

Textkodierung und Auszeichnung

Matthias Bethke bethke@linguistik.uni-erlangen.de
Besim Kabashi bmkabash@linguistik.uni-erlangen.de

Linguistische Informatik
Universität Erlangen-Nürnberg

Sommersemester 2006

1 Organisatorisches

2 Einführung

- Warum?
- Beispiele
- Historische Anfänge
- Von EBCDIC bis ASCII

Organisatorisches

- ① Termin
- ② Voraussetzungen und Ablauf
 - Kodierung: Zeichensätze, Unicode, Fonts, Locales, ...
 - Auszeichnung: SGML, HTML, XML, TEI, \LaTeX , ...
- ③ Literatur
- ④ WWW: <http://www.linguistik.uni-erlangen.de/~msbethke/teaching/SS2006-Kodierung.html>

Warum das Ganze?

- Ein Großteil der Computerlinguistik befasst sich mit der Verarbeitung von geschriebenen Texten
- Vorgefundene Texte insbesondere in der Korpuslinguistik
- Unterschiedlichste Quellen, Sprachen, Medien, . . .
- Ziel: korrekte Erhaltung/Konvertierung und Verarbeitung von textuellen und Metainformationen

Darum: ein tieferes Verständnis von den zugrundeliegenden Verfahren ist für Computerlinguisten wichtig.

Beispiele (1)

Mail von Assunção Verônica Álvares an Kryštůfek Březový

Subject: Test

From: =?iso-8859-1?Q?Assun=E7=E3o_Ver=F4nica_
=C1lvares?= <veronica@invalid>

To: =?iso-8859-2?Q?Kry=B9t=F9fek_R=0C9=08eov=FD=?=
<brezovy@nowhere>

Content-Type: text/plain; charset="iso-8859-2"

Pøíli ¾lu»ouèký kùò úpìl ïábelské ódy

P?íli? ?lu?ou?ký kùò úp?l ?ábelské ódy

Příliš žluťoučký kůň úpěl ďábelské ódy

Beispiele (2)

Ausschnitt aus einer Webseite

```
<tr class="igalltxt"><td>Many of the islands have grottos.  
&#272;&#x1ec7;ng Thiên Cung (Heavenly Palace Cave) is not the  
largest but probably the most beautiful .</td></tr>
```

Probleme und Fragen zu Zeichensätzen

- 1 Welche Sprachen will ich verarbeiten und welchen Zeichenvorrat brauche ich dafür?
- 2 Woher stammt mein Material (z.B. Korpora) dafür, und in welcher Darstellung liegt es vor?
- 3 Wie bekomme ich es unverfälscht vom Speichermedium in mein Programm und von dort zur Darstellung auf dem Bildschirm/Drucker?

Beispiele (3)

Eintrag eines maschinenlesbaren Lexikons

```
<!DOCTYPE LEXICON SYSTEM "LEXICON-1.DTD">
<LEXICON><LETTER NAME="a"><ENTR LEX="Aa1" ID="738.01"
TYP="NORM"><WRT>Aa1</WRT><SPZ><GRA>m. 1</GRA>
<POL NR="1"><EXPL>langer, schlangen&auml;hnlicher Fisch: Anguilla
vulgaris </EXPL></POL> </SPZ><ETY>&Ety2;ahd., as&auml;chs.
<X>al,</X></ETY></ENTR></LETTER></LEXICON>
```

Probleme und Fragen zum Textmarkup

- 1 Was ist *Markup*, und warum und wie verwende ich das?
- 2 Welches Markup für welchen Zweck?
- 3 Haben meine Textquellen irgendeine erkennbare Struktur, und inwiefern muss ich diese erhalten/erkennen/verarbeiten/umwandeln?

Von Zahlen zu Buchtaben

- Vor 1949: fast nur numerische Anwendungen für Computer (Codebrecher für Militär u.ä.)
- Kommerzielle Anwendungen zuerst von IBM → Notwendigkeit einer Zeichendarstellung
- Zunächst *alphameric*: Zahlen und Großbuchstaben in einem 6-bit-Code
- 1963: Entwicklung des 8-bit *Extended Binary Coded Decimal Interchange Code* (EBCDIC), der immer noch auf einigen IBM-Architekturen benutzt wird.
- Nachteil: „eigenwillige“ Verteilung der Buchstaben und anderen Zeichen wg. historischer Entwicklung aus Lochkarten-Code
- Spätere Weiterentwicklung: begrenzte Internationalisierung

Codetabelle für EBCDIC

0	10	20	30	40	50	60	70	80	90	A0	B0	C0	D0	E0	F0	
NUL(1)		DS(2)		SP(3)	&(4)	_(5)									0	0
		SOS				/		a	j			A	J			1
		FS						b	k	s		B	K	S	2	2
	TM							c	l	t		C	L	T	3	3
PF	RES	BYP	PN					d	m	u		D	M	U	4	4
HT	NL	LF	RS					e	n	v		E	N	V	5	5
LC	BS	EOB	UC					f	o	w		F	O	6	6	
DEL	IL	PRE														
								h								
								i								
				¢	!											
				.	\$											
				<	*											
				()											
				+	;											
						?										
				12												

ASCII

- 1963: ANSI-Standard *ASCII* (American Standard Code for Information Interchange).
- 7-bit Code (128 Positionen)
- Zunächst nur Großbuchstaben, Kleinbuchstaben ab 1967
- Beschränkt auf den Zeichenvorrat des Englischen plus relativ viele, mittlerweile größtenteils obsolete, Steuerzeichen („*CO*“ *control codes*, Pos. 0–31)

Codetabelle für ASCII

1. Ziffer →
2. Ziffer ↓

	0	1	2	3	4	5	6	7
0	NUL	DLE	SPC	0	@	P	'	p
1	SOH	DC1	!	1	A	Q	a	q
2	STX	DC2	"	2	B	R	b	r
3	ETX	DC3	#	3	C	S	c	s
4	EOT	DC4	\$	4	D	T	d	t
5	ENQ	NAK	%	5	E	U	e	u
6	ACK	SYN	&	6	F	V	f	v
7	BEL	ETB	'	7	G	W	g	w
8	BS	CAN	(8	H	X	h	x
9	TAB	EM)	9	I	Y	i	y
A	LF	SUB	*	:	J	Z	j	z
B	VT	ESC	+	;	K	[k	{
C	FF	FS	,	<	L	\	l	
D	CR	GS	-	=	M]	m	}
E	SO	RS	.	>	N	^	n	~
F	SI	US	/	?	O	_	o	DEL