Università di Padova
Dipartimenti di Studi Linguistici e Letterari

Tecnologie per la Traduzione 2020/2021

Codifica del testo

Giorgio Maria Di Nunzio

Obiettivi

- Comprendere la definizione della struttura dei file utilizzati dai software di traduzione assistita.

- Vedere ad esempio
  - https://www.w3.org/2002/02/01-i18n-workshop/LocFormats

- Software che utilizzano file tmx (memorie di traduzione)
  - https://en.wikipedia.org/wiki/Translation_Memory_eXchange

- Software che utilizzano file tbx (terminologia)
  - https://en.wikipedia.org/wiki/TermBase_eXchange

Codifica di un documento testuale

- Dalla codifica dei caratteri alla codifica del testo
- Che cosa è uno “standard” di codifica
- Standard di codifica dei caratteri
- Standard di codifica del testo

Struttura vs Contenuto

- Separazione tra contenuto, ovvero dalla successione di parole che lo compongono, e presentazione.

- La presentazione, e cioè il modo in cui il contenuto viene strutturato e organizzato, è diventata sempre più importante

  - Monitor e stampanti

  - “Leggibilità” di un documento a seconda del dispositivo

- Una primo approccio verso la strutturazione del testo era presente già nella codifica ASCII.

Elementi base della struttura di un testo

- Formato dei singoli caratteri
- Formato Interlinea
- Formato dei paragrafi
- Formato delle pagine

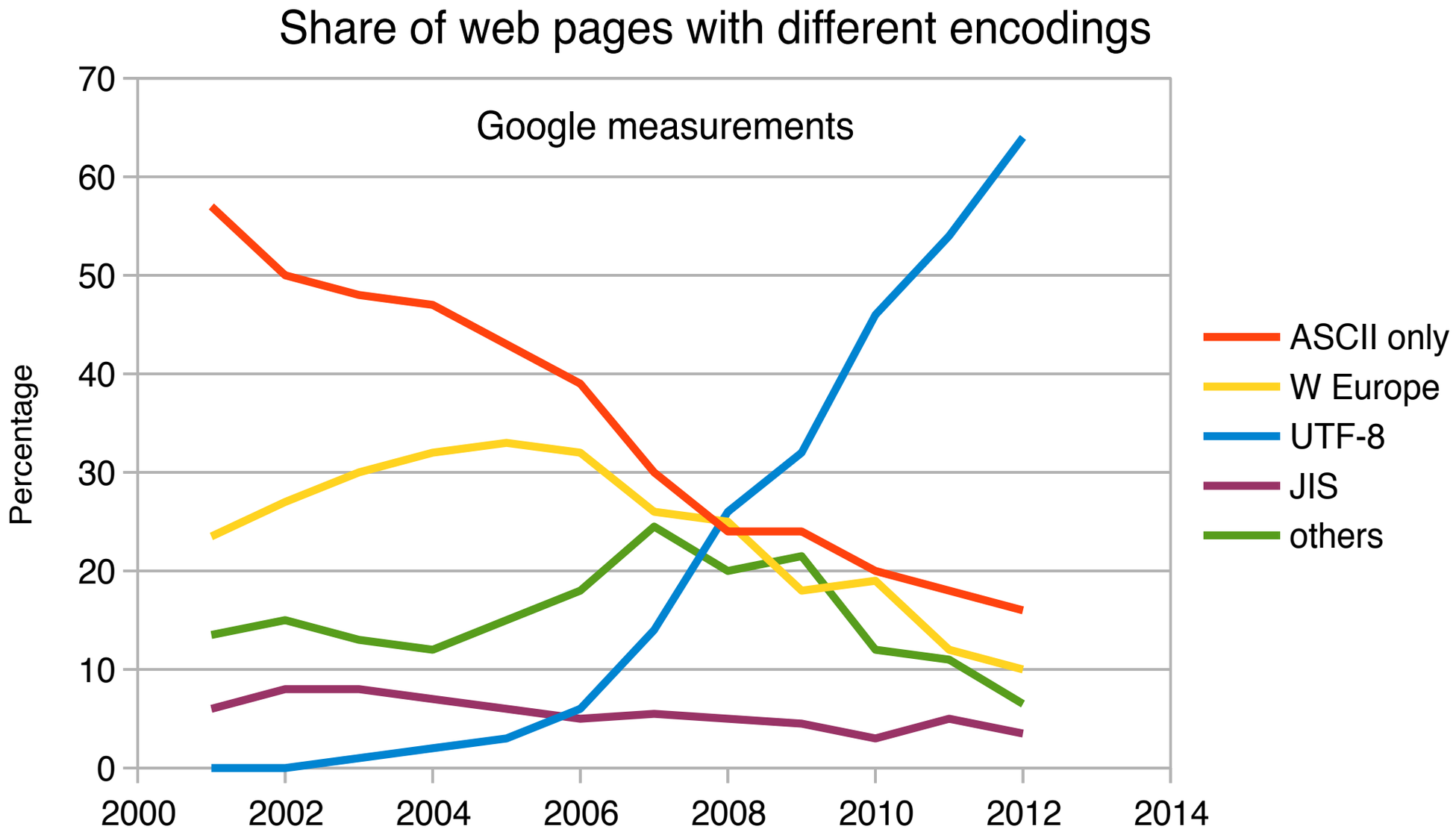
Standard di codifica

Uno standard è uno schema o un complesso di norme stabilito da un’autorità o basato su un consenso generale che definisce un modello o un esempio di riferimento al quale uniformarsi.

Codifica dei caratteri

- 1963 ASCII (American Standard Code for Information Interchange), 7 bit
  - 1973, ISO-7 (International Standard Organization)
- 1987 ASCII Extended ISO 8859-x, 8 bit
- 1991 Unicode (Universal Code character set), 16 bit
- 1993 UTF-8 (Uniform Transformation Format), (da 8 a 32 bit)

Utilizzo della codifica UTf-8



Standard di codifica del testo

- Dallo *Standard Generalized Markup Language* (SGML)
- All'Hypertext Markup Language (HTML)
- All'*eXtensible Markup Language* (XML)
- La *Text Encoding Initiative* (TEI)

L’origine dello SGML

- 1974 SGML, Charles Goldfarb

“We were trying to do an automated law-office application. I had been a lawyer (in fact, I still am). Lawyers must do research on existing case law, decisions of court, and so on, to find out which ones are applicable to a given situation, find out what the previous legal rulings have been, and then merge that with text that the lawyer has written himself. Eventually, if it's, say, a brief for the court, he must then compose it and print it. At the time, which was 1969 or 1970, there weren't any systems available that did these three things. So in order to get the systems to share the data we had to come up with a way to represent it that was independent of any of those applications. It was a very small research project...”