄjġġai'rÄġtÉ™mÄ.

.. vahiÄjtÄ.)

Y 35 3

taṭ. aṭ. vairīmaidī.¹ ahurā.²
mazdā. aṣā. srīrā. hiiat.³ ī.
mainimadicā.⁴ vaocōimācā.⁵ vərəzimācā.⁶
yā. hātaṃ. ṣiiaoḡananāṃ.⁷ vahištā.
xiiāt.⁸ ubōibiiā. ahubiiā.⁹

Y 35 4

¹gauuōi. adāiš. tāiš. ṣiiaoḡanāiš.²
yāiš. vahištāiš. fraēṣiiāmahi.³
rāmācā. vāstrēmca. dazdiiāi. surunuuatascā.⁴
asurunuuatascā.⁵ xṣaiiaṇtascā.⁶
axṣaiiaṇtascā.⁷

Y 35 5

¹huxṣaḡrōtēmāi. bāt.² xṣaḡrēm. ahmaṭ.
hiiat.³ aibī.⁴ dadəmahicā.⁵ cīṣmahicā.⁶
huuqamahicā.⁷ hiiat.⁸ mazdāi.
ahurāi. aṣāicā. vahištāi. : (huxṣaḡrōtēmāi.
.. vahištāi.)

N.B.

- We have not talked about original scripts yet!

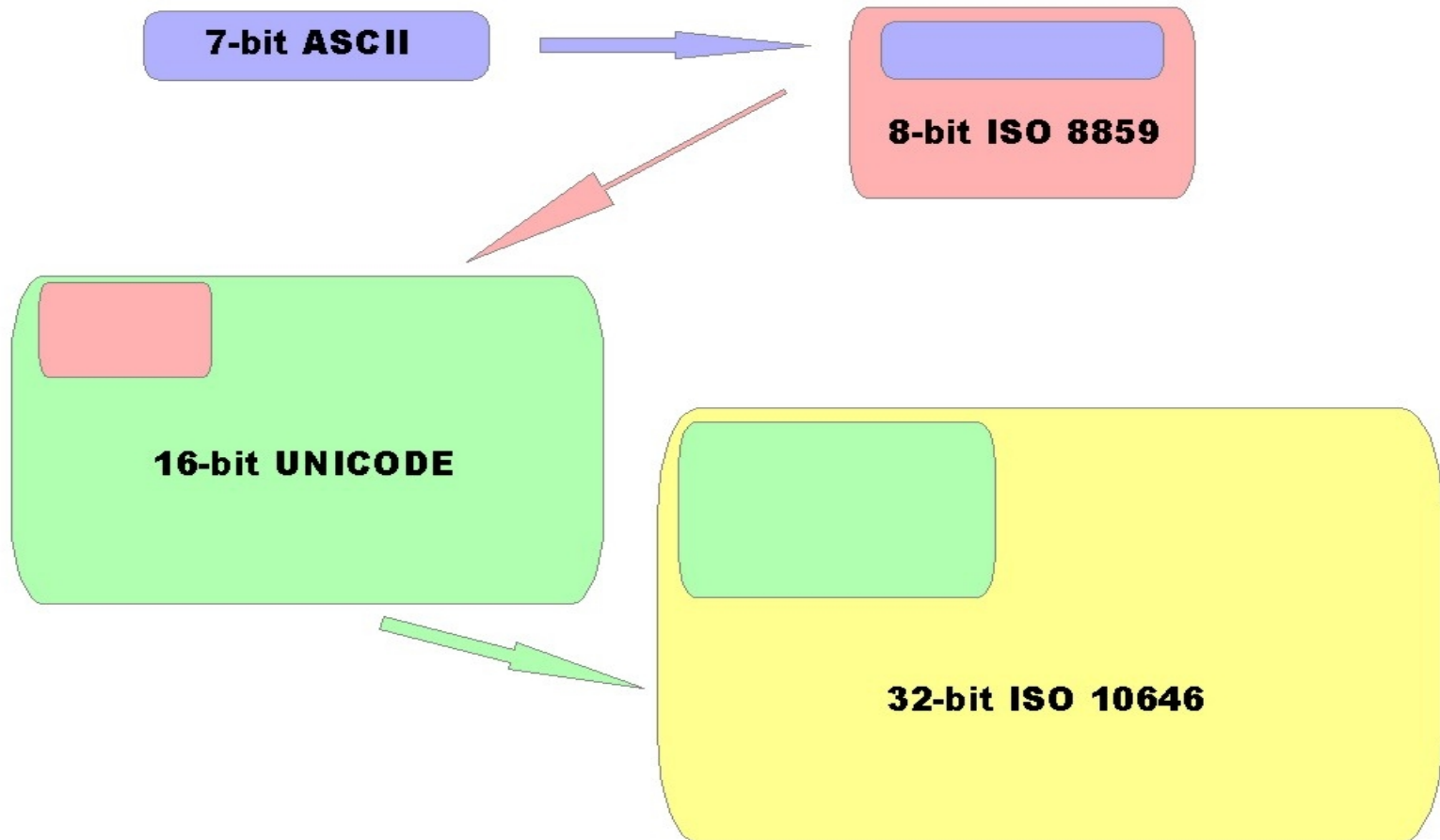
Encoding of Characters summarized

- Restrictions of encoding standards
 - 7-bit: max. 128 different characters
 - 8-bit: max. 256 diff. characters (>> “1-byte”)
 - 16-bit: max. 65536 diff. chars. (>> “2-byte”)
 - 32-bit: max. 4,294,967,296 d.c. (>> “4-byte”)

Encoding of Characters

- Scope of encoding standards
 - 7-bit: “EBCDIC”, “ASCII”
 - Mainframe computers, first generation of PCs, web applications (URLs, e-mail) until recently
 - 8-bit: “IBM”, “Mac-OS”, “ANSI”, “ISO 8859”:
“Codepages”
 - PCs (DOS, Windows < NT 4), Macs (< OS 10), web applications (URLs, web pages, e-mail) of today
 - 16-bit: “WordPerfect 5 ff.” (incomplete), “UNICODE”
 - PCs (Windows > NT 3), Macs (> OS 9), web applications (URLs since 2003, web pages since 1995, e-mail)
 - 32-bit: “ISO 10646”, “extended UNICODE”
 - Restricted use

Encoding of Characters: Portation of standards



Encoding of Characters: Character repertoires

7-bit encoding: ASCII

	0											1										
	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9		
000											TAB	LF										
020											ESC		␣	!	"	#	\$	%	&	'		
040	()	*	+	,	-	.	/	0	1	2	3	4	5	6	7	8	9	:	;		
060	<	=	>	?	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O		
080	P	Q	R	S	T	U	V	W	X	Y	Z	[\]	^	_	‘	a	b	c		
100	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w		
120	x	y	z	{		}	~	␣	(end)													
	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9		
	0											1										

Encoding of Characters: Character repertoires

8-bit encoding: "Extended ANSI / "WINDOWS-encoding" / Codepage 1252)

	0											1										
	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9		
000											TAB	LF										
020											ESC											
040	()	*	+	,	-	.	/	0	1	2	3	4	5	6	7	8	9	:	;		
060	<	=	>	?	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O		
080	P	Q	R	S	T	U	V	W	X	Y	Z	[\]	^	_	`	a	b	c		
100	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w		
120	x	y	z	{		}	~					,	f	„	...	†	‡	^	%	Š	<	
140	Œ							‘	’	“	”	•	—	—	~	™	š	>	œ	ÿ		
160		ı	¢	£	⚙	¥		§	”	©	^a	«	¬	-	®	-	°	±	²	³		
180	,	μ	¶	—	,	¹	^o	»	¼	½	¾	¿	À	Á	Â	Ã	Ä	Å	Æ	Ç		
200	È	É	Ê	Ë	Ì	Í	Î	Ï	Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û		
220	Ü	Ý	Þ	ß	à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï		
240	ð	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	ÿ	(end)					
	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9		
	0											1										

Encoding of Characters:

“Official” font mapping: ISO 8859 Codepages

32		!	"	#	\$	%	&	'	()	*	+	,	-	.	/	47
48	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?	63
64	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	79
80	P	Q	R	S	T	U	V	W	X	Y	Z	[\]	^	_	95
96	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	111
112	p	q	r	s	t	u	v	w	x	y	z	{		}	~	□	127
160		ı	¢	£	¤	¥	¦	§	¨	©	ª	«	¬	-	®	¯	175
176	°	±	²	³	´	µ	¶	·	,	¹	º	»	¼	½	¾	¿	191
192	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï	207
208	Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ	ß	223
224	à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï	239
240	ð	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	ÿ	255

32		!	"	#	\$	%	&	'	()	*	+	,	-	.	/	47
48	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?	63
64	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	79
80	P	Q	R	S	T	U	V	W	X	Y	Z	[\]	^	_	95
96	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	111
112	p	q	r	s	t	u	v	w	x	y	z	{		}	~	□	127
160		Ë	Ђ	Ѓ	Є	Ѕ	І	Ї	Ј	Љ	Њ	Ћ	Ќ	-	Ў	Џ	175
176	А	Б	В	Г	Д	Е	Ж	З	И	Й	К	Л	М	Н	О	П	191
192	Р	С	Т	У	Ф	Х	Ц	Ч	Ш	Щ	Ъ	Ы	Ь	Э	Ю	Я	207
208	а	б	в	г	д	е	ж	з	и	й	к	л	м	н	о	п	223
224	р	с	т	у	ф	х	ц	ч	ш	щ	ъ	ы	ь	э	ю	я	239
240	№	ё	ђ	ѓ	є	ѕ	і	ї	ј	љ	њ	ћ	ќ	џ	џ	џ	255

Encoding of Characters:

“Official” font mapping: ISO 8859

ISO-8859-subset	ISO name	Name in Netscape 4.x	Name in MS Internet Explorer 4.x
ISO-8859-1	Latin-1	Western (ISO-8859-1)	Western Alphabet (USA/Western Europe)
ISO-8859-2	Latin-2	Central European (ISO-8859-2)	Central European Alphabet (ISO)
ISO-8859-3	Latin-3		
ISO-8859-4	Latin-4		
ISO-8859-5	Cyrillic	Cyrillic (ISO-8859-5)	Cyrillic (ISO)
ISO-8859-6	Arabic		
ISO-8859-7	Greek	Greek (ISO-8859-7)	
ISO-8859-8	Hebrew		
ISO-8859-9	Latin-5	Turkish (ISO-8859-9)	
ISO-8859-10	Latin-6		

- <http://titus.uni-frankfurt.de/unicode/iso8859/iso8859.htm>
- N.B. Other encoding standards for Cyrillic, Greek etc. acknowledged

Encoding of Characters:

Scripts covered by UNICODE (5.0)

The Unicode Character Code Charts By Script				
SYMBOLS AND PUNCTUATION NAME INDEX HELP AND LINKS				
European Alphabets	African Scripts	Indic Scripts	East Asian Scripts	Central Asian Scripts
(see also Comb. Marks)	Ethiopic	Bengali	Han Ideographs	Kharoshthi
Armenian	Ethiopic	Devanagari	Unified CJK Ideographs (5MB)	Mongolian
Armenian	Ethiopic Supplement	Gujarati	CJK Ideographs Ext. A (2MB)	Phags-Pa
<i>Armenian Ligatures</i>	Ethiopic Extended	Gurmukhi	CJK Ideographs Ext. B (13MB)	Tibetan
Coptic	Other African scripts	Kannada	Compatibility Ideographs (.5MB)	
Coptic	N'Ko	Limbu	... Supplement (.5MB)	
<i>Coptic in Greek block</i>	Tifinagh	Malayalam	Kanbun	
Cyrillic	Middle Eastern Scripts	Oriya	(see also Unihan Database)	Ancient Scripts
Cyrillic	Arabic	Sinhala	Radicals and Strokes	Ancient Greek
Cyrillic Supplement	Arabic	Syloti Nagri	CJK Radicals	Ancient Greek Numbers
Georgian	Arabic Supplement	Tamil	KangXi Radicals	Ancient Greek Musical
Georgian	Arabic Presentation Forms A	Telugu	CJK Strokes	Cuneiform
Georgian Supplement	Arabic Presentation Forms B		Ideographic Description	Cuneiform
Greek	Hebrew	Philippine Scripts	Chinese-specific	Cuneiform Numbers
Greek	Hebrew	Buhid	Bopomofo	Old Persian
Greek Extended	<i>Hebrew Presentation Forms</i>	Hanunoo	Bopomofo Extended	Ugaritic
(see also Ancient Greek)	Syriac	Tagalog	Japanese-specific	Linear B
Latin	Syriac	Tagbanwa	Hiragana	Linear B Syllabary
Basic Latin	Thaana		Katakana,	Linear B Ideograms
Latin-1	Thaana	South East Asian	Katakana Phonetic Ext.	Other Ancient Scripts
Latin Extended A	American scripts	Buginese	<i>Halfwidth Katakana</i>	Aegean Numbers
Latin Extended B	Canadian Syllabics	Balinese	Korean-specific	Counting Rod Numerals
Latin Extended C	Cherokee	Khmer	Hangul Syllables (4MB)	Cypriot Syllabary
Latin Extended D	Deseret	Khmer Symbols	Hangul Jamo	Gothic
Latin Extended Additional	Other Scripts	Lao	Hangul Compatibility Jamo	Old Italic
<i>Latin Ligatures</i>	Shavian	Myanmar	<i>Halfwidth Jamo</i>	Ogham
<i>Fullwidth Latin Letters</i>	Osmanya	New Tai Lue	Yi	Runic
Small Forms	Glagolitic	Tai Le	Yi (.6MB)	Phoenician
(see also Phonetic Symbols)		Thai	Yi Radicals	

- <http://www.unicode.org/>

Encoding of Characters: UNICODE blocks

<u>00</u>	<u>01</u>	<u>02</u>	<u>03</u>	<u>04</u>	<u>05</u>	<u>06</u>	-	-	<u>09</u>	<u>0A</u>	<u>0B</u>	<u>0C</u>	<u>0D</u>	<u>0E</u>	<u>0F</u>
<u>10</u>	<u>11</u>	<u>12</u>	<u>13</u>	<u>14</u>	<u>15</u>	<u>16</u>	-	-	-	-	-	-	-	<u>1E</u>	<u>1F</u>
<u>20</u>	<u>21</u>	<u>22</u>	<u>23</u>	<u>24</u>	<u>25</u>	<u>26</u>	<u>27</u>	28	-	-	-	-	-	-	-
<u>30</u>	<u>31</u>	<u>32</u>	<u>33</u>	<u>34</u>	<u>35</u>	<u>36</u>	<u>37</u>	<u>38</u>	<u>39</u>	<u>3A</u>	<u>3B</u>	<u>3C</u>	<u>3D</u>	<u>3E</u>	<u>3F</u>
<u>40</u>	<u>41</u>	<u>42</u>	<u>43</u>	<u>44</u>	<u>45</u>	<u>46</u>	<u>47</u>	<u>48</u>	<u>49</u>	<u>4A</u>	<u>4B</u>	<u>4C</u>	<u>4D</u>	<u>4E</u>	<u>4F</u>
<u>50</u>	<u>51</u>	<u>52</u>	<u>53</u>	<u>54</u>	<u>55</u>	<u>56</u>	<u>57</u>	<u>58</u>	<u>59</u>	<u>5A</u>	<u>5B</u>	<u>5C</u>	<u>5D</u>	<u>5E</u>	<u>5F</u>
<u>60</u>	<u>61</u>	<u>62</u>	<u>63</u>	<u>64</u>	<u>65</u>	<u>66</u>	<u>67</u>	<u>68</u>	<u>69</u>	<u>6A</u>	<u>6B</u>	<u>6C</u>	<u>6D</u>	<u>6E</u>	<u>6F</u>
<u>70</u>	<u>71</u>	<u>72</u>	<u>73</u>	<u>74</u>	<u>75</u>	<u>76</u>	<u>77</u>	<u>78</u>	<u>79</u>	<u>7A</u>	<u>7B</u>	<u>7C</u>	<u>7D</u>	<u>7E</u>	<u>7F</u>
<u>80</u>	<u>81</u>	<u>82</u>	<u>83</u>	<u>84</u>	<u>85</u>	<u>86</u>	<u>87</u>	<u>88</u>	<u>89</u>	<u>8A</u>	<u>8B</u>	<u>8C</u>	<u>8D</u>	<u>8E</u>	<u>8F</u>
<u>90</u>	<u>91</u>	<u>92</u>	<u>93</u>	<u>94</u>	<u>95</u>	<u>96</u>	<u>97</u>	<u>98</u>	<u>99</u>	<u>9A</u>	<u>9B</u>	<u>9C</u>	<u>9D</u>	<u>9E</u>	<u>9F</u>
-	-	-	-	-	-	-	-	-	-	-	-	<u>AC</u>	<u>AD</u>	<u>AE</u>	<u>AF</u>
<u>B0</u>	<u>B1</u>	<u>B2</u>	<u>B3</u>	<u>B4</u>	<u>B5</u>	<u>B6</u>	<u>B7</u>	<u>B8</u>	<u>B9</u>	<u>BA</u>	<u>BB</u>	<u>BC</u>	<u>BD</u>	<u>BE</u>	<u>BF</u>
<u>C0</u>	<u>C1</u>	<u>C2</u>	<u>C3</u>	<u>C4</u>	<u>C5</u>	<u>C6</u>	<u>C7</u>	<u>C8</u>	<u>C9</u>	<u>CA</u>	<u>CB</u>	<u>CC</u>	<u>CD</u>	<u>CE</u>	<u>CF</u>
<u>D0</u>	<u>D1</u>	<u>D2</u>	<u>D3</u>	<u>D4</u>	<u>D5</u>	<u>D6</u>	<u>D7</u>	-	-	-	-	-	-	-	-
<u>E0</u>	<u>E1</u>	<u>E2</u>	<u>E3</u>	<u>E4</u>	<u>E5</u>	<u>E6</u>	<u>E7</u>	<u>E8</u>	<u>E9</u>	<u>EA</u>	<u>EB</u>	<u>EC</u>	<u>ED</u>	<u>EE</u>	<u>EF</u>
<u>F0</u>	<u>F1</u>	<u>F2</u>	<u>F3</u>	<u>F4</u>	<u>F5</u>	<u>F6</u>	<u>F7</u>	<u>F8</u>	<u>F9</u>	<u>FA</u>	<u>FB</u>	<u>FC</u>	<u>FD</u>	<u>FE</u>	<u>FF</u>

- <http://titus.uni-frankfurt.de/unicode/unitest.htm>
- <http://www.unicode.org/>

Encoding of Characters: UNICODE blocks (examples)

<	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F	>
000																	
001																	
002		!	"	#	\$	%	&	'	()	*	+	,	-	.	/	
003	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?	
004	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	
005	P	Q	R	S	T	U	V	W	X	Y	Z	[\]	^	_	
006	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	
007	p	q	r	s	t	u	v	w	x	y	z	{		}	~		
008																	
009																	
00A		ı	¢	£	¤	¥	¦	§	¨	©	ª	«	¬	—	®	¯	
00B	°	±	²	³	´	µ	¶	·	¸	¹	º	»	¼	½	¾	¿	
00C	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï	
00D	Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ	ß	
00E	à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï	
00F	ð	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	ÿ	

[illegible]

Encoding of Characters: UNICODE blocks (examples)

<	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F	>
040	È	É	Ê	Ë	Ì	Í	Î	Ï	Ð	Ñ	Ò	Ó	Ô	Õ	Ö	Ù	Ú
041	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï	Ð
042	Р	С	Т	У	Ф	Х	Ц	Ч	Ш	Щ	Ъ	Ы	Ь	Э	Ю	Я	
043	а	б	в	г	д	е	ж	з	и	й	к	л	м	н	о	п	
044	р	с	т	у	ф	х	ц	ч	ш	щ	ъ	ы	ь	э	ю	я	
045	è	é	ê	ë	ì	í	î	ï	ð	ñ	ò	ó	ô	õ	ö	ù	ú
046	Œ	œ	ƒ	℥	ℳ	℥	ℳ	℥	ℳ	℥	ℳ	℥	ℳ	℥	ℳ	℥	ℳ
047	Ψ	ψ	Θ	θ	V	v	Û	ü	Ÿ	ÿ	Ω	ω	Ω	ω	Ω	ω	Ω
048	Ç	ç	*	†	‡	§	¶	•	◊	◊	Й	й	Ъ	ъ	Р	р	
049	Г	г	Г	г	Б	б	Ж	ж	З	з	К	к	К	к	К	к	
04A	К	к	Н	н	Н	н	П	п	Q	q	С	с	Т	т	У	у	
04B	Y	y	X	x	Ц	ц	Ч	ч	Ч	ч	h	h	е	е	е	е	
04C	I	Ж	ж	Б	б	Л	л	Н	н	Н	н	Ч	ч	М	м		
04D	Ă	ă	Ä	ä	Æ	æ	Ë	ë	Ə	ə	Ë	ë	Ж	ж	Э	э	
04E	З	з	И	й	Й	й	Ö	ö	Ө	ө	Ө	ө	Э	э	У	у	
04F	Û	ü	Û	ü	Č	č					Б	б					

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F	
060																	
061																	
062		ء	أ	إ	ؤ	ئ	ب	ة	ت	ث	ج	ح	خ	د			
063	ذ	ر	ز	س	ش	ص	ض	ط	ظ	ع	غ						
064	-	ق	ف	ك	ل	م	ن	و	ه	ي							
065																	
066	٠	١	٢	٣	٤	٥	٦	٧	٨	٩	٪	,	,	*	ب	و	
067		أ	إ		أ	ؤ	ئ	ب	ة	ت	ث	ج	ح	خ	د		
068	پ	خ	ج	چ	خ	چ	چ	د	د	د	د	پ	پ	پ	پ	پ	پ
069	ظ	ظ	ظ	ظ	ظ	ظ	ظ	ظ	ظ	ظ	ظ	ظ	ظ	ظ	ظ	ظ	ظ
06A	گ	گ	گ	گ	گ	گ	گ	گ	گ	گ	گ	گ	گ	گ	گ	گ	گ
06B	خ	ه	ن	ن	ن	ن	ن	ن	ن	ن	ن	ن	ن	ن	ن	ن	ن
06C	و	ی	ی	ی	ی	ی	ی	ی	ی	ی	ی	ی	ی	ی	ی	ی	ی
06D	ی	ی	ی	ی	ی	ی	ی	ی	ی	ی	ی	ی	ی	ی	ی	ی	ی
06E																	
06F	٠	١	٢	٣	٤	٥	٦	٧	٨	٩	ش	ض	غ	ء	م		

Encoding of Characters: UNICODE blocks (examples)

00	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F	00
FB0	ff	fi	fl	ffi	ffl	ft	st	Ⓢ	Ⓣ	Ⓤ	Ⓥ	Ⓦ	Ⓧ	Ⓨ	Ⓩ	ⓐ	
FB1	ⓐ	ⓑ	ⓔ	ⓕ	ⓖ	ⓗ	ⓙ	ⓚ	ⓛ	ⓞ	ⓟ	ⓠ	ⓡ	ⓢ	ⓤ	ⓦ	
FB2	ⓧ	ⓨ	ⓩ	⓪	⓫	⓬	⓭	⓮	⓯	⓰	⓱	⓲	⓳	⓴	⓵	⓶	
FB3	⓷	⓸	⓹	⓺	⓻	⓼	⓽	⓾	⓿	ⓠ	ⓡ	⓴	⓶	⓸	⓺	⓼	
FB4	⓽	⓿	ⓠ	ⓡ	⓴	⓶	⓸	⓺	⓼	⓶	⓸	⓺	⓼	⓶	⓸	⓺	
FB5	⓽	⓿	ⓠ	ⓡ	⓴	⓶	⓸	⓺	⓼	⓶	⓸	⓺	⓼	⓶	⓸	⓺	
FB6	⓽	⓿	ⓠ	ⓡ	⓴	⓶	⓸	⓺	⓼	⓶	⓸	⓺	⓼	⓶	⓸	⓺	
FB7	⓽	⓿	ⓠ	ⓡ	⓴	⓶	⓸	⓺	⓼	⓶	⓸	⓺	⓼	⓶	⓸	⓺	
FB8	⓽	⓿	ⓠ	ⓡ	⓴	⓶	⓸	⓺	⓼	⓶	⓸	⓺	⓼	⓶	⓸	⓺	
FB9	⓽	⓿	ⓠ	ⓡ	⓴	⓶	⓸	⓺	⓼	⓶	⓸	⓺	⓼	⓶	⓸	⓺	
FBA	⓽	⓿	ⓠ	ⓡ	⓴	⓶	⓸	⓺	⓼	⓶	⓸	⓺	⓼	⓶	⓸	⓺	
FBB	⓽	⓿	ⓠ	ⓡ	⓴	⓶	⓸	⓺	⓼	⓶	⓸	⓺	⓼	⓶	⓸	⓺	
FBC	⓽	⓿	ⓠ	ⓡ	⓴	⓶	⓸	⓺	⓼	⓶	⓸	⓺	⓼	⓶	⓸	⓺	
FBD	⓽	⓿	ⓠ	ⓡ	⓴	⓶	⓸	⓺	⓼	⓶	⓸	⓺	⓼	⓶	⓸	⓺	
FBE	⓽	⓿	ⓠ	ⓡ	⓴	⓶	⓸	⓺	⓼	⓶	⓸	⓺	⓼	⓶	⓸	⓺	
FBF	⓽	⓿	ⓠ	ⓡ	⓴	⓶	⓸	⓺	⓼	⓶	⓸	⓺	⓼	⓶	⓸	⓺	

00	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F	00
FC0	ⓐ	ⓑ	ⓔ	ⓕ	ⓖ	ⓗ	ⓙ	ⓚ	ⓛ	ⓞ	ⓟ	ⓠ	ⓡ	⓴	⓶	⓸	
FC1	ⓐ	ⓑ	ⓔ	ⓕ	ⓖ	ⓗ	ⓙ	ⓚ	ⓛ	ⓞ	ⓟ	ⓠ	ⓡ	⓴	⓶	⓸	
FC2	ⓐ	ⓑ	ⓔ	ⓕ	ⓖ	ⓗ	ⓙ	ⓚ	ⓛ	ⓞ	ⓟ	ⓠ	ⓡ	⓴	⓶	⓸	
FC3	ⓐ	ⓑ	ⓔ	ⓕ	ⓖ	ⓗ	ⓙ	ⓚ	ⓛ	ⓞ	ⓟ	ⓠ	ⓡ	⓴	⓶	⓸	
FC4	ⓐ	ⓑ	ⓔ	ⓕ	ⓖ	ⓗ	ⓙ	ⓚ	ⓛ	ⓞ	ⓟ	ⓠ	ⓡ	⓴	⓶	⓸	
FC5	ⓐ	ⓑ	ⓔ	ⓕ	ⓖ	ⓗ	ⓙ	ⓚ	ⓛ	ⓞ	ⓟ	ⓠ	ⓡ	⓴	⓶	⓸	
FC6	ⓐ	ⓑ	ⓔ	ⓕ	ⓖ	ⓗ	ⓙ	ⓚ	ⓛ	ⓞ	ⓟ	ⓠ	ⓡ	⓴	⓶	⓸	
FC7	ⓐ	ⓑ	ⓔ	ⓕ	ⓖ	ⓗ	ⓙ	ⓚ	ⓛ	ⓞ	ⓟ	ⓠ	ⓡ	⓴	⓶	⓸	
FC8	ⓐ	ⓑ	ⓔ	ⓕ	ⓖ	ⓗ	ⓙ	ⓚ	ⓛ	ⓞ	ⓟ	ⓠ	ⓡ	⓴	⓶	⓸	
FC9	ⓐ	ⓑ	ⓔ	ⓕ	ⓖ	ⓗ	ⓙ	ⓚ	ⓛ	ⓞ	ⓟ	ⓠ	ⓡ	⓴	⓶	⓸	
FCA	ⓐ	ⓑ	ⓔ	ⓕ	ⓖ	ⓗ	ⓙ	ⓚ	ⓛ	ⓞ	ⓟ	ⓠ	ⓡ	⓴	⓶	⓸	
FCB	ⓐ	ⓑ	ⓔ	ⓕ	ⓖ	ⓗ	ⓙ	ⓚ	ⓛ	ⓞ	ⓟ	ⓠ	ⓡ	⓴	⓶	⓸	
FCC	ⓐ	ⓑ	ⓔ	ⓕ	ⓖ	ⓗ	ⓙ	ⓚ	ⓛ	ⓞ	ⓟ	ⓠ	ⓡ	⓴	⓶	⓸	
FCD	ⓐ	ⓑ	ⓔ	ⓕ	ⓖ	ⓗ	ⓙ	ⓚ	ⓛ	ⓞ	ⓟ	ⓠ	ⓡ	⓴	⓶	⓸	
FCE	ⓐ	ⓑ	ⓔ	ⓕ	ⓖ	ⓗ	ⓙ	ⓚ	ⓛ	ⓞ	ⓟ	ⓠ	ⓡ	⓴	⓶	⓸	
FCF	ⓐ	ⓑ	ⓔ	ⓕ	ⓖ	ⓗ	ⓙ	ⓚ	ⓛ	ⓞ	ⓟ	ⓠ	ⓡ	⓴	⓶	⓸	

Recent additions: UNICODE blocks (examples)

2D00

Georgian Supplement

2D2F

	2D0	2D1	2D2
0	Ⴀ	Ⴁ	Ⴂ
1	Ⴃ	Ⴄ	Ⴅ
2	Ⴆ	Ⴇ	Ⴈ
3	Ⴉ	Ⴊ	Ⴋ
4	Ⴌ	Ⴍ	Ⴎ
5	Ⴏ	Ⴐ	Ⴑ
6	Ⴒ	Ⴓ	Ⴔ
7	Ⴕ	Ⴖ	Ⴗ
8	Ⴘ	Ⴙ	Ⴚ
9	Ⴛ	Ⴜ	Ⴝ
A	Ⴞ	Ⴟ	Ⴀ
B	Ⴁ	Ⴂ	Ⴃ
C	Ⴄ	Ⴅ	Ⴆ
D	Ⴇ	Ⴈ	Ⴉ
E	Ⴊ	Ⴋ	Ⴌ
F	Ⴍ	Ⴎ	Ⴏ

Small letters (Khutsuri)

This is the lowercase of the old ecclesiastical alphabet. See the Georgian block for uppercase *Aso margvuli*.

2D00	Ⴀ	GEORGIAN SMALL LETTER AN
2D01	Ⴁ	GEORGIAN SMALL LETTER BAN
2D02	Ⴂ	GEORGIAN SMALL LETTER GAN
2D03	Ⴃ	GEORGIAN SMALL LETTER DON
2D04	Ⴄ	GEORGIAN SMALL LETTER EN
2D05	Ⴅ	GEORGIAN SMALL LETTER VIN
2D06	Ⴆ	GEORGIAN SMALL LETTER ZEN
2D07	Ⴇ	GEORGIAN SMALL LETTER TAN
2D08	Ⴈ	GEORGIAN SMALL LETTER IN
2D09	Ⴉ	GEORGIAN SMALL LETTER KAN
2D0A	Ⴊ	GEORGIAN SMALL LETTER LAS
2D0B	Ⴋ	GEORGIAN SMALL LETTER MAN
2D0C	Ⴌ	GEORGIAN SMALL LETTER NAR
2D0D	Ⴍ	GEORGIAN SMALL LETTER ON
2D0E	Ⴎ	GEORGIAN SMALL LETTER PAR
2D0F	Ⴏ	GEORGIAN SMALL LETTER ZHAR
2D10	Ⴐ	GEORGIAN SMALL LETTER RAE
2D11	Ⴑ	GEORGIAN SMALL LETTER SAA
2D12	Ⴒ	GEORGIAN SMALL LETTER TAR
2D13	Ⴓ	GEORGIAN SMALL LETTER UN
2D14	Ⴔ	GEORGIAN SMALL LETTER PHAR
2D15	Ⴕ	GEORGIAN SMALL LETTER KHAR
2D16	Ⴖ	GEORGIAN SMALL LETTER GHAN
2D17	Ⴗ	GEORGIAN SMALL LETTER QAR
2D18	Ⴘ	GEORGIAN SMALL LETTER SHIN
2D19	Ⴙ	GEORGIAN SMALL LETTER CHIN
2D1A	Ⴚ	GEORGIAN SMALL LETTER CAN
2D1B	Ⴛ	GEORGIAN SMALL LETTER JIL
2D1C	Ⴜ	GEORGIAN SMALL LETTER CIL
2D1D	Ⴝ	GEORGIAN SMALL LETTER CHAR
2D1E	Ⴞ	GEORGIAN SMALL LETTER XAN
2D1F	Ⴟ	GEORGIAN SMALL LETTER JHAN
2D20	Ⴀ	GEORGIAN SMALL LETTER HAE
2D21	Ⴁ	GEORGIAN SMALL LETTER HE
2D22	Ⴂ	GEORGIAN SMALL LETTER HIE
2D23	Ⴃ	GEORGIAN SMALL LETTER WE
2D24	Ⴄ	GEORGIAN SMALL LETTER HAR
2D25	Ⴅ	GEORGIAN SMALL LETTER HOE

2C00

































Glagolitic

2C5F

	2C0	2C1	2C2	2C3	2C4	2C5
0	ᐁ	ᐂ	ᐃ	ᐄ	ᐅ	ᐆ
1	ᐇ	ᐈ	ᐉ	ᐊ	ᐋ	ᐌ
2	ᐏ	ᐐ	ᐑ	ᐒ	ᐓ	ᐔ
3	ᐗ	ᐘ	ᐙ	ᐚ	ᐛ	ᐜ
4	ᐞ	ᐟ	ᐠ	ᐡ	ᐢ	ᐣ
5	ᐥ	ᐦ	ᐧ	ᐨ	ᐩ	ᐪ
6	ᐬ	ᐭ	ᐮ	ᐯ	ᐰ	ᐱ
7	ᐳ	ᐴ	ᐵ	ᐶ	ᐷ	ᐸ
8	ᐹ	ᐺ	ᐻ	ᐼ	ᐽ	ᐾ
9	ᐿ	ᑀ	ᑁ	ᑂ	ᑃ	ᑄ
A	ᑆ	ᑇ	ᑈ	ᑉ	ᑊ	ᑋ
B	ᑎ	ᑏ	ᑐ	ᑑ	ᑒ	ᑓ
C	ᑕ	ᑖ	ᑗ	ᑘ	ᑙ	ᑚ
D	ᑝ	ᑞ	ᑟ	ᑐ	ᑠ	ᑡ
E	ᑣ	ᑤ	ᑥ	ᑦ	ᑧ	ᑨ
F	ᑩ	ᑪ	ᑫ	ᑬ	ᑭ	ᑮ

































































Ancient languages:

“Extended” UNICODE blocks (examples)

10330	Gothic		1034F
	1033	1034	
0			
1			
2			
3			
4			
5			
6			
7			
8			
9			
A			
B			
C			
D			
E			
F			

Letters

10330	A	GOthic LETTER AHSA
10331	B	GOthic LETTER BAIRKAN
10332	C	GOthic LETTER GIBA
10333	D	GOthic LETTER DAGS
10334	E	GOthic LETTER AIHVUS
10335	F	GOthic LETTER QARÞHRA
10336	G	GOthic LETTER IJJA
10337	H	GOthic LETTER HAGL
10338	I	GOthic LETTER THRUÞH
10339	J	GOthic LETTER IES
1033A	K	GOthic LETTER KUSMA
1033B	L	GOthic LETTER LAGUS
1033C	M	GOthic LETTER NAINNA
1033D	N	GOthic LETTER NAUTHS
1033E	O	GOthic LETTER JER
1033F	P	GOthic LETTER URUS
10340	Q	GOthic LETTER PAIRÞHRA
10341	R	GOthic LETTER NINEITY
10342	S	GOthic LETTER RAIDA
10343	T	GOthic LETTER SAUL
10344	U	GOthic LETTER TEIWS
10345	V	GOthic LETTER WINJA
10346	W	GOthic LETTER FAIHU
10347	X	GOthic LETTER SIGWS
10348	Y	GOthic LETTER HWAIR
10349	Z	GOthic LETTER OTHAL
1034A	T	GOthic LETTER NINE HUNDRED

103A0	Old Persian				103DF
	103A	103B	103C	103D	
0					
1					
2					
3					
4					
5					
6					
7					
8					
9					
A					
B					
C					
D					
E					
F					

Independent vowels

103A0	𐎠	OLD PERSIAN SIGN A
103A1	𐎡	OLD PERSIAN SIGN I
103A2	𐎢	OLD PERSIAN SIGN U

Consonants

103A3	𐎣	OLD PERSIAN SIGN KA
103A4	𐎤	OLD PERSIAN SIGN KU
103A5	𐎥	OLD PERSIAN SIGN GA
103A6	𐎦	OLD PERSIAN SIGN GU
103A7	𐎧	OLD PERSIAN SIGN XA
103A8	𐎨	OLD PERSIAN SIGN CA
103A9	𐎩	OLD PERSIAN SIGN JA
103AA	𐎪	OLD PERSIAN SIGN CI
103AB	𐎫	OLD PERSIAN SIGN TA
103AC	𐎬	OLD PERSIAN SIGN TU
103AD	𐎭	OLD PERSIAN SIGN DA
103AE	𐎮	OLD PERSIAN SIGN DI
103AF	𐎯	OLD PERSIAN SIGN DU
103B0	𐎰	OLD PERSIAN SIGN THA
103B1	𐎱	OLD PERSIAN SIGN PA
103B2	𐎲	OLD PERSIAN SIGN BA
103B3	𐎳	OLD PERSIAN SIGN FA
103B4	𐎴	OLD PERSIAN SIGN NA
103B5	𐎵	OLD PERSIAN SIGN NU
103B6	𐎶	OLD PERSIAN SIGN MA
103B7	𐎷	OLD PERSIAN SIGN ME
103B8	𐎸	OLD PERSIAN SIGN MU
103B9	𐎹	OLD PERSIAN SIGN YA
103BA	𐎺	OLD PERSIAN SIGN VA
103BB	𐎻	OLD PERSIAN SIGN VI
103BC	𐎼	OLD PERSIAN SIGN RA
103BD	𐎽	OLD PERSIAN SIGN RU
103BE	𐎾	OLD PERSIAN SIGN LA
103BF	𐎿	OLD PERSIAN SIGN SA
103C0	𐏀	OLD PERSIAN SIGN ZA
103C1	𐏁	OLD PERSIAN SIGN SHA
103C2	𐏂	OLD PERSIAN SIGN SSA
103C3	𐏃	OLD PERSIAN SIGN HA

Various signs

103C8	𐏈	OLD PERSIAN SIGN AURAMAZDAA
103C9	𐏉	OLD PERSIAN SIGN AURAMAZDAA-2
103CA	𐏊	OLD PERSIAN SIGN AURAMAZDAAHA
103CB	𐏋	OLD PERSIAN SIGN XSHAA YATHYA
103CC	𐏌	OLD PERSIAN SIGN DAHYAAUSH
103CD	𐏍	OLD PERSIAN SIGN DAHYAAUSH-2
103CE	𐏎	OLD PERSIAN SIGN BAGA
103CF	𐏏	OLD PERSIAN SIGN BUUMISH

Punctuation

103D0	𐏐	OLD PERSIAN WORD DIVIDER
-------	---	--------------------------

Numbers

103D1	𐏑	OLD PERSIAN NUMBER ONE
103D2	𐏒	OLD PERSIAN NUMBER TWO
103D3	𐏓	OLD PERSIAN NUMBER TEN
103D4	𐏔	OLD PERSIAN NUMBER TWENTY
103D5	𐏕	OLD PERSIAN NUMBER HUNDRED

Yet to be implemented:

- Avestan
- Pahlavi
- Manichean
- Sogdian
- etc.

Co-existence of encoding standards:

Bad case scenario: Non-standard fonts

- File to be exchanged (Georgian word list, MS-Word 6)

0000200T 2 აგება [ააგებს, ააგო, აუგებიეს]
0000240T 2 აგვა [აპგვის, აგავა, აუგავს]
0000250T 2 აგზება [ააგზებს, ააგზო, აუგზებიეს]
0000270T 2 აგზნება [ააგზნებს, ააგზნო, აუგზნებიეს]
0000290I 1 აგრეკნება [აგრეკინდების, აგრეკინდა, აგრეკნებულ არს]
0000420* * ადგილ-გება [ადგილს-აგებს, ადგილ-აგო, ადგილ-უგებიეს]
0000510T 2 ადგილის-ჰყრობა [ადგილს-იჰყრობს, ადგილს-იჰყრა, ადგილ-უჰყრიეს]
0000520T 3 ადგილის-ცემა [ადგილს-სცემს, ადგილს-სცა, ადგილ-უცემაეს]
0000570T 2 ადგინება [ადგინებს, აადგინა, აუდგინებიეს]
0000580T 2 ადგმა [აადგამს, აადგა, აუდგამს]
0000590I 1 ადგომა [ადგების, ადგა, ამდგარა]
0000650T 2 ადვილ-ყოფა [ადვილ-ჰყოფს, ადვილ-ყო, ადვილ-უყოფიეს]
0000860I 1 ავაზაკობა [ავაზაკობს, იავაზაკა, უავაზაკებიეს]
0001120T 2 აზელა [აზელს, აზილა, აუზელიეს]
0001190I 1 აზრახება [აზრახების, აზრახდა, აზრახებულარს]
0001200T 2 აზლუდვა [აპზლუდავს, აზლუდა, აუზლუდავს]
0001430T 2 ათრთოდება [ათრთოდებს, აათრთოდა, აუთრთოდებიეს]
0001490T 2 ათუადვა [ასთუადავს, ათუადა, აუთუადავს]
0001570I 1 ათხელიება [ათხელიების, ათხელიდა, ათხელიებულ არს]

Co-existence of encoding standards:

Bad case scenario: Non-standard fonts

- File opened with MS-Word XP (assumption: Japanese)

```
0000200T 2 タットチタ [タタットチモ, タタツマ, タヨットチノトモ]
0000240T 2 タツナタ [タ萃ナノモ, タツタナタ, タヨツタナモ]
0000250T 2 タツニトチタ [タタツニトチモ, タタツニマ, タヨツニトチノトモ]
0000270T 2 タツニヘトチタ [タタツニヘトチモ, タタツニヘマ, タヨツニヘトチノトモ]
0000290I 1 タツメツユヘトチタ [タツメツユノヘトチノモ, タツメツユノヘタ, タツメツユヘトチヨヒ タメモ]
0000420* * タテツノテータットチタ [タテツノテモータットチモ, タテツノテータツマ, タテツノテータツノトモ]
0000510T 2 タテツノヒノモミレメマチタ [タテツノヒノモミレメマチモ, タテツノヒノモミレメタ, タテツノヒノモミレメノトモ]
0000520T 3 タテツノヒノモントフタ [タテツノヒノモントフモ, タテツノヒノモントタ, タテツノヒノモントフノトモ]
0000570T 2 タテツノヘトチタ [タテツノヘトチモ, タタテツノヘタ, タヨテツノヘトチノトモ]
0000580T 2 タテツタ [タタテツタモ, タタテツタ, タヨテツタモ]
0000590I 1 タテツマフタ [タテツタチノモ, タテツタ, タフテツタメタ]
0000650T 2 タテナノヒレマラタ [タテナノヒレマラモ, タテナノヒレマ, タテナノヒレマラノトモ]
0000860I 1 タナタニタハマチタ [タナタニタハマチモ, ノタナタニタハタ, ヨタナタニタハトチノトモ]
0001120T 2 タニトヒタ [タニトヒモ, タニノヒタ, タヨニトヒノトモ]
0001190I 1 タニメタ昧チタ [タニメタ昧トチノモ, タニメタ昧タ, タニメタ昧チヨヒタメモ]
0001200T 2 タニルヨテナタ [タ蒂ルヨテナモ, タニルヨテナ, タヨニルヨテナモ]
0001430T 2 タネメネマヒトチタ [タタネメネマヒトチモ, タタネメネマヒタ, タヨネメネマヒトチノトモ]
0001490T 2 タネヨタヒナタ [タネヨタヒナモ, タネヨタヒタ, タヨネヨタヒナモ]
0001570I 1 タネ昧ヒトチタ [タネ昧ヒトチノモ, タネ昧ヒタ, タネ昧ヒトチヨヒ タメモ]
```

Co-existence of encoding standards:

Bad case scenario: Non-standard fonts

- File opened with Open Office 1.1 (assumption: Roman)

```
0000200T 2 ÀÃÄÅ [ÀÃÄÅÁÓ, ÀÃÄÏ, ÀÖÄÄÄÉÄÓ]
0000240T 2 ÀÃÄÄ [ÀäÄÄÉÓ, ÀÃÄÄÄ, ÀÖÄÄÄÓ]
0000250T 2 ÀÄÆÄÄÄ [ÀÄÄÆÄÄÁÓ, ÀÄÄÆÏ, ÀÖÄÆÄÄÉÄÓ]
0000270T 2 ÀÄÆÍÄÄÄ [ÀÄÄÆÍÄÄÁÓ, ÀÄÄÆÏÏ, ÀÖÄÆÍÄÄÉÄÓ]
0000290I 1 ÄÄÖÄÖÍÄÄÄ [ÄÄÖÄÖÉÍÄÄÄÉÓ, ÄÄÖÄÖÉÍÄÄ, ÄÄÖÄÖÍÄÄÄÖÄÖ]
0000420* * ÄÄÄÉÄ-ÄÄÄÄ [ÄÄÄÉÄÖ-ÄÄÄÄÁÓ, ÄÄÄÉÄ-ÄÄÏ, ÄÄÄÉÄ-ÖÄÉÄÓ]
0000510T 2 ÄÄÄÉÉÉÖ-ÐÚÒÍÄÄ [ÄÄÄÉÉÖ-ÉÐÚÒÍÄÄÁÓ, ÄÄÄÉÉÖ-ÉÐÚÒÄ, ÄÄÄÉÉ-ÖÐÚÒÉÄÖ]
0000520T 3 ÄÄÄÉÉÉÖ-ÝÄÌÄ [ÄÄÄÉÉÖ-ÓÝÄÌÓ, ÄÄÄÉÉÖ-ÓÝÄ, ÄÄÄÉÉ-ÖÝÄÉÄÓ]
0000570T 2 ÄÄÄÉÍÄÄÄ [ÄÄÄÉÍÄÄÁÓ, ÄÄÄÄÉÍÄ, ÀÖÄÄÉÍÄÄÉÄÓ]
0000580T 2 ÄÄÄÌÄ [ÄÄÄÄÄÌÓ, ÄÄÄÄÄ, ÀÖÄÄÄÌÓ]
0000590I 1 ÄÄÄÏÄ [ÄÄÄÄÄÉÓ, ÄÄÄÄÄ, ÄÌÄÄÄÒÄ]
0000650T 2 ÄÄÄÉÉ-ÚÏ×Ä [ÄÄÄÉÉ-äÚÏ×Ó, ÄÄÄÉÉ-ÚÏ, ÄÄÄÉÉ-ÖÚÏ×ÉÄÓ]
0000860I 1 ÄÄÄÆÄÉÍÄÄ [ÄÄÄÆÄÉÍÄÄÁÓ, ÉÄÄÄÆÄÉÄ, ÖÄÄÄÆÄÉÄÄÉÄÁÓ]
0001120T 2 ÄÆÄÉÄ [ÄÆÄÉÖ, ÄÆÉÉÄ, ÀÖÆÄÉÉÄÁÓ]
0001190I 1 ÄÆÖÄäÄÄÄ [ÄÆÖÄäÄÄÄÉÓ, ÄÆÖÄäÄÄÄ, ÄÆÖÄäÄÄÄÖÄÖÄÖ]
0001200T 2 ÄÆÜÖÄÄÄ [ÄäÆÜÖÖÄÄÄÁÓ, ÄÆÜÖÖÄÄ, ÀÖÆÜÖÖÄÄÄÁÓ]
0001430T 2 ÄÈÖÈÏÄÄÄ [ÄÄÈÖÈÏÄÄÄÁÓ, ÄÄÈÖÈÏÄÄ, ÀÖÈÖÈÏÄÄÄÉÄÁÓ]
0001490T 2 ÄÈÖÄÉÄÄ [ÄÖÈÖÄÉÄÄÁÓ, ÄÈÖÄÉÄ, ÀÖÈÖÄÉÄÄÁÓ]
0001570I 1 ÄÈäÄÉÄÄÄ [ÄÈäÄÉÄÄÄÉÓ, ÄÈäÄÉÄÄÄ, ÄÈäÄÉÄÄÄÖÄÖÄÖ]
```

Co-existence of encoding standards:

Bad case scenario: Non-standard fonts

- Same after applying correct Georgian 8-bit font

```
0000200T 2 აგება [ააგებს, ააგო, აუგებიეს]
0000240T 2 აგვა [აპგვის, აგავა, აუგავს]
0000250T 2 აგზება [ააგზებს, ააგზო, აუგზებიეს]
0000270T 2 აგზნება [ააგზნებს, ააგზნო, აუგზნებიეს]
0000290I 1 აგრგუნება [აგრგუნდების, აგრგუნდა, აგრგუნებულ არს]
0000420* * ადგიდ-გება [ადგიდს-აგებს, ადგიდ-აგო, ადგიდ-უგებს]
0000510T 2 ადგილის-ჰყრობა [ადგიდს-იჰყრობს, ადგიდს-იჰყრა, ადგიდ-უჰყრიეს]
0000520T 3 ადგილის-ცემა [ადგიდს-სცემს, ადგიდს-სცა, ადგიდ-უცემიეს]
0000570T 2 ადგინება [ადგინებს, აადგინა, აუდგინებიეს]
0000580T 2 ადგმა [აადგამს, აადგა, აუდგამს]
0000590I 1 ადგომა [ადგების, ადგა, ამდგარა]
0000650T 2 ადვილ-ყოფა [ადვილ-ჰყოფს, ადვილ-ყო, ადვილ-უყოფიეს]
0000860I 1 ავაზაკობა [ავაზაკობს, იავაზაკა, უავაზაკებიეს]
0001120T 2 აზელა [აზელს, აზილა, აუზელიეს]
0001190I 1 აზრახება [აზრახდების, აზრახდა, აზრახებულარს]
0001200T 2 აზღუდვა [აპზღუდავს, აზღუდა, აუზღუდავს]
0001430T 2 ათრთოღება [აათრთოღებს, აათრთოღა, აუთრთოღებიეს]
0001490T 2 ათუაღვა [ასთუაღავს, ათუაღა, აუთუაღავს]
0001570I 1 ათხეღება [ათხეღდების, ათხეღდა, ათხეღებულ არს]
```

Co-existence of encoding standards:

Bad case scenario: Non-standard fonts

- Same after applying equivalent transcriptional 8-bit font

```
0000200T 2 ageba [aagebs, aago, augebies]
0000240T 2 agva [ahgvis, agava, augavs]
0000250T 2 agzeba [aagzebs, aagzo, augzebies]
0000270T 2 agzneba [aagznebs, aagzno, augznebies]
0000290I 1 agrgwineba [agrgwiindebis, agrgwiinda, agrgwinebul ars]
0000420* * adgid-geba [adgids-agebs, adgid-ago, adgid-ugies]
0000510T 2 adgilis-pqroba [adgils-ipqrobs, adgils-ipqra, adgil-upqries]
0000520T 3 adgilis-cema [adgils-scems, adgils-sca, adgil-ucemies]
0000570T 2 adgineba [adginebs, aadgina, audginebies]
0000580T 2 adgma [aadgams, aadga, audgams]
0000590I 1 adgoma [adgebis, adga, amdgara]
0000650T 2 advil-qopa [advil-hqops, advil-qo, advil-ugopies]
0000860I 1 avazaḱoba [avazaḱobs, iavazaḱa, uavazaḱebies]
0001120T 2 azela [azels, azila, auzelies]
0001190I 1 azraxeba [azraxdebis, azraxda, azraxebulars]
0001200T 2 azyudva [ahzyudavs, azyuda, auzyudavs]
0001430T 2 atrtoleba [aatrtolebs, atrtola, autrtolebies]
0001490T 2 atualva [astualavs, atuala, autualavs]
0001570I 1 atxeleba [atxeldebis, atxelda, atxelebul ars]
```

Co-existence of encoding standards: Worst case scenario: Mixture of 8- and 16-bit

- MS-Word XP after applying correct Georgian 8-bit font

0000200T 2 ტატიტა [ტატატიტო, ტატა, ტაოტიტოტო]
0000240T 2 ტაჟა [ტაჟაჟა, ტაჟაჟა, ტაოჟაჟა]
0000250T 2 ტაჟიტი [ტატაჟიტი, ტატაჟი, ტაოჟიტიტო]
0000270T 2 ტაჟიტი [ტატაჟიტი, ტატაჟი, ტაოჟიტიტო]
0000290I 1 ტაჟიტი [ტატაჟიტი, ტატაჟი, ტაოჟიტიტო] ტაჟი
0000420* * ტაჟიტი [ტატაჟიტი, ტატაჟი, ტაოჟიტიტო]
0000510T 2 ტაჟიტი-მილამა [ტატაჟიტი-მილამა, ტატაჟიტი-მილამა, ტატაჟიტი-მილამა]
0000520T 3 ტაჟიტი-ნოტი [ტატაჟიტი-ნოტი, ტატაჟიტი-ნოტი, ტატაჟიტი-ნოტი]
0000570T 2 ტაჟიტი [ტატაჟიტი, ტატაჟიტი, ტაოჟიტიტო]
0000580T 2 ტაჟიტი [ტატაჟიტი, ტატაჟიტი, ტაოჟიტიტო]
0000590I 1 ტაჟიტი [ტატაჟიტი, ტატაჟი, ტაოჟიტიტო]
0000650T 2 ტაჟიტი-ლამა [ტატაჟიტი-ლამა, ტატაჟიტი-ლამა, ტატაჟიტი-ლამა]
0000860I 1 ტაჟიტი-ლამა [ტატაჟიტი-ლამა, ტატაჟიტი-ლამა, ტაოჟიტიტო]
0001120T 2 ტაჟიტი [ტატაჟიტი, ტატაჟიტი, ტაოჟიტიტო]
0001190I 1 ტაჟიტი [ტატაჟიტი, ტატაჟიტი, ტაოჟიტიტო]
0001200T 2 ტაჟიტი [ტატაჟიტი, ტატაჟიტი, ტაოჟიტიტო]
0001430T 2 ტაჟიტი-მილამა [ტატაჟიტი-მილამა, ტატაჟიტი-მილამა, ტაოჟიტიტო]
0001490T 2 ტაჟიტი-მილამა [ტატაჟიტი-მილამა, ტატაჟიტი-მილამა, ტაოჟიტიტო]
0001570I 1 ტაჟიტი-მილამა [ტატაჟიტი-მილამა, ტატაჟიტი-მილამა, ტაოჟიტიტო]

Co-existence of encoding standards:

Worst case scenario: Mixture of 8- and 16-bit

- The MS-Word strategy:
 - Checks whether the document is Unicode-encoded
 - If not, checks whether the character distribution might meet the “typical” distribution of one of the known codepages
 - If yes, assumes that codepage to be represented
 - Converts the 8-bit characters of the codepage into the equivalent characters of Unicode
 - Stores the Unicode characters in memory
 - Applying 8-bit fonts will be no remedy as they do not meet the Unicode encoding assumed and applied

Example: Comparison of Shoebox and Toolbox (Unicode) encoding:

\id	10054
\lm	
\fg	ხვაშადი
\sv	ხეჲშინ, ჰომ
\ru	тайна, секрет;
\fg	შეგინდე და დამიძალე
\sv	შემინდ ი ხვაშინ მაყრ.
\sc	BN_GSR

\id 10054
\lm
\fg xvaêiadi
\sv xvóêin, hom
\ru nfqyf> ctrhtn&
\fg êeginde da damimale
\sv êemind i xvaêin maqr.
\sc BN_GSR

Comparison of Shoebox and Toolbox (Unicode) encoding:

\c	71.5
\tl	Шәара шәца, арахь ааишья шәымам, шәарей сарей уаха хабла еихәапшраны хаҕам!
\ts	š°ara š°ca, arax' aaiša š°əmam, š°arei sarei uaha habla eix°apš´ranə haḡam!
\fg	თქვენ წადით, აქეთ ვერ მოაღწევთ, ჩვენ ერთმანეთს ველარასოდეს ვნახავთო.
\fe	
\c	71.6
\tl	Шәара зака шәааскьо акара, сара убриакара нак сагоит.
\ts	š°ara zaḡa š°aaskʼo, sara ubriaḡara naḡ sagoiṭ.
\fg	რამდენს მომახლოვდებით, იმდენს დაგვაშორებენ.
\fe	

\c 71.5

\tl Шәара шәца, арахь ааишья шәымам, шәарей сарей уаха хабла еихәапшраны хаҕам!

\ts š°ara š°ca, arax' aaiša š°əmam, š°arei sarei uaha habla eix°apš´ranə haḡam!

\fg თქვენ წადით, აქეთ ვერ მოაღწევთ, ჩვენ ერთმანეთს ველარასოდეს ვნახავთო.

\fe

\c 71.6

\tl Шәара зака шәааскьо акара, сара убриакара нак сагоит.

\ts š°ara zaḡa š°aaskʼo, sara ubriaḡara naḡ sagoiṭ.

\fg რამდენს მომახლოვდებით, იმდენს დაგვაშორებენ.

\fe

How to avoid the worst case scenario

- Requirements for text data exchange:
 - If 8-bit encoding is required, mixing up several fonts with a different encoding in the document should be avoided
 - Keep track of font-and-encoding
 - Inform users about all this and provide fonts (if legal)
 - TRY TO USE UNICODE ENCODING WHEREVER POSSIBLE

Encoding of Characters:

How to avoid the worst case scenario

- Recommended strategy for data storage (archiving):
 - Convert all 8-bit documents into 16-bit Unicode documents
 - Avoid storage of proprietary formats (e.g., MS-Word)

Problems of Unicode encoding

- Byte storage:
 - UTF-16 vs. UTF-8 vs. UTF-7 (vs. U+FFFF ...)
- Problems of “Non-Uniqueness”
 - Problems of the “Private Use Area”
 - Problems of “Normalization”
- Problems of “Too-Uniqueness”
 - Problems of bidirectionality

Byte storage:

Desired output

English

Südöstlich

русский

ქართული

Byte storage: UTF-8

English

SÃ¼dÃ¶stlich

Ñ€ÑƒÑ¼Ñ½Ð°Ð,Ð¹

áf¥áfłáf áf—áf£áfšáf~

Byte storage: UTF-16

E n g l i s h

S ü d ö s t l i c h

@ C A A : 8 9

å ð à × ã ú ø

UTF-8 vs. UTF-16 encoding:

45 6E 67 6C 69 73 68 OD 0A

53 C3 BC 64 C3 B6 73 74 6C 69 63 68 0D 0A

D1 80 D1 83 D1 81 D1 81 D0 BA D0 B8 D0 B9 0D 0A

E1 83 A5 E1 83 90 E1 83 A0 E1 83 97 E1 83 A3 E1

83 9A E1 83 98 0D 0A

EnglishHJ

Stilistisch

[illegible]

BÂN BĂE BĂA BĂU BĂÚ

âÜâÿMJ

```
45 00 6E 00 67 00 6C 00 69 00 73 00 68 00 0D 00
0A
```

00 53 00 FC 00 64 00 F6 00 73 00 74 00 6C 00 69

00 63 00 68 00 0D 00 0A

```
00 40 04 43 04 41 04 41 04 3A 04 38 04 39 04 0D
```

00 0A

```
00 E5 10 D0 10 E0 10 D7 10 E3 10 DA 10 D8 10 0D
```

EnergiescheW
J

0\$0^0@d÷@s@t@l@i

@ceh@M@J

@@DCDADAD:D8D9DM

01

α σ ρ \parallel ρ α ρ \parallel ρ π ρ \perp ρ π

Characteristics

- UTF-16
 - constant byte rate for Latin and other scripts
 - ANSI elements readable as such (depending on viewer capabilities)
- UTF-8
 - low byte rate with Latin-based text
 - ASCII elements readable as such
 - high byte rate with “exotic” scripts
 - supported by many viewers, browsers...

Recommendation:

- Prefer UTF-8 for the storage of textual data that are meant for instant retrieval
- Prefer UTF-16 for the long-time storage of data

Non-Uniqueness of Unicode:

Font mapping anew: the PUA

<	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F	>
E10	Ġ	G̈	G̉	G̊	G̋	Ǧ	G̍	G̎	G̏	G̐	G̑	G̒	G̓	G̔	G̕	G̖	G̗
E11	G̘	G̙	G̚		Ḣ	Ḧ		H̉	H̊	H̋	Ȟ	H̍	H̎	H̏	H̐	H̑	H̒
E12	H̓	H̔	H̕	H̖	H̗	H̘	H̙	H̚		İ̇	İ̈	İ̉	İ̊	İ̋	İ̌	İ̍	İ̎
E13	İ̏	İ̐	İ̑	İ̒	İ̓	İ̔	İ̕	İ̖	İ̗	İ̘	İ̙	İ̚	İ̛	İ̜	İ̝	İ̞	İ̟
E14	İ̠	İ̡	İ̢	Ị̇	İ̤	İ̥	İ̦	İ̧	Į̇	İ̩	İ̪	İ̫	İ̬	İ̭	İ̮	İ̯	
E15	Ḭ̇	İ̱	İ̲	İ̳	İ̴	İ̵	İ̶	İ̷	İ̸	İ̹	İ̺	İ̻	İ̼	İ̽	İ̾	İ̿	
E16			Ĵ	Ĵ̇	Ĵ̈	Ĵ̉	Ĵ̊	Ĵ̋	Ĵ̌	Ĵ̍	Ĵ̎	Ĵ̏	Ĵ̐	Ĵ̑	Ĵ̒	Ĵ̓	Ĵ̔
E17	Ĵ̕	Ĵ̖	Ĵ̗	Ĵ̘	Ĵ̙	Ĵ̚	Ĵ̛	Ĵ̜	Ĵ̝	Ĵ̞	Ĵ̟	Ĵ̠	Ĵ̡	Ĵ̢	Ĵ̣	Ĵ̤	Ĵ̥
E18	Ĵ̦	Ĵ̧	Ĵ̨	Ĵ̩	Ĵ̪	Ĵ̫	Ĵ̬	Ĵ̭	Ĵ̮	Ĵ̯	Ĵ̰	Ĵ̱	Ĵ̲	Ĵ̳	Ĵ̴	Ĵ̵	Ĵ̶
E19	Ĵ̷	Ĵ̸	Ĵ̹	Ĵ̺	Ĵ̻	Ĵ̼	Ĵ̽	Ĵ̾	Ĵ̿	Ĵ̠	Ĵ̡	Ĵ̢	Ĵ̣	Ĵ̤	Ĵ̥	Ĵ̦	Ĵ̧
E1A	Ĵ̨	Ĵ̩	Ĵ̪	Ĵ̫	Ĵ̬	Ĵ̭	Ĵ̮	Ĵ̯	Ĵ̰	Ĵ̱	Ĵ̲	Ĵ̳	Ĵ̴	Ĵ̵	Ĵ̶	Ĵ̷	Ĵ̸
E1B			Ṁ	Ṁ̇	Ṁ̈	Ṁ̉	Ṁ̊	Ṁ̋	Ṁ̌	Ṁ̍	Ṁ̎	Ṁ̏	Ṁ̐	Ṁ̑	Ṁ̒	Ṁ̓	Ṁ̔
E1C	Ṁ̕	Ṁ̖	Ṁ̗	Ṁ̘	Ṁ̙	Ṁ̚	Ṁ̛	Ṁ̜	Ṁ̝	Ṁ̞	Ṁ̟	Ṁ̠	Ṁ̡	Ṁ̢	Ṃ̇	Ṁ̤	Ṁ̥
E1D	Ṁ̦	Ṁ̧	Ṁ̨	Ṁ̩	Ṁ̪	Ṁ̫	Ṁ̬	Ṁ̭	Ṁ̮	Ṁ̯	Ṁ̰	Ṁ̱	Ṁ̲	Ṁ̳	Ṁ̴	Ṁ̵	Ṁ̶
E1E	Ṁ̷	Ṁ̸	Ṁ̹	Ṁ̺	Ṁ̻	Ṁ̼	Ṁ̽	Ṁ̾	Ṁ̿	Ṁ̠	Ṁ̡	Ṁ̢	Ṃ̇	Ṁ̤	Ṁ̥	Ṁ̦	Ṁ̧
E1F	Ṁ̨	Ṁ̩	Ṁ̪	Ṁ̫	Ṁ̬	Ṁ̭	Ṁ̮	Ṁ̯	Ṁ̰	Ṁ̱	Ṁ̲	Ṁ̳	Ṁ̴	Ṁ̵	Ṁ̶	Ṁ̷	Ṁ̸

≤	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F	≥
E10	𐀀	𐀁	𐀂	𐀃	𐀄	𐀅	𐀆	𐀇	𐀈	𐀉	𐀊	𐀋	𐀌	𐀍	𐀎	𐀏	𐀐
E11	𐀑	𐀒	𐀓	𐀔	𐀕	𐀖	𐀗	𐀘	𐀙	𐀚	𐀛	𐀜	𐀝	𐀞	𐀟	𐀠	𐀡
E12	𐀢	𐀣	𐀤	𐀥	𐀦	𐀧	𐀨	𐀩	𐀪	𐀫	𐀬	𐀭	𐀮	𐀯	𐀰	𐀱	𐀲
E13	𐀳	𐀴	𐀵	𐀶	𐀷	𐀸	𐀹	𐀺	𐀻	𐀼	𐀽	𐀾	𐀿	𐁀	𐁁	𐁂	𐁃
E14	𐁄	𐁅	𐁆	𐁇	𐁈	𐁉	𐁊	𐁋	𐁌	𐁍	𐁎	𐁏	𐁐	𐁑	𐁒	𐁓	𐁔
E15	𐁕	𐁖	𐁗	𐁘	𐁙	𐁚	𐁛	𐁜	𐁝	𐁞	𐁟	𐁠	𐁡	𐁢	𐁣	𐁤	𐁥
E16	𐁦	𐁧	𐁨	𐁩	𐁪	𐁫	𐁬	𐁭	𐁮	𐁯	𐁰	𐁱	𐁲	𐁳	𐁴	𐁵	𐁶
E17	𐁷	𐁸	𐁹	𐁺	𐁻	𐁼	𐁽	𐁾	𐁿	𐂀	𐂁	𐂂	𐂃	𐂄	𐂅	𐂆	𐂇
E18	𐂈	𐂉	𐂊	𐂋	𐂌	𐂍	𐂎	𐂏	𐂐	𐂑	𐂒	𐂓	𐂔	𐂕	𐂖	𐂗	𐂘
E19	𐂙	𐂚	𐂛	𐂜	𐂝	𐂞	𐂟	𐂠	𐂡	𐂢	𐂣	𐂤	𐂥	𐂦	𐂧	𐂨	𐂩
E1A	𐂪	𐂫	𐂬	𐂭	𐂮	𐂯	𐂰	𐂱	𐂲	𐂳	𐂴	𐂵	𐂶	𐂷	𐂸	𐂹	𐂺
E1B	𐂻	𐂼	𐂽	𐂾	𐂿	𐃀	𐃁	𐃂	𐃃	𐃄	𐃅	𐃆	𐃇	𐃈	𐃉	𐃊	𐃋
E1C	𐃌	𐃍	𐃎	𐃏	𐃐	𐃑	𐃒	𐃓	𐃔	𐃕	𐃖	𐃗	𐃘	𐃙	𐃚	𐃛	𐃜
E1D	𐃝	𐃞	𐃟	𐃠	𐃡	𐃢	𐃣	𐃤	𐃥	𐃦	𐃧	𐃨	𐃩	𐃪	𐃫	𐃬	𐃭
E1E	𐃮	𐃯	𐃰	𐃱	𐃲	𐃳	𐃴	𐃵	𐃶	𐃷	𐃸	𐃹	𐃺	𐃻	𐃼	𐃽	𐃾
E1F	𐃿	𐄀	𐄁	𐄂	𐄃	𐄄	𐄅	𐄆	𐄇	𐄈	𐄉	𐄊	𐄋	𐄌	𐄍	𐄎	𐄏

Non-Uniqueness in Unicode

Multiple code points: Arabic numerals

Arabic-Indic digits

These digits are used with Arabic proper; for languages of Iran, Pakistan, and India, see the Eastern Arabic-Indic digits at 06F0..06F9.

0660	٠	ARABIC-INDIC DIGIT ZERO
0661	١	ARABIC-INDIC DIGIT ONE
0662	٢	ARABIC-INDIC DIGIT TWO
0663	٣	ARABIC-INDIC DIGIT THREE
0664	٤	ARABIC-INDIC DIGIT FOUR
0665	٥	ARABIC-INDIC DIGIT FIVE
0666	٦	ARABIC-INDIC DIGIT SIX
0667	٧	ARABIC-INDIC DIGIT SEVEN
0668	٨	ARABIC-INDIC DIGIT EIGHT
0669	٩	ARABIC-INDIC DIGIT NINE

Eastern Arabic-Indic digits

These digits are used with Arabic-script languages of Iran, Pakistan, and India (Persian, Sindhi, Urdu, etc.). For details of variations in preferred glyphs, see the block description for the Arabic script.

06F0	٠	EXTENDED ARABIC-INDIC DIGIT ZERO
06F1	١	EXTENDED ARABIC-INDIC DIGIT ONE
06F2	٢	EXTENDED ARABIC-INDIC DIGIT TWO
06F3	٣	EXTENDED ARABIC-INDIC DIGIT THREE
06F4	٤	EXTENDED ARABIC-INDIC DIGIT FOUR <ul style="list-style-type: none">• Persian has a different glyph than Sindhi and Urdu
06F5	٥	EXTENDED ARABIC-INDIC DIGIT FIVE <ul style="list-style-type: none">• Persian, Sindhi, and Urdu share glyph different from Arabic
06F6	٦	EXTENDED ARABIC-INDIC DIGIT SIX <ul style="list-style-type: none">• Persian, Sindhi, and Urdu have glyphs different from Arabic
06F7	٧	EXTENDED ARABIC-INDIC DIGIT SEVEN <ul style="list-style-type: none">• Urdu and Sindhi have glyphs different from Arabic
06F8	٨	EXTENDED ARABIC-INDIC DIGIT EIGHT
06F9	٩	EXTENDED ARABIC-INDIC DIGIT NINE

Non-Uniqueness in Unicode

Multiple code points: Arabic “presentation forms”

0627	ا	ARABIC LETTER ALEF
0628	ب	ARABIC LETTER BEH
0629	ة	ARABIC LETTER TEH MARBUTA
062A	ت	ARABIC LETTER TEH
062B	ث	ARABIC LETTER THEH
062C	ج	ARABIC LETTER JEEM
062D	ح	ARABIC LETTER HAH
062E	خ	ARABIC LETTER KHAH
062F	د	ARABIC LETTER DAL
0630	ذ	ARABIC LETTER THAL
0631	ر	ARABIC LETTER REH
0632	ز	ARABIC LETTER ZAIN
0633	س	ARABIC LETTER SEEN
0634	ش	ARABIC LETTER SHEEN
0635	ص	ARABIC LETTER SAD
0636	ض	ARABIC LETTER DAD
0637	ط	ARABIC LETTER TAH
0638	ظ	ARABIC LETTER ZAH
0639	ع	ARABIC LETTER AIN

FE9	FEA	FEB	FEC
ب FE90	ج FEA0	ز FEB0	ظ FEC0
د FE91	ح FEA1	س FEB1	ط FEC1
ذ FE92	ح FEA2	س FEB2	ط FEC2
ة FE93	ح FEA3	س FEB3	ط FEC3
ة FE94	ج FEA4	س FEB4	ط FEC4

Non-Uniqueness in Unicode

“Precomposed” characters

Canonical Equivalence

1	Å	U+212B ANGSTROM SIGN
	Å	U+00C5 LATIN CAPITAL LETTER A WITH RING ABOVE
	A ◌̊	U+0041 LATIN CAPITAL LETTER A, U+030A COMBINING RING ABOVE
2	x ◌̊ ◌̇	U+0078 LATIN SMALL LETTER X, U+031B COMBINING HORN, U+0323 COMBINING DOT BELOW
	x ◌̇ ◌̊	U+0078 LATIN SMALL LETTER X, U+0323 COMBINING DOT BELOW, U+031B COMBINING HORN
3	ʊ̊̇	U+1EF1 LATIN SMALL LETTER U WITH HORN AND DOT BELOW
	ʊ̊̇	U+1EE5 LATIN SMALL LETTER U WITH DOT BELOW, U+031B COMBINING HORN
	u ◌̊ ◌̇	U+0075 LATIN SMALL LETTER U, U+031B COMBINING HORN, U+0323 COMBINING DOT BELOW
	ʊ̊̇	U+01B0 LATIN SMALL LETTER U WITH HORN, U+0323 COMBINING DOT BELOW
	u ◌̇ ◌̊	U+0075 LATIN SMALL LETTER U, U+0323 COMBINING DOT BELOW, U+031B COMBINING HORN

“Normalization” problems

Figure 5. Multiple Combining Marks

Source		NFD	NFC
Š	:	Š ̇ ̈	Š
1E69		0073 0323 0307	1E69
đ	:	đ ̇ ̈	đ ̈
1E0B 0323		0064 0323 0307	1E0D 0307
q̇	:	q ̇ ̈	q ̇ ̈
0071 0307 0323		0071 0323 0307	0071 0323 0307

“Normalization” problems

Figure 2. Compatibility Equivalence

Font variants	Œ	H
Breaking differences	—	
Cursive forms	ا	ن
Circled	①	
Width, size, rotated	カ	{
Superscripts/subscripts	9	9
Squared characters	アバ —ト	
Fractions	¼	
Others	dz	

“Normalization” problems

Figure 6. Compatibility Composites

Source		NFD		NFC		NFKD		NFKC
fi FB01	:	fi FB01		fi FB01		f i 0066 0069		f i 0066 0069
2 ⁵ 0032 2075	:	2 5 0032 2075		2 5 0032 2075		2 5 0032 0035		2 5 0032 0035
fi 1E9B 0323	:	f ̇ ̇ 017F 0323 0307		fi ̇ 1E9B 0323		s ̇ ̇ 0073 0323 0307		š 1E69

“Normalization”: What to store?

- Recommended strategy for precomposed characters:
 - a) Total decomposition (NFD)
 - b) Maximal composition (NFC)

“Normalization” strategies

Table 6. Basic Examples

	Original	NFD, NFKD	NFC, NFKC	Notes
a	D-dot_above	D + dot_above	D-dot_above	Both decomposed and precomposed canonical sequences produce the same result.
b	D + dot_above	D + dot_above	D-dot_above	
c	D-dot_below + dot_above	D + dot_below + dot_above	D-dot_below + dot_above	The <i>dot_above</i> cannot be combined with the D because the D has already combined with the intervening <i>dot_below</i> .
d	D-dot_above + dot_below	D + dot_below + dot_above	D-dot_below + dot_above	
e	D + dot_above + dot_below	D + dot_below + dot_above	D-dot_below + dot_above	
f	D + dot_above + horn + dot_below	D + horn + dot_below + dot_above	D-dot_below + horn + dot_above	There may be intervening combining marks, so long as the result of the combination is canonically equivalent.
g	E-macron-grave	E + macron + grave	E-macron-grave	Multiple combining characters are combined with the base character.
h	E-macron + grave	E + macron + grave	E-macron-grave	
i	E-grave + macron	E + grave + macron	E-grave + macron	Characters will <i>not</i> be combined if they would not be canonical equivalents because of their ordering.
j	angstrom_sign	A + ring	A-ring	Because Å (A-ring) is the preferred composite, it is the form produced for both characters.
k	A-ring	A + ring	A-ring	

“Normalization”: What to store?

- Recommended strategy for “compatibility equivalents”?
 - a) Total decomposition (NFKD)
 - b) Maximal canonical composition (NFKC)

“Normalization” strategies

Title	Description
Normalization Form D (NFD)	Canonical Decomposition
Normalization Form C (NFC)	Canonical Decomposition, followed by Canonical Composition
Normalization Form KD (NFKD)	Compatibility Decomposition
Normalization Form KC (NFKC)	Compatibility Decomposition, followed by Canonical Composition

“Normalization” strategies

Table 7. NFD and NFC Applied to Compatibility-Equivalent Strings

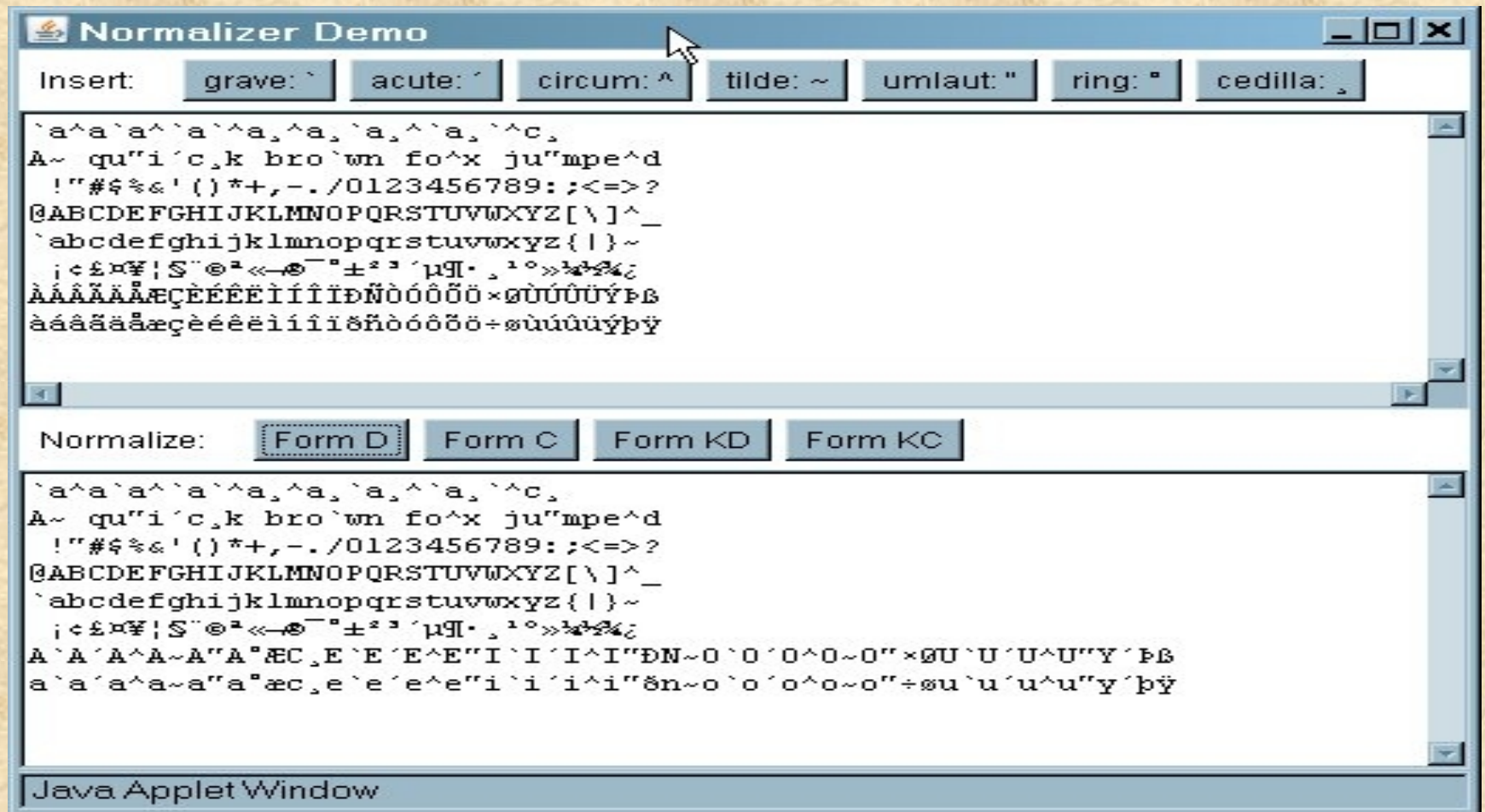
	Original	NFD	NFC	Notes
l	"Äffin"	"A\u0308ffin"	"Äffin"	The <i>ffi_ligature</i> (U+FB03) is <i>not</i> decomposed, because it has a compatibility mapping, not a canonical mapping. (See Table 8.)
m	"Ä\uFB03n"	"A\u0308\u0303n"	"Ä\uFB03n"	
n	"Henry IV"	"Henry IV"	"Henry IV"	Similarly, the ROMAN NUMERAL IV (U+2163) is <i>not</i> decomposed.
o	"Henry \u2163"	"Henry \u2163"	"Henry \u2163"	
p	ga	ka + ten	ga	Different compatibility equivalents of a single Japanese character will <i>not</i> result in the same string in NFC.
q	ka + ten	ka + ten	ga	
r	hw_ka + hw_ten	hw_ka + hw_ten	hw_ka + hw_ten	
s	ka + hw_ten	ka + hw_ten	ka + hw_ten	
t	hw_ka + ten	hw_ka + ten	hw_ka + ten	Hangul syllables are maintained under normalization.
u	kaks	ki + a _m + ks _f	kaks	

“Normalization” strategies

Table 8. NFKD and NFKC Applied to Compatibility-Equivalent Strings

	Original	NFKD	NFKC	Notes
l'	"Äffin"	"A\u0308ffin"	"Äffin"	The <i>ffi_ligature</i> (U+FB03) is decomposed in NFKC (where it is not in NFC).
m'	"Ä\uFB03n"	"A\u0308ffin"	"Äffin"	
n'	"Henry IV"	"Henry IV"	"Henry IV"	Similarly, the resulting strings here are identical in NFKC.
o'	"Henry \u2163"	"Henry IV"	"Henry IV"	
p'	ga	ka + ten	ga	Different compatibility equivalents of a single Japanese character <i>will</i> result in the same string in NFKC.
q'	ka + ten	ka + ten	ga	
r'	hw_ka + hw_ten	ka + ten	ga	
s'	ka + hw_ten	ka + ten	ga	
t'	hw_ka + ten	ka + ten	ga	
u'	kaks	ki + am + ksf	kaks	Hangul syllables are maintained under normalization.*

“Normalization” strategies



“Normalization” strategies

Normalizer Demo

Insert:

``a^a`a^`a`^a,^a,`a,^`a,`^c,
A~ qu"i'c,k bro`wn fo^x ju"mpe^d
!"#$%&'()*+,-./0123456789:;<=>?
@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_
`abcdefghijklmnopqrstuvwxyz{|}~
¡¢£¥¦§¨ª«¬®¯°±²³´µ¶·¸¹º»¼½¾¿
ÀÁÂÃÄÅÆÇÈÉÊËÌÍÎÏÐÑÒÓÔÕÖ×ØÙÚÛÜÝÞß
àáâãäåæçèéêëìíîïðñòóôõö÷øùúûüýþÿ`

Normalize:

``ââââ^â,â,â,â,^ç
Ã quïçk bròwn fôx jümpêd
!"#$%&'()*+,-./0123456789:;<=>?
@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_
`abcdefghijklmnopqrstuvwxyz{|}~
¡¢£¥¦§¨ª«¬®¯°±²³´µ¶·¸¹º»¼½¾¿
ÀÁÂÃÄÅÆÇÈÉÊËÌÍÎÏÐÑÒÓÔÕÖ×ØÙÚÛÜÝÞß
àáâãäåæçèéêëìíîïðñòóôõö÷øùúûüýþÿ`

Java Applet Window

“Normalization” strategies

Normalizer Demo

Insert:

Java Applet Window

“Normalization” strategies

Normalizer Demo

Insert:

``a^a`a^`a`^a,^a,`a,^`a,`^c,
A~ qu"i'c,k bro`wn fo^x ju"mpe^d
!"#$%&'()*+,-./0123456789:;<=>?
@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_
`abcdefghijklmnopqrstuvwxyz{|}~
¡¢£¥¦§¨ª«¬®¯°±²³´µ¶·¸¹º»¼½¾¿
ÀÁÂÃÄÅÆÇÈÉÊËÌÍÎÏÐÑÒÓÔÕÖ×ØÙÚÛÜÝÞß
àáâãäåæçèéêëìíîïðñòóôõö÷øùúûüýþÿ`

Normalize:

``ââââ^â,â,â,â,^ç
Ã qüiçk bròwn fôx jümpêd
!"#$%&'()*+,-./0123456789:;<=>?
@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_ |
`abcdefghijklmnopqrstuvwxyz{|}~
¡¢£¥¦§¨ª«¬®¯°±²³´µ¶·¸¹º»¼½¾¿
ÀÁÂÃÄÅÆÇÈÉÊËÌÍÎÏÐÑÒÓÔÕÖ×ØÙÚÛÜÝÞß
àáâãäåæçèéêëìíîïðñòóôõö÷øùúûüýþÿ`

Java Applet Window

Recommendations as to “Normalization”:

- Question of storage vs. question of retrieval?
 - File size?
 - Sorting
 - Searching
 - Comparing
- “On the fly” interpretation of data
 - today?
 - tomorrow?

“Too-Uniqueness” problems: Bidirectionality: Arabic punctuation

Punctuation

060C	ﷂ	ARABIC COMMA
		• also used with Thaana and Syriac in modern text → 002C , comma
060D	٫	ARABIC DATE SEPARATOR

Punctuation

06D4	-	ARABIC FULL STOP
		• Urdu

Punctuation

061B	؛	ARABIC SEMICOLON
		• also used with Thaana and Syriac in modern text → 003B ; semicolon
061C	ⵜ	<reserved>
061D	ⵓ	<reserved>
061E	ⵓ	ARABIC TRIPLE DOT PUNCTUATION MARK
061F	؟	ARABIC QUESTION MARK
		• also used with Thaana and Syriac in modern text → 003F ? question mark

- Missing (> Latin equivalents):
 - “Normal” full stop
 - Exclamation mark
 - Quotation marks
 - Parentheses, brackets, braces...

“Bidirectionality” character types

Table 4. Bidirectional Character Types

Category	Type	Description	General Scope
Strong	L	Left-to-Right	LRM, most alphabetic, syllabic, Han ideographs, non-European or non-Arabic digits, ...
	LRE	Left-to-Right Embedding	LRE
	LRO	Left-to-Right Override	LRO
	R	Right-to-Left	RLM, Hebrew alphabet, and related punctuation
	AL	Right-to-Left Arabic	Arabic, Thaana, and Syriac alphabets, most punctuation specific to those scripts, ...
	RLE	Right-to-Left Embedding	RLE
	RLO	Right-to-Left Override	RLO
Weak	PDF	Pop Directional Format	PDF
	EN	European Number	European digits, Eastern Arabic-Indic digits, ...
	ES	European Number Separator	PLUS SIGN, MINUS SIGN
	ET	European Number Terminator	DEGREE SIGN, currency symbols, ...
	AN	Arabic Number	Arabic-Indic digits, Arabic decimal and thousands separators, ...
	CS	Common Number Separator	COLON, COMMA, FULL STOP (<i>period</i>), NO-BREAK SPACE, ...
	NSM	Nonspacing Mark	Characters marked Mn (Nonspacing_Mark) and Me (Enclosing_Mark) in the Unicode Character Database
	BN	Boundary Neutral	Most formatting and control characters, other than those explicitly given types above
Neutral	B	Paragraph Separator	PARAGRAPH SEPARATOR, appropriate Newline Functions, higher-level protocol paragraph determination
	S	Segment Separator	<i>Tab</i>
	WS	Whitespace	SPACE, FIGURE SPACE, LINE SEPARATOR, FORM FEED, General Punctuation spaces, ...
	ON	Other Neutrals	All other characters, including OBJECT REPLACEMENT CHARACTER

“Bidirectionality” character types

Misbehaviour problems

τοῦ ἐσταυρωμένου
ἀπεξήρανεν·

අප දින කීයෝ . අප දින කීයෝ
කඩුවට තදවැටුණා.

τὸν δὲ τὸν ἄρχοντα
τῶν δαιμόνων
ἠρώτησα ὀργίσας

120

א.א.א.

αὐτὸν τοῦ σταυροῦ

၁၆၃

हुल्ल

τὴν δύναμιν ὥστε εἰπεῖν μοι τὴν δύναμιν τοῦ σημείου·

കുറവ്.

കുറുപ്പുകൾ

Recommendations:

- This is an implementation problem, not a storage problem!
 - Arguing with software providers for a correct treatment?
 - Arguing with UNICODE.ORG for an addition of RLM punctuation marks?


“Original” scripts vs. transcription / transliteration?

- “Original” scripts preferred by native speakers / communities?
- Transcriptions preferred by linguists?

Twofold ELAN output

Elan - BAV03_04.EAF

Datei Bearbeiten Annotation Zeile Typ Suche Ansicht Optionen Fenster Hilfe



00:02:23.975

Auswahl: 00:00:00.000 - 00:00:00.000 0

Tabelle	Text	Untertitel	Steuerung
tr1@AS	ბატკახი ლევანათერ მენ, ბაცბი ცო მიწრენა ახაჲცნ ხალხ ცო ახ ეზე ბჭარჩე' ესე დახერ, 'ალიხ ბარეჰა დარ ებეე, ხკულიხ ლოლუმს იხორ.		
tl1@AS	baqqaxi levanater men, bacbi co mičrena yaxken xalx co ya ebe bwarče' ese daxer, 'alix bareħa dar ebee, xkuliḥ l oumā ixor.		
vfg@AS	უფროსები ამბობდნენ, რომ ბაცებები არსაიდან მოსული ხალხი არ არის, აი ესენი დღენიადაგ აქ ცხოვრობდნენ. ზამთარში ბარში იყვნენ ესენი და ზაფხულში მთაში დადიოდნენ.		
ve@AS	The old people used to say that the Batsbis are not a people that has come from somewhere, they have lived here ever since, in the winter they were in the plain and in the summer they went up into the mountain(s).		

ref@AS

ref@AS	tr@AS	tl@AS	vfg@AS	vfg1@AS	ve@AS
02 24.000 00:02:25.000 00:02:26.000 00:02:27.000 00:02:28.000 00:02:29.000 00:02:30.000 00:02:31.000 00:02:32.000	ოკუნდალა დარ მე, ბაცავ სტაკ ბჭა გარე მხრობახ ვარ, 'ალიხ ბარე	oquyndalla dar me, bacav ştaḥ bwa gare mxrobax var, 'alix bare	ოკუნდალა დარ მე, ბაცავი (თუშ) კაცი ყოველთვის გარე მხარეებში იყო, ზამთარში ბარში გარეთ დადიოდა ცხვრის უკან, საქონლის უკან, ცხენების უკან,	amiŋom iqo, rom, bacavi (tuşi) ḳacı qoveltvis gare mxareebši iqo, zamtarşi barşi garet dadioda cxvris uḳan, sakonlis uḳan, cxenebis uḳan,	Therefore the Bateman has always been in outer regions; in the winter he went into the plain after the sheep, after the cattle, after the horses

“Original” script vs. transcription / transliteration?

- Ideal scenario:
 - transcription automatically derivable from rendering in original script and
 - vice versa
- No problems with Cyrillic vs. Latin
- Manifold problems with Arabic vs. Latin

(Narrow) Transcription

<> Cyrillic script (no font mapping!)



Satz / sentence 1

Нартæн уæд сæ хистæр Уæрхæг уыдис.

Nartæn uæd sæ xistær Uærxæg uydīs.

Nartæn sæ xistær wæd Wærxæg wydīs.

The Narts' eldest was Wærxæg then.

<i>Nartæn</i>	<i>sæ</i>	<i>xistær</i>	<i>wæd</i>	<i>Wærxæg</i>	<i>wydīs</i>
Narts	their	eldest	then	Wærxæg (PN)	was
dat.pl.	poss.pron.procl.	nom.sg.	adv.	nom.sg.	3.sg.pret.



Satz / sentence 2

Уæрхæгæн райгуырдис дыууæ лæппуйы, фаззæттæ.

Uærxægæn rajguyrdis dyuuæ læppujy, fazzættæ.

Wærxægæn rajgwyrdis dywwæ læppujy, dywwæ læppujy, dywwæ fazzony.

Two boys were born to Wærxæg, two boys, two twins.

<i>Wærxægæn</i>	<i>rajgwyrdis</i>	<i>dywwæ</i>	<i>læppujy,</i>	<i>dywwæ</i>	<i>læppujy,</i>	<i>dywwæ</i>	<i>fazzony</i>
Wærxæg (PN)	was born	two	boys	two	boys	two	twins
dat.sg.	3.sg.pret.	num.	gen.sg.	num.	gen.sg.	num.	gen.sg.

(Narrow) Transliteration

<> Arabic script

Page of edition: 1

Chapter: 1

bsm 'llh 'lrhmn 'lrhym

Strophe: 1

Verse: a sp's w 'fryn 'n p'dš' r'

Verse: b kh gyty r' pdyd 'wrd w m'r'

Strophe: 2

Verse: a bdw zyb'st mlk w p'dš'yy

Verse: b kh hr gz n'yd 'z mlkš ġd'yy

Strophe: 3

Verse: a xd'y p'k w by hmt' w by y'r

Verse: b hm 'z 'ndyšh dwr w hm z dyd'r

Strophe: 4

Verse: a nh btw'nd mrw r' čšm dydn

Verse: b nh 'ndyšh drw d'nd rsydn

Strophe: 5

Verse: a nh nqš'ny pdyrd hmčw ġwhr

Verse: b nh z'n grdd mrw r' ħ'l dygr

Page of edition: 1

Chapter: 1

بسم الله الرحمن الرحيم

Strophe: 1

Verse: a سپاس و آفرین آن پادشا را

Verse: b که گیتی را پدید آورد و مارا

Strophe: 2

Verse: a بدو زیباست ملک و پادشایی

Verse: b که هرگز ناید از ملکش جدایی

Strophe: 3

Verse: a خدای پاک و بی همتا و بی یار

Verse: b هم از اندیشه دور و هم ز دیدار

Strophe: 4

Verse: a نه بتواند مرو را چشم دیدن

Verse: b نه اندیشه درو داند رسیدن

Strophe: 5

Verse: a نه نقصانی پذیرد همچو جوهر

Verse: b نه زان گردد مرو را حال دیگر

(Broad) Transliteration

<> Arabic script

Chapter: 1

Section / Image: 1

Page: 31

Line: 1

'btd'y kt'b sndb'g

Line: 2 čnyn gwynd r'wy'ni hdyt w hgd'wnd'ni t'ryh ky dr mwšy'yy'm
 Line: 3 w sw'lfī 'w'm dr 'qlymi hndwst'n y'dš'hy bwgh 'st kwr dys n'm ky šh'yf
 Line: 4 m'ly ġh'nd'ry r' bmk'rmi 'hl'qhi hmydh mwsh grd'nydh bwz. wrd'y mf'hri
 Line: 5 p'dš'hy r' bm'tri 'r'qi krym mtrrz krdh, w rwzg'ri 'w bğm'li 'dl 'r'sth
 Line: 6 w 'wš'fi w bkm'li fšl mšhwr šđh, dwlty m't' w hšmty m'ty'm, mddty
 Line: 7 twyl w mmlkty 'rys, dsti tn'wli h'sd'n w t't'wli q'sd'n 'z mmlkti
 Line: 8 'w bsth w kwt'h, w čšmi 'tm'i f'sdh' m't'ddy'n dr dwlt 'w pwšydh w fr'z,
 Line: 9 hmyšh mut'b'i 'dl w m't'w'i 'ql bwgy, w 't'r w 'hb'ri rftg'n w sunan wi w slayri
 Line: 10 'yn'n šnwgy, w đlkr ħusni šiyam w šyti m't'w'ti hdm w ħašami 'w bsm'
 Line: 11 sl'tyni wqt rsydh, w zb'ni ruw't w by'ni tiq't 'w'zh' r'fhyti r'yt
 Line: 12 w ħšb w'mni wlyti 'w bgwši hl'yq rs'nydh, w 'z bdwi šibay ky 'mrh'
 Line: 13 'mr ġrri' dhr 'st' t' t'l'w'i šb'hi šayb ky ħbr dhndh' w d'i' hy'tst ġz dr
 Line: 14 mnħgi r'yt w mslki tħfyf w tħfyhi š'fy wlyt qdm nzdh bwg, w 'z
 Page: 32 Line: 15 br'y 'kt'bi 'mw'l g'my dr ħaħ' wizr w wb'l nnh'gh bwg, pywsth 'htm'm
 Line: 16 br 'tm'mi mš'ħi r'y'y dwlt mwfwr my d'st, w br w bħri mulkt r' b'fšti
 Line: 17 nšft w 'š'ti m'dlt m'mwr my grd'nyd, dwlti 'wr' s'di 'kbri 'qlymi
 Line: 18 zħl my gftnd, w mlwki 'f'q mk'rmi 'hl'qi 'w br ħ'šyh' ġrydh' sy'st
 Line: 19 t'lyq my krdnd, w 'z fš'yli 'lmi 'w 'qtb's my nmwgdnd w dr
 Line: 20 n't w wšfi 'w my gftnd:

Chapter: 1

Section / Image: 1

Page: 31

Line: 1

ابتدای کتاب سندباد

Line: 2 چنین گویند راویان حدیث و خداوندان تاریخ کی در موصیایام
 Line: 3 و سوالفی اعوام در اقلیم هندوستان یازشاهی بوزه است کوردیس نام کی صحایف
 Line: 4 معالی جهانداری را بمکارم اخلاق حمیده موسخ گردانیده بوز، وردای مفاخر
 Line: 5 پادشاهی را بمآثر اعراق کریم مطرز کرده، و روزگار او بجمال عدل آراسته
 Line: 6 و اوصاف و بکمال فصل مشهور شده، دولتی مطاع و حشمتی مطیع، مدتی
 Line: 7 طویل و مملکتی عریض، دست تناول حاسدان و تناول قاصدان از مملکت
 Line: 8 او بسته و کوتاه، و چشم اطماع فاسده متعبدان در دولت او پوشیده و فراز،
 Line: 9 همیشه متابع عدل و مطاوع عقل بوزی، و آثار و اخبار رفتگان و ستن و وسیر
 Line: 10 اینان شنودی، و ذکر حسن شیم و صیت مطاوعت خدم و حشم او بسمع
 Line: 11 سلاطین وقت رسیده، و زبان رواں و بیان ثقات آوازه رفاهیت رعیت
 Line: 12 و خصی و امن ولایت او بگوش خلاق رسانیده، و از بدو صیتی کی عمره
 Line: 13 عمر غره دهر است تا طلوع صباح شیب کی خبر دهنده و داع حیانتست جز در
 Line: 14 منہج رعایت و مسلک تخفیف و ترفیه صغای ولایت قدم نزده بود، و از
 Page: 32 Line: 15 برای اکتاب اموال گامی در خطه وزر و وبال نهاده بود، پیوسته اهتمام
 Line: 16 بر اتمام مصالح رعابای دولت موفور می داشت، و بر و بحر ملک را بافاصت
 Line: 17 نصفت و اشاعت معدلت معمور می گردانید، دولت اورا سعید اکبر اقلیم
 Line: 18 زحل می گفتند، و ملوک آفاق مکارم اخلاق او بر حاشیه جریده سیاست
 Line: 19 تعلیق می کردند، و از فصایل علم او اقتباس می نمودند و در
 Line: 20 نعت و وصف او می گفتند:

(Broad) Transcription

<> Arabic script

Verse: 1

Half verse: 1 *ay dil am bahr e qabūl ī ba tanā y at mašgūl*

Half verse: 2 *rad našud har ke ba dargāh e to gardīd qabūl*

Verse: 2

Half verse: 1 *bar dar e dark e kamāl e to nabāšad rah e 'aql*

Half verse: 2 *ke ba juz 'ajz nadārand dar īn bāb 'uqūl*

Verse: 3

Half verse: 1 *'ārifān rā ke zi dāniš ba falak bar šude and*

Half verse: 2 *hast dar manzil e nādānī y īšān az to nuzūl*

Verse: 4

Half verse: 1 *ḡayr e iqrār ba tawḥīd e to ay šānī' e pāk*

Half verse: 2 *'āqilān rā hame tawjīh buwad nāma'qūl*

Verse: 5

Half verse: 1 *li l lahi l ḥamd ke šud dar naẓar e ahl e kamāl*

Half verse: 2 *az qabūl e naẓar at naẓm e qabūlī maqbūl*

Verse: 1

Half verse: 1 *أَي دِل اَم بَهْر اَقْبُولِي بَ تَنَاي اَت مَشْغُول*

Half verse: 2 *رَد نَشُد هَر كَ بَ دَر گَاه اَت گَرْدِيد قَبُول*

Verse: 2

Half verse: 1 *بَر دَر اَدَر كَ اَكْمَال اَت نَبَاشَد رَه اَعْقَل*

Half verse: 2 *كَ بَ جُز عَجَز نَدَارَنَد دَرِيْن بَاب عُقُول*

Verse: 3

Half verse: 1 *عَارِفَان رَا كَ ز دَانِش بَ قَلَك بَر شُد اَنَد*

Half verse: 2 *هَسْت دَر مَنَزَل اَنَادَانِيْ اِيْشَان اَز تُوْزُل*

Verse: 4

Half verse: 1 *غَيْر اَقْرَار بَ تَوْحِيد اَت اَي صَانِع اُپَاك*

Half verse: 2 *عَاقِلَان رَا هَم تَوْجِيه بُود نَامَعْقُول*

Verse: 5

Half verse: 1 *لِ ل لِه ل حَمْد كَ شُد دَر نَظَر اَ اَهْل اَكْمَال*

Half verse: 2 *اَز قَبُول اَنَظَر اَت نَظْم اَقْبُولِي مَقْبُول*

“Original” script vs. transcription / transliteration?

- Recommendations:
 - For data storage choose the most informative rendering available
 - with a unique representation of all consonant and vowel phonemes (“broad” transcription type)
 - do care for convertibility
 - N.B. a plain ASCII-based encoding may suffice (!)

Summary

- The struggle for a reliable encoding basis is approaching its end with the UNICODE standard developing
BUT
- Inconsistencies of the UNICODE standard should be considered carefully for data storage right from the beginning