

# IB047

## Úvod do korpusové lingvistiky a počítačové lexikografie

Pavel Rychlý

pary@fi.muni.cz

25. února 2022

# Formáty korpusů

archiv/kolekce různé formáty, podle zdroje/typu

*Oxford Text Archive*

textové banky jednotný formát a základní struktura

dokumenty/texty, základní metadata

*Project Gutenberg*

vertikální text

binární data v aplikaci pomocné data pro rychlejší zpracování

- indexy
- statistiky

## soubory/adresáře

- dokumenty/texty
- 1:1 (soubor  $\approx$  dokument)
- 1:n (soubor  $\approx$  n dokumentů)
- n:1 (n souborů  $\approx$  dokument)  
značkování, statistiky, ...
- pro hodně velké korpusy každý soubor 100 MB
- doplňková data samostatně (zvuky k textu)

## Co je v korpusu uloženo?

- text
- metadata
  - autor, rok publikace, pohlaví cílové skupiny
- struktura dokumentu
  - odstavce, nadpisy, verše, věty
- značkování
  - informace o slovech
  - morfologie, základní tvary

- 8 bitů 256 znaků
  - ASCII – základ 7 bitů
  - kódování pro češtinu
    - ISO-Latin-2, Windows-1250, 852
- Unicode
  - Unicode 13 (2021)
  - 31bitů na znak, kódy zatím jen do 0xE01EF (0x10FFFF)
  - asi 130 tisíc znaků (2600 emoji)
    - klasifikace (číslíce, znaky), převody, lokalizace
  - UTF-8
    - 1 až 4 bytů na znak
  - UTF-16
    - 2 až 4 byty na znak
    - Byte Order Mark, koplíkové

- kompatibilita s ASCII
- jednotné na rôznych platformách (Little/Big Endian)
- snadno zistíme kde začína znak

Bits	Last code point	Byte 1	Byte 2	Byte 3	Byte 4	Byte 5	Byte 6
7	U+007F	0xxxxxxx					
11	U+07FF	110xxxxx	10xxxxxx				
16	U+FFFF	1110xxxx	10xxxxxx	10xxxxxx			
21	U+1FFFFF	11110xxx	10xxxxxx	10xxxxxx	10xxxxxx		
26	U+3FFFFFFF	111110xx	10xxxxxx	10xxxxxx	10xxxxxx	10xxxxxx	
31	U+7FFFFFFF	1111110x	10xxxxxx	10xxxxxx	10xxxxxx	10xxxxxx	10xxxxxx

- 'a': U+0061, UTF-8: 61 \x61      \u0061
- 'à': U+00E0, UTF-8: c3 a0 \xc3\xa0  
combining character: ' U+0300  
a + ' – 61 cc 80 \x61\xcc\x80

- escape-sekvence
  - speciální znak mění význam následujících znaků
    - `\n`, `\t`, `&amp;`; `<tag>`
- SGML
  - Standard Generalised Markup Language
  - ISO 8879:1986(E)
- XML
  - Extensible Markup Language
  - W3C, 1998

- struktura popsána v DTD
- elementy
  - počáteční, koncová značka
  - `<doc>`, `<head>`, `</head>`, `<g/>`
- atributy elementů/značek
  - `<doc title="Jak pejsek ..." author="Čapek">`
  - `<head type="main">`
- entity
  - `&gt;`; `&lt;`; `&amp;`; `&eacute;`;



- SGML/XML
- TEI
  - Text Encoding Initiative
  - TEI Guidelines for Electronic Text Encoding and Interchange
  - 3. verze (TEI P3), 1993, 39 kapitol
  - 23. kapitola – Language Corpora
  - 4. verze (TEI P4), 2001–2004, podpora XML
  - aktuálně – TEI P5 – 2007, více XML (vnoření jiných sad: MathML), kontroly
  - 15. kapitola – Language Corpora
- CES, XCES
  - Corpus Encoding Standard
  - XCES 1.0.4. (2008) – odpovídá TEI P5
- definují sadu elementů a atributů pro strukturu a metadata

## Rozdělení textu do pozic

- token (pozice) = základní prvek korpusu
- většinou slovo, číslo, interpunkce
- může silně ovlivnit výsledky

## Příklady:

- |         |             |
|---------|-------------|
| bude-li | ■ bude-li   |
|         | ■ bude -li  |
|         | ■ bude - li |
| don't   | ■ don't     |
|         | ■ don ' t   |
|         | ■ do n't    |

## Výceslovné výrazy

- samostatné tokeny + struktury

```
<phr word="ping pong">  
ping  
pong  
</phr>
```

- jeden token + vícehodnotový atribut pro vyhledávání

```
ping pong<TAB>ping,pong
```

- jednoduchý formát i jeho zpracování
  - každý token na samostatném řádku
  - struktury formou XML elementů
  - značkování odděleno tabulátorem
- podrobnosti
  - <http://nlp.fi.muni.cz/>
  - Informace pro současné a potenciální spolupracovníky
  - Textové korpusy
  - Popis vertikálů

# PRE-vertikální text

- pre-vertikální = “před” vertikálním textem
- struktury jako ve vertikálním textu:  
každá značka (otevírací, zavírací) na samostatném řádku
- `<doc>` dokument, `<p>` odstavec, (automaticky `<s>` věta)
- text není tokenizován – každý odstavec na samostatném řádku

```
<doc genre="fiction" title="1984" author="G. Orwell">
<chapter no="1.1">
<p>
It was a bright cold day in April, and the clocks were striking thirteen. Winston Smith, his
</p>
<p>
The hallway smelt of boiled cabbage and old rag mats. At one end of it a coloured poster, too
</p>
</chapter>
</doc>
```

- kolekce nástrojů na `http://corpus.tools`
- Chared – detekce kódování dle jazyka
- uninorm – normalizace kódování a znaků (uvozovky, pomlčky)
- unitok – tokenizace dle jazyka