

# BBN-ANG-183 Typography

## Lecture 3: Electronic texts

Zoltán G. Kiss & Péter Szigetvári

Dept of English Linguistics, Eötvös Loránd University

# outline

## encoding characters

- ASCII

- nonASCII

- alternatives

- Unicode

## encoding form

- markup

- physical vs. logical markup

- logical markup at its best

- the rationale of markup

## WYSIWYG

- comparison of markup vs. WYSIWYG

# encoding

- ▶ text is not encoded as an image (dot-by-dot)

# encoding

- ▶ text is not encoded as an image (dot-by-dot)
- ▶ but as characters

# encoding

- ▶ text is not encoded as an image (dot-by-dot)
- ▶ but as characters
- ▶ that is, the following

à    **à**    a    a    *a*

are all encoded as an 'a'

## encoding

- ▶ text is not encoded as an image (dot-by-dot)
- ▶ but as characters
- ▶ that is, the following

à    **à**    a    **a**    *a*

are all encoded as an 'a'

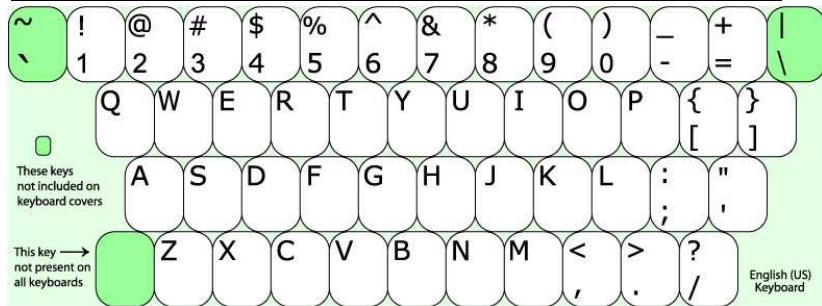
- ▶ the following characters (and most others)

à 3 % ä 1 ç ñ “

are each represented by a number

## encoding characters: ASCII

32– 47	␣	!	"	#	\$	%	&	'	(	)	*	+	,	-	.	/
48– 63	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
64– 79	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
80– 95	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
96–111	'	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
112–126	p	q	r	s	t	u	v	w	x	y	z	{		}	~	



## encoding characters: outside ASCII

character	windows-1250	ibm850	IBM437
ä	228	132	132
ı	—	213	—
ç	231	135	135
ñ	241	164	164
“	147	—	—



## encoding characters: alternatives

character	HTML	T <sub>E</sub> X/L <sub>A</sub> T <sub>E</sub> X
ä	<code>&amp;auml;</code>	<code>\"a</code>
ı	<code>&amp;#305;</code> or <code>&amp;#x131;</code>	<code>\i</code>
ç	<code>&amp;ccedil;</code>	<code>\c c</code>
ñ	<code>&amp;ntilde;</code>	<code>\~n</code>
“	<code>&amp;ldquo;</code>	<code>“</code>

# Unicode

Unicode uses more than a single byte to represent characters (much like digraphs in writing)

# Unicode

Unicode uses more than a single byte to represent characters (much like digraphs in writing)

number of bytes	possibilities
1	256
2	65,536
3	16,777,216
4	4,294,967,296

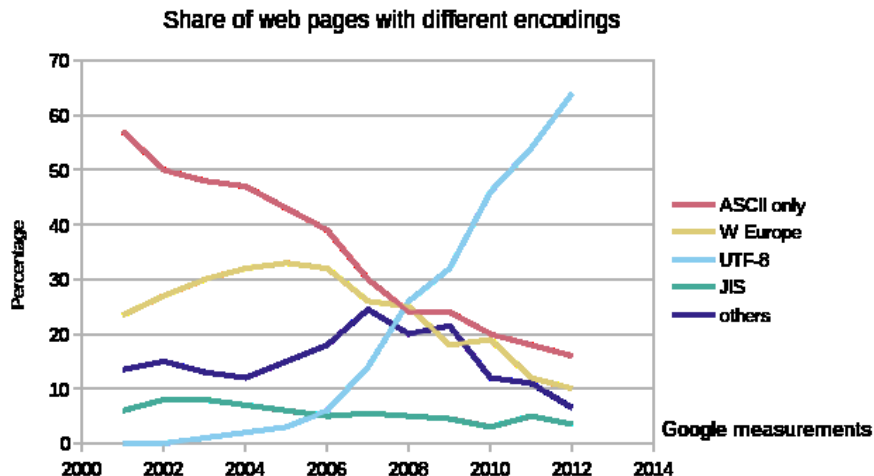
# Unicode

Unicode uses more than a single byte to represent characters (much like digraphs in writing)

number of bytes	possibilities
1	256
2	65,536
3	16,777,216
4	4,294,967,296

character	hexadecimal	decimal	size in bytes
á	0xE1	225	1 byte
Ł	0x141	321	2 bytes
ə	0x259	601	2 bytes
я	0x44F	1103	2 bytes
𐌆	0x30DC	12508	2 bytes
🍷	0x1F60D	128525	3 bytes

# the growth of Unicode



## encoding form

▶ a a a a a

## encoding form

▶ a a a a a

▶ an A is an A is an A...

## encoding form

▶ a a a a a

▶ an A is an A is an A... well, not exactly



## encoding form

▶ a a a a a

- ▶ an A is an A is an A... well, not exactly
- ▶ how to mark font types?

# markup

▶ sample text:

This example contains *italics* and **boldface**, as well as a formula:  
 $2^3 < 3^2$ .

# markup

- ▶ sample text:

This example contains *italics* and **boldface**, as well as a formula:  
 $2^3 < 3^2$ .

- ▶ marked up versions

# markup

- ▶ sample text:

This example contains *italics* and **boldface**, as well as a formula:  
 $2^3 < 3^2$ .

- ▶ marked up versions

- ▶ HTML

This example contains `<i>italics</i>` and `<b>boldface,</b>` as well as a formula:  
`2<sup>3</sup>&lt;3<sup>2</sup>`.

# markup

- ▶ sample text:

This example contains *italics* and **boldface**, as well as a formula:  
 $2^3 < 3^2$ .

- ▶ marked up versions

- ▶ HTML

This example contains `<i>italics</i>` and `<b>boldface,</b>` as well as a formula:  
`2<sup>3</sup>&lt;3<sup>2</sup>`.

- ▶ T<sub>E</sub>X

This example constains `{\it italics}` and `{\bf boldface,}` as well as a formula: `$2^3<3^2$`.

# markup

- ▶ sample text:

This example contains *italics* and **boldface**, as well as a formula:  
 $2^3 < 3^2$ .

- ▶ marked up versions

- ▶ HTML

This example contains `<i>italics</i>` and `<b>boldface,</b>` as well as a formula:  
`2<sup>3</sup>&lt;3<sup>2</sup>`.

- ▶ T<sub>E</sub>X

This example constains `{\it italics}` and `{\bf boldface,}` as well as a formula: `$2^3<3^2$`.

- ▶ Wiki markup ([http://en.wikipedia.org/wiki/Help:Wiki\\_markup](http://en.wikipedia.org/wiki/Help:Wiki_markup))

This example constains `''italics''` and `'''boldface,'''` as well as a formula: `<math>2^3<3^2</math>`.

## types of markup

- ▶ physical markup describes the appearance of the marked text  
“this bit is in italics, that bit is in boldface”  
e.g., `<i>sample word</i>`, `<b>section title</b>`

## types of markup

- ▶ physical markup describes the appearance of the marked text  
“this bit is in italics, that bit is in boldface”  
e.g., `<i>sample word</i>`, `<b>section title</b>`
- ▶ logical markup describes the function of the marked text  
“this bit is a sample word to be **emphasized**, that bit is a section **heading of level 2**”  
e.g., `<em>sample word</em>`, `<h2>section title</h2>`



## comparison of markup types

### logical

- ▶ depends heavily on later interpretation (esp. in web documents)

### physical

- ▶ firmer control over output

## comparison of markup types

### logical

- ▶ depends heavily on later interpretation (esp. in web documents)
- ▶ interpretation of markup has to be customized

### physical

- ▶ firmer control over output
- ▶ less customization necessary

## comparison of markup types

### logical

- ▶ depends heavily on later interpretation (esp. in web documents)
- ▶ interpretation of markup has to be customized
- ▶ flexible on format: e.g., `\emph{}` produces italics in a roman context, and roman in an italic context

### physical

- ▶ firmer control over output
- ▶ less customization necessary
- ▶ (often) premature stance on format

## comparison of markup types

### logical

- ▶ depends heavily on later interpretation (esp. in web documents)
- ▶ interpretation of markup has to be customized
- ▶ flexible on format: e.g., `\emph{}` produces italics in a roman context, and roman in an italic context
- ▶ style easily modifiable later

### physical

- ▶ firmer control over output
- ▶ less customization necessary
- ▶ (often) premature stance on format
- ▶ style modifiable by extensive replacement of markup

## a chunk of SGML code

```

<szocikk>
  <admin>
    <szerk></szerk>
    <forr>OL</forr>
    <statusz st="nyers">
  </admin>
  <foalak>
    <cszo>quiz</cszo>
    <kiejt>kw&pisc;z</kiejt>
  </foalak>
  <joszt>
    <nytan>
      <szf>n</szf>
    </nytan>
    <gralak>
      <anev>pl</anev>
      <alak>quizzes</alak>
    </gralak>
    <jvalt>
      <jarny>
        <jel>
          <ekv>találós játék</ekv>
          <ekv>rejtvény</ekv>
        </jel>
      </jarny>
    </jvalt>
  </joszt>
  <jvalt>
    <jarny>
      <jel>
        <ekv><min>US</min><min>isk</min>
          <min>biz</min>szóbeli
          <elh> vizsga</elh></ekv>
          <ekv>vizsgáztatás</ekv>
        </jel>
      </jarny>
    </jvalt>
  </szocikk>

```

## the entry for *quiz* printed

**quiz** [kwiz] **I.** *fn tsz* **quizzes** **1. a)** találós játék, rejtvény **b)** rég tréfa, móka, ugratás **2. a)** *US, okt, biz* szóbeli (vizsga), vizsgáztatás **b)** *US, okt, biz* osztálykérdés **3.** fogas/nehéz (vizsgai) kérdés **4.** ~ (**game/programme/show**) vetélkedő **5.** rég furcsa figura, fura szerzet **6.** rég tréfacsináló **II. -zz-** **A. tsi** **1. a)** fogas/nehéz kérdéseket tesz fel [*osztálynak*], vizsgáztat [*osztályt*] **b)** kérdez, faggat, vizsgáztat [*vizsgálót*] **2.** megtréfál, ugrat **3. a)** *GB, rég* kihívóan/kíváncsian/feltűnően/fürkészve néz/bámul/mustrál, szemüvegen/lornyonon át vizsgálgat **b)** *GB, rég* gúnyosan/csúfondárosan néz (vkt, vkre) **B. tni** bolondozik, másokat beugrat

Figure: the printed entry for *quiz* in Ország–Magay’s English–Hungarian dictionary

# why have markup?

- ▶ separation of contents and form

## why have markup?

- ▶ separation of contents and form
- ▶ easily modifiable form



## why have markup?

- ▶ separation of contents and form
- ▶ easily modifiable form
- ▶ the form of a text is for *humans*, but

## why have markup?

- ▶ separation of contents and form
- ▶ easily modifiable form
- ▶ the form of a text is for *humans*, but
- ▶ electronic text is not read only by humans, but also by machines (e.g., search engines, for blind people)

# WYSIWYG

- ▶ “What You See Is What You Get”

# WYSIWYG

- ▶ “What You See Is What You Get”
- ▶ e.g., Microsoft Word, Open Office Writer (now called Libre Office)

# WYSIWYG

- ▶ “What You See Is What You Get”
- ▶ e.g., Microsoft Word, Open Office Writer (now called Libre Office)
- ▶ the screen shows (more or less) what comes out of the printer

# WYSIWYG

- ▶ “What You See Is What You Get”
- ▶ e.g., Microsoft Word, Open Office Writer (now called Libre Office)
- ▶ the screen shows (more or less) what comes out of the printer
- ▶ BUT

# WYSIWYG

- ▶ “What You See Is What You Get”
- ▶ e.g., Microsoft Word, Open Office Writer (now called Libre Office)
- ▶ the screen shows (more or less) what comes out of the printer
- ▶ **BUT**
  - ▶ our eyes are not that perfect

# WYSIWYG

- ▶ “What You See Is What You Get”
- ▶ e.g., Microsoft Word, Open Office Writer (now called Libre Office)
- ▶ the screen shows (more or less) what comes out of the printer
- ▶ **BUT**
  - ▶ our eyes are not that perfect
  - ▶ we don't want to print all our documents!



## properties of texts

- ▶ inherent properties of text

## properties of texts

- ▶ inherent properties of text
  - ▶ characters

## properties of texts

- ▶ inherent properties of text
  - ▶ characters
  - ▶ potential breaking points (to be discussed next week)

## properties of texts

- ▶ inherent properties of text
  - ▶ characters
  - ▶ potential breaking points (to be discussed next week)
  - ▶ emphasis

## properties of texts

- ▶ inherent properties of text
  - ▶ characters
  - ▶ potential breaking points (to be discussed next week)
  - ▶ emphasis
  - ▶ structure (sections, subsections)

## properties of texts

- ▶ inherent properties of text
  - ▶ characters
  - ▶ potential breaking points (to be discussed next week)
  - ▶ emphasis
  - ▶ structure (sections, subsections)
- ▶ noninherent properties of text

## properties of texts

- ▶ inherent properties of text
  - ▶ characters
  - ▶ potential breaking points (to be discussed next week)
  - ▶ emphasis
  - ▶ structure (sections, subsections)
- ▶ noninherent properties of text
  - ▶ paper size, margin widths

## properties of texts

- ▶ inherent properties of text
  - ▶ characters
  - ▶ potential breaking points (to be discussed next week)
  - ▶ emphasis
  - ▶ structure (sections, subsections)
- ▶ noninherent properties of text
  - ▶ paper size, margin widths
  - ▶ actual breaking points



## properties of texts

- ▶ inherent properties of text
  - ▶ characters
  - ▶ potential breaking points (to be discussed next week)
  - ▶ emphasis
  - ▶ structure (sections, subsections)
- ▶ noninherent properties of text
  - ▶ paper size, margin widths
  - ▶ actual breaking points
  - ▶ paragraph shape

## properties of texts

- ▶ inherent properties of text
  - ▶ characters
  - ▶ potential breaking points (to be discussed next week)
  - ▶ emphasis
  - ▶ structure (sections, subsections)
- ▶ noninherent properties of text
  - ▶ paper size, margin widths
  - ▶ actual breaking points
  - ▶ paragraph shape
  - ▶ font properties (family, size)

## properties of texts

- ▶ inherent properties of text
  - ▶ characters
  - ▶ potential breaking points (to be discussed next week)
  - ▶ emphasis
  - ▶ structure (sections, subsections)
- ▶ noninherent properties of text
  - ▶ paper size, margin widths
  - ▶ actual breaking points
  - ▶ paragraph shape
  - ▶ font properties (family, size)
- ▶ yet WYSIWYG technology forces decisions in the case of the latter items, too

# comparison of markup and WYSIWYG

## markup

- ▶ daunting at first sight

## WYSIWYG

- ▶ intuitive, easy at first sight

# comparison of markup and WYSIWYG

## markup

- ▶ daunting at first sight
- ▶ powerful

## WYSIWYG

- ▶ intuitive, easy at first sight
- ▶ “what you see is **all** you get”

# comparison of markup and WYSIWYG

## markup

- ▶ daunting at first sight
- ▶ powerful
- ▶ persuades user more effectively to use logical markup

## WYSIWYG

- ▶ intuitive, easy at first sight
- ▶ “what you see is **all** you get”
- ▶ allows user to use primitive formatting techniques

# a horrendous example

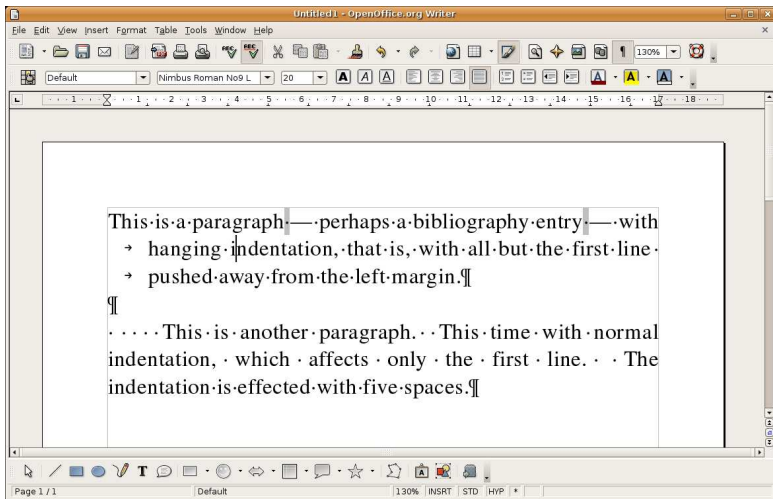


Figure: hanging and normal indentation: never do it this way!

# comparison of markup and WYSIWYG

## markup

- ▶ daunting at first sight
- ▶ powerful
- ▶ persuades user more effectively to use logical markup
- ▶ uses less computer resources

## WYSIWYG

- ▶ intuitive, easy at first sight
- ▶ “what you see is **all** you get”
- ▶ allows user to use primitive formatting techniques
- ▶ uses huge computer resources



# comparison of markup and WYSIWYG

## markup

- ▶ daunting at first sight
- ▶ powerful
- ▶ persuades user more effectively to use logical markup
- ▶ uses less computer resources
- ▶ user sees everything in the file

## WYSIWYG

- ▶ intuitive, easy at first sight
- ▶ “what you see is **all** you get”
- ▶ allows user to use primitive formatting techniques
- ▶ uses huge computer resources
- ▶ data in the file are hidden from user

# comparison of markup and WYSIWYG

## markup

- ▶ daunting at first sight
- ▶ powerful
- ▶ persuades user more effectively to use logical markup
- ▶ uses less computer resources
- ▶ user sees everything in the file

## WYSIWYG

- ▶ intuitive, easy at first sight
- ▶ “what you see is **all** you get”
- ▶ allows user to use primitive formatting techniques
- ▶ uses huge computer resources
- ▶ data in the file are hidden from user