

Korpusy textů a jejich využití

Pavel Rychlý, Aleš Horák

E-mail: hales@fi.muni.cz

http://nlp.fi.muni.cz/poc_lingv/

Obsah:

- Co to je korpus?
- Anglické a národní korpusy
- Formáty korpusů
- Korpusové manažery

Co to je korpus?

- Co to je text, dokument?

- lecos

- Různé typy korpusů

- textové
 - mluvené

- Pro potřeby NLP

- textový korpus

Co to je korpus?

- Co to je text, dokument?

- lecos

- Různé typy korpusů

- textové
 - mluvené

- Pro potřeby NLP

- textový korpus

Co to je korpus?

- Co to je text, dokument?
 - lecos
- Různé typy korpusů
 - textové
 - mluvené
- Pro potřeby NLP
 - textový korpus

Co to je korpus?

- Co to je text, dokument?
 - lecos
- Různé typy korpusů
 - textové
 - mluvené
- Pro potřeby NLP
 - textový korpus

Co to je korpus?

- Co to je text, dokument?
 - lecos
- Různé typy korpusů
 - textové
 - mluvené
- Pro potřeby NLP
 - textový korpus

Co to je korpus?

- Co to je text, dokument?
 - lecos
- Různé typy korpusů
 - textové
 - mluvené
- Pro potřeby NLP
 - textový korpus

Co to je korpus?

- Co to je text, dokument?
 - lecos
- Různé typy korpusů
 - textové
 - mluvené
- Pro potřeby NLP
 - textový korpus

Textový korpus

- soubor textů
- charakteristiky
 - rozsáhlý (stovky mil. až mld. pozic/slov)
 - v jednotném formátu
 - strukturovaný
 - v elektronické podobě

Textový korpus

- soubor textů
- charakteristiky
 - rozsáhlý (stovky mil. až mld. pozic/slov)
 - v jednotném formátu
 - stukturovaný
 - v elektronické podobě

Textový korpus

- soubor textů
- charakteristiky
 - rozsáhlý (stovky mil. až mld. pozic/slov)
 - v jednotném formátu
 - stukturowaný
 - v elektronické podobě

Textový korpus

- soubor textů
- charakteristiky
 - rozsáhlý (stovky mil. až mld. pozic/slov)
 - v jednotném formátu
 - stukturovaný
 - v elektronické podobě

Textový korpus

- soubor textů
- charakteristiky
 - rozsáhlý (stovky mil. až mld. pozic/slov)
 - v jednotném formátu
 - stukturovaný
 - v elektronické podobě

Textový korpus

- soubor textů
- charakteristiky
 - rozsáhlý (stovky mil. až mld. pozic/slov)
 - v jednotném formátu
 - stukturovaný
 - v elektronické podobě

Typy korpusů

- vždy záleží na účelu a způsobu použití
- možnosti
 - jazyk
 - typy textů
 - zdroj dat
 - značkování
 - ...

Typy korpusů

- vždy záleží na účelu a způsobu použití
- možnosti
 - jazyk
 - typy textů
 - zdroj dat
 - značkování
 - ...

Typy korpusů

- vždy záleží na účelu a způsobu použití
- možnosti
 - jazyk
 - typy textů
 - zdroj dat
 - značkování
 - ...

Typy korpusů

- vždy záleží na účelu a způsobu použití
- možnosti
 - jazyk
 - typy textů
 - zdroj dat
 - značkování
 - ...

Typy korpusů

- vždy záleží na účelu a způsobu použití
- možnosti
 - jazyk
 - typy textů
 - zdroj dat
 - značkování
 - ...

Typy korpusů

- vždy záleží na účelu a způsobu použití
- možnosti
 - jazyk
 - typy textů
 - zdroj dat
 - značkování

... .

Typy korpusů

- vždy záleží na účelu a způsobu použití
- možnosti
 - jazyk
 - typy textů
 - zdroj dat
 - značkování
 - ...

První korpus

Brown

- americká angličtina (1961)
- Brown University, 1964
- gramatické značkování, 1979
- 500 textů, 1 mil. slov
- W. N. Francis & H. Kučera
 - * první statistické charakteristiky angličtiny
 - * relativní četnosti slov a slovních druhů

První korpus

Brown

- americká angličtina (1961)
- Brown University, 1964
- gramatické značkování, 1979
- 500 textů, 1 mil. slov
- W. N. Francis & H. Kučera
 - * první statistické charakteristiky angličtiny
 - * relativní četnosti slov a slovních druhů

První korpus

Brown

- americká angličtina (1961)
- Brown University, 1964
- gramatické značkování, 1979
- 500 textů, 1 mil. slov
- W. N. Francis & H. Kučera
 - * první statistické charakteristiky angličtiny
 - * relativní četnosti slov a slovních druhů

První korpus

Brown

- americká angličtina (1961)
- Brown University, 1964
- gramatické značkování, 1979
- 500 textů, 1 mil. slov
- W. N. Francis & H. Kučera
 - * první statistické charakteristiky angličtiny
 - * relativní četnosti slov a slovních druhů

První korpus

Brown

- americká angličtina (1961)
- Brown University, 1964
- gramatické značkování, 1979
- 500 textů, 1 mil. slov
- W. N. Francis & H. Kučera
 - první statistické charakteristiky angličtiny
 - relativní četnosti slov a slovních druhů

První korpus

Brown

- americká angličtina (1961)
- Brown University, 1964
- gramatické značkování, 1979
- 500 textů, 1 mil. slov
- W. N. Francis & H. Kučera
 - první statistické charakteristiky angličtiny
 - relativní četnosti slov a slovních druhů

První korpus

Brown

- americká angličtina (1961)
- Brown University, 1964
- gramatické značkování, 1979
- 500 textů, 1 mil. slov
- W. N. Francis & H. Kučera
 - první statistické charakteristiky angličtiny
 - relativní četnosti slov a slovních druhů

SUSANNE

SUSANNE

- autor Geoffrey Sampson, Sussex University
- kniha *English for the Computer*
- část korpusu Brown ($\frac{1}{4}$)
- nové gramatické značkování
- syntaktické značkování

SUSANNE

SUSANNE

- autor Geoffrey Sampson, Sussex University
- kniha *English for the Computer*
- část korpusu Brown ($\frac{1}{4}$)
- nové gramatické značkování
- syntaktické značkování

SUSANNE

SUSANNE

- autor Geoffrey Sampson, Sussex University
- kniha *English for the Computer*
- část korpusu Brown ($\frac{1}{4}$)
- nové gramatické značkování
- syntaktické značkování

SUSANNE

SUSANNE

- autor Geoffrey Sampson, Sussex University
- kniha *English for the Computer*
- část korpusu Brown ($\frac{1}{4}$)
- nové gramatické značkování
- syntaktické značkování

SUSANNE

SUSANNE

- autor Geoffrey Sampson, Sussex University
- kniha *English for the Computer*
- část korpusu Brown ($\frac{1}{4}$)
- nové gramatické značkování
- syntaktické značkování

British National Corpus

- britská angličtina, 10% mluva
- první velký korpus pro lexikografy
- vydavatelé slovníků (OUP) + univerzity
- 1991–1994, World Edition 2000
- ≈3000 textů, 100 mil. slov
- gramatické značkování automatickým nástrojem

British National Corpus

- britská angličtina, 10% mluva
- první velký korpus pro lexikografy
- vydavatelé slovníků (OUP) + univerzity
- 1991–1994, World Edition 2000
- ≈3000 textů, 100 mil. slov
- gramatické značkování automatickým nástrojem

British National Corpus

- britská angličtina, 10% mluva
- první velký korpus pro lexikografy
- vydavatelé slovníků (OUP) + univerzity
- 1991–1994, World Edition 2000
- ≈3000 textů, 100 mil. slov
- gramatické značkování automatickým nástrojem

British National Corpus

- britská angličtina, 10% mluva
- první velký korpus pro lexikografy
- vydavatelé slovníků (OUP) + univerzity
- 1991–1994, World Edition 2000
- ≈3000 textů, 100 mil. slov
- gramatické značkování automatickým nástrojem

British National Corpus

- britská angličtina, 10% mluva
- první velký korpus pro lexikografy
- vydavatelé slovníků (OUP) + univerzity
- 1991–1994, World Edition 2000
- ≈3000 textů, 100 mil. slov
- gramatické značkování automatickým nástrojem

British National Corpus

- britská angličtina, 10% mluva
- první velký korpus pro lexikografy
- vydavatelé slovníků (OUP) + univerzity
- 1991–1994, World Edition 2000
- ≈3000 textů, 100 mil. slov
- gramatické značkování automatickým nástrojem

Bank of English

- britská angličtina
- COBUILD (HarperCollins), University of Birmingham
- 1991, stále rozšiřován
- 2002, \approx 450 mil. slov

Bank of English

- britská angličtina
- COBUILD (HarperCollins), University of Birmingham
- 1991, stále rozšiřován
- 2002, \approx 450 mil. slov

Bank of English

- britská angličtina
- COBUILD (HarperCollins), University of Birmingham
- 1991, stále rozšiřován
- 2002, \approx 450 mil. slov

Bank of English

- britská angličtina
- COBUILD (HarperCollins), University of Birmingham
- 1991, stále rozšiřován
- 2002, \approx 450 mil. slov

Další národní korpusy

- Český národní korpus
 - ÚČNK, FF UK
 - SYN2000: 100 mil. slov
 - Litera, Synek, BMK, ...
- Slovenský, Maďarský, Chorvatský, ...
- Americký

Další národní korpusy

- Český národní korpus
 - ÚČNK, FF UK
 - SYN2000: 100 mil. slov
 - Litera, Synek, BMK, ...
- Slovenský, Maďarský, Chorvatský, ...
- Americký

Další národní korpusy

- Český národní korpus
 - ÚČNK, FF UK
 - SYN2000: 100 mil. slov
 - Litera, Synek, BMK, ...
- Slovenský, Maďarský, Chorvatský, ...
- Americký

Další národní korpusy

- Český národní korpus
 - ÚČNK, FF UK
 - SYN2000: 100 mil. slov
 - Litera, Synek, BMK, ...
- Slovenský, Maďarský, Chorvatský, ...
- Americký

Další národní korpusy

- Český národní korpus
 - ÚČNK, FF UK
 - SYN2000: 100 mil. slov
 - Litera, Synek, BMK, ...
- Slovenský, Maďarský, Chorvatský, ...
- Americký

Další národní korpusy

- Český národní korpus
 - ÚČNK, FF UK
 - SYN2000: 100 mil. slov
 - Litera, Synek, BMK, ...
- Slovenský, Maďarský, Chorvatský, ...
- Americký

Korpusy na FI

vytvořené na FI, příklady:

- Desam

- 1996, ručně značkovaný (desambiguovaný)
- ≈1 mil. slov

- WWW

- periodika z webu, z let 1996–1998
- ≈100 mil.

- Chyby

- práce studentů předmětu Základy odb. stylu s vyznačenými chybami
- ≈400 tis.

Korpusy na FI

vytvořené na FI, příklady:

- Desam
 - 1996, ručně značkovaný (desambiguovaný)
 - ≈1 mil. slov
- WWW
 - periodika z webu, z let 1996–1998
 - ≈100 mil.
- Chyby
 - práce studentů předmětu Základy odb. stylu s vyznačenými chybami
 - ≈400 tis.

Korpusy na FI

vytvořené na FI, příklady:

- Desam
 - 1996, ručně značkovaný (desambiguovaný)
 - ≈1 mil. slov
- WWW
 - periodika z webu, z let 1996–1998
 - ≈100 mil.
- Chyby
 - práce studentů předmětu Základy odb. stylu s vyznačenými chybami
 - ≈400 tis.

Korpusy na FI

vytvořené na FI, příklady:

- Desam
 - 1996, ručně značkovaný (desambiguovaný)
 - ≈1 mil. slov
- WWW
 - periodika z webu, z let 1996–1998
 - ≈100 mil.
- Chyby
 - práce studentů předmětu Základy odb. stylu s vyznačenými chybami
 - ≈400 tis.

Korpusy na FI

vytvořené na FI, příklady:

- Desam
 - 1996, ručně značkovaný (desambiguovaný)
 - ≈1 mil. slov
- WWW
 - periodika z webu, z let 1996–1998
 - ≈100 mil.
- Chyby
 - práce studentů předmětu Základy odb. stylu s vyznačenými chybami
 - ≈400 tis.

Korpusy na FI

vytvořené na FI, příklady:

- Desam
 - 1996, ručně značkovaný (desambiguovaný)
 - ≈1 mil. slov
- WWW
 - periodika z webu, z let 1996–1998
 - ≈100 mil.
- Chyby
 - práce studentů předmětu Základy odb. stylu s vyznačenými chybami
 - ≈400 tis.

Korpusy na FI

vytvořené na FI, příklady:

- Desam
 - 1996, ručně značkovaný (desambiguovaný)
 - ≈1 mil. slov
- WWW
 - periodika z webu, z let 1996–1998
 - ≈100 mil.
- Chyby
 - práce studentů předmětu Základy odb. stylu s vyznačenými chybami
 - ≈400 tis.

Korpusy na FI

vytvořené na FI, příklady:

- Desam
 - 1996, ručně značkovaný (desambiguovaný)
 - ≈1 mil. slov
- WWW
 - periodika z webu, z let 1996–1998
 - ≈100 mil.
- Chyby
 - práce studentů předmětu Základy odb. stylu s vyznačenými chybami
 - ≈400 tis.

Korpusy na FI

vytvořené na FI, příklady:

- Desam
 - 1996, ručně značkovaný (desambiguovaný)
 - ≈1 mil. slov
- WWW
 - periodika z webu, z let 1996–1998
 - ≈100 mil.
- Chyby
 - práce studentů předmětu Základy odb. stylu s vyznačenými chybami
 - ≈400 tis.

Korpusy na FI

spolupráce

- Dopisy
- Mluv
- Kačenka
- ČNPK
- 1984
- Otto
- Italian
- Giga Chinese
- Francouzský, Slovinský, Britská angličtina, ...

Korpusy na FI

spolupráce

- Dopisy
- Mluv
- Kačenka
- ČNPK
- 1984
- Otto
- Italian
- Giga Chinese
- Francouzský, Slovinský, Britská angličtina, ...

Korpusy na FI

spolupráce

- Dopisy
- Mluv
- Kačenka
- ČNPK
- 1984
- Otto
- Italian
- Giga Chinese
- Francouzský, Slovinský, Britská angličtina, ...

Korpusy na FI

spolupráce

- Dopisy
- Mluv
- Kačenka
- ČNPK
- 1984
- Otto
- Italian
- Giga Chinese
- Francouzský, Slovinský, Britská angličtina, ...

Korpusy na FI

spolupráce

- Dopisy
- Mluv
- Kačenka
- ČNPK
- 1984
- Otto
- Italian
- Giga Chinese
- Francouzský, Slovinský, Britská angličtina, ...

Korpusy na FI

spolupráce

- Dopisy
- Mluv
- Kačenka
- ČNPK
- 1984
- Otto
- Italian
- Giga Chinese
- Francouzský, Slovinský, Britská angličtina, ...

Korpusy na FI

spolupráce

- Dopisy
- Mluv
- Kačenka
- ČNPK
- 1984
- Otto
- Italian
- Giga Chinese
- Francouzský, Slovinský, Britská angličtina, ...

Korpusy na FI

spolupráce

- Dopisy
- Mluv
- Kačenka
- ČNPK
- 1984
- Otto
- Italian
- Giga Chinese
- Francouzský, Slovinský, Britská angličtina, ...

Korpusy na FI

spolupráce

- Dopisy
- Mluv
- Kačenka
- ČNPK
- 1984
- Otto
- Italian
- Giga Chinese
- Francouzský, Slovinský, Britská angličtina, ...

Formáty korpusů

- archiv/kolekce
 - různé formáty, podle zdroje/typu
- textové banky
 - jednotný formát a základní struktura
 - dokumenty/texty, základní metainformace
- vertikální text
- binární data v aplikaci
 - pomocná data pro rychlejší zpracování

Formáty korpusů

- archiv/kolekce
 - různé formáty, podle zdroje/typu
- textové banky
 - jednotný formát a základní struktura
 - dokumenty/texty, základní metainformace
- vertikální text
- binární data v aplikaci
 - pomocná data pro rychlejší zpracování

PDF

DOCX

CSV

XML

JSON

DB

DBF

DB2

DB3

DB4

DB5

DB6

DB7

DB8

DB9

DB10

DB11

DB12

DB13

DB14

DB15

DB16

DB17

DB18

DB19

DB20

DB21

DB22

DB23

DB24

DB25

DB26

DB27

DB28

DB29

DB30

DB31

DB32

DB33

DB34

DB35

DB36

DB37

DB38

DB39

DB40

DB41

DB42

DB43

DB44

DB45

DB46

DB47

DB48

DB49

DB50

DB51

DB52

DB53

DB54

DB55

DB56

DB57

DB58

DB59

DB60

DB61

DB62

DB63

DB64

DB65

DB66

DB67

DB68

DB69

DB70

DB71

DB72

DB73

DB74

DB75

DB76

DB77

DB78

DB79

DB80

DB81

DB82

DB83

DB84

DB85

DB86

DB87

DB88

DB89

DB90

DB91

DB92

DB93

DB94

DB95

DB96

DB97

DB98

DB99

DB100

DB101

DB102

DB103

DB104

DB105

DB106

DB107

DB108

DB109

DB110

DB111

DB112

DB113

DB114

DB115

DB116

DB117

DB118

DB119

DB120

DB121

DB122

DB123

DB124

DB125

DB126

DB127

DB128

DB129

DB130

DB131

DB132

DB133

DB134

DB135

DB136

DB137

DB138

DB139

DB140

DB141

DB142

DB143

DB144

DB145

DB146

DB147

DB148

DB149

DB150

DB151

DB152

DB153

DB154

DB155

DB156

DB157

DB158

DB159

DB160

DB161

DB162

DB163

DB164

DB165

DB166

DB167

DB168

DB169

DB170

DB171

DB172

DB173

DB174

DB175

DB176

DB177

DB178

DB179

DB180

DB181

DB182

DB183

DB184

DB185

DB186

DB187

DB188

DB189

DB190

DB191

DB192

DB193

DB194

DB195

DB196

DB197

DB198

DB199

DB200

DB201

DB202

DB203

DB204

DB205

DB206

DB207

DB208

DB209

DB210

DB211

DB212

DB213

DB214

DB215

DB216

DB217

DB218

DB219

DB220

DB221

DB222

DB223

DB224

DB225

DB226

DB227

DB228

DB229

DB230

DB231

DB232

DB233

DB234

DB235

DB236

DB237

DB238

DB239

DB240

DB241

DB242

DB243

DB244

DB245

DB246

DB247

DB248

DB249

DB250

DB251

DB252

DB253

DB254

DB255

DB256

DB257

DB258

DB259

DB260

DB261

DB262

DB263

DB264

DB265

DB266

DB267

DB268

DB269

DB270

DB271

DB272

DB273

DB274

DB275

DB276

DB277

DB278

DB279

DB280

DB281

DB282

DB283

DB284

DB285

DB286

DB287

DB288

DB289

DB290

DB291

DB292

DB293

DB294

DB295

DB296

DB297

DB298

DB299

DB300

Formáty korpusů

- archiv/kolekce
 - různé formáty, podle zdroje/typu
- textové banky
 - jednotný formát a základní struktura
 - dokumenty/texty, základní metainformace
- vertikální text
- binární data v aplikaci
 - pomocná data pro rychlejší zpracování

Formáty korpusů

- archiv/kolekce
 - různé formáty, podle zdroje/typu
- textové banky
 - jednotný formát a základní struktura
 - dokumenty/texty, základní metainformace
- vertikální text
- binární data v aplikaci
 - pomocná data pro rychlejší zpracování

Formáty korpusů

- archiv/kolekce
 - různé formáty, podle zdroje/typu
- textové banky
 - jednotný formát a základní struktura
 - dokumenty/texty, základní metainformace
- vertikální text
- binární data v aplikaci
 - pomocná data pro rychlejší zpracování

Formáty korpusů

- archiv/kolekce
 - různé formáty, podle zdroje/typu
- textové banky
 - jednotný formát a základní struktura
 - dokumenty/texty, základní metainformace
- vertikální text
- binární data v aplikaci
 - pomocná data pro rychlejší zpracování

Formáty korpusů

- archiv/kolekce
 - různé formáty, podle zdroje/typu
- textové banky
 - jednotný formát a základní struktura
 - dokumenty/texty, základní metainformace
- vertikální text
- binární data v aplikaci
 - pomocná data pro rychlejší zpracování
 - indexy
 - statistiky

Formáty korpusů

- archiv/kolekce
 - různé formáty, podle zdroje/typu
- textové banky
 - jednotný formát a základní struktura
 - dokumenty/texty, základní metainformace
- vertikální text
- binární data v aplikaci
 - pomocná data pro rychlejší zpracování
 - indexy
 - statistiky

Formáty korpusů

- archiv/kolekce
 - různé formáty, podle zdroje/typu
- textové banky
 - jednotný formát a základní struktura
 - dokumenty/texty, základní metainformace
- vertikální text
- binární data v aplikaci
 - pomocná data pro rychlejší zpracování
 - indexy
 - statistiky

Formáty korpusů

- archiv/kolekce
 - různé formáty, podle zdroje/typu
- textové banky
 - jednotný formát a základní struktura
 - dokumenty/texty, základní metainformace
- vertikální text
- binární data v aplikaci
 - pomocná data pro rychlejší zpracování
 - indexy
 - statistiky

Kódování znaků

- 8 bitů \approx 256 znaků

- ASCII – základ 7 bitů
- kódování pro češtinu
 - ISO-Latin-2, Windows-1250, 852

- Unicode

- 32bitů na znak
- UTF-8
 - ISO-10646, Unicode
- UTF-16
 - ISO-10646, Unicode

Kódování znaků

- 8 bitů \approx 256 znaků
 - ASCII – základ 7 bitů
 - kódování pro češtinu
 - ISO-Latin-2, Windows-1250, 852
- Unicode
 - 32bitů na znak
 - UTF-8
 - ISO-10646, Unicode
 - UTF-16
 - ISO-10646, Unicode

Kódování znaků

- 8 bitů \approx 256 znaků
 - ASCII – základ 7 bitů
 - kódování pro češtinu
 - ISO-Latin-2, Windows-1250, 852

- Unicode
 - 32bitů na znak
 - UTF-8
 - ISO-10646, Unicode
 - UTF-16
 - ISO-10646, Unicode

Kódování znaků

- 8 bitů \approx 256 znaků
 - ASCII – základ 7 bitů
 - kódování pro češtinu
 - ISO-Latin-2, Windows-1250, 852

• Unicode

- 32bitů na znak

- UTF-8

- 16bitů na znak

- UTF-16

- 32bitů na znak

Kódování znaků

- 8 bitů \approx 256 znaků
 - ASCII – základ 7 bitů
 - kódování pro češtinu
 - ISO-Latin-2, Windows-1250, 852
- Unicode
 - 32bitů na znak
 - UTF-8
 - 1 až 4 byty na znak
 - UTF-16
 - 2 až 4 byty na znak

Kódování znaků

- 8 bitů \approx 256 znaků
 - ASCII – základ 7 bitů
 - kódování pro češtinu
 - ISO-Latin-2, Windows-1250, 852
- Unicode
 - 32bitů na znak
 - UTF-8
 - 1 až 4 byty na znak
 - UTF-16
 - 2 až 4 byty na znak

Kódování znaků

- 8 bitů \approx 256 znaků
 - ASCII – základ 7 bitů
 - kódování pro češtinu
 - ISO-Latin-2, Windows-1250, 852
- Unicode
 - 32bitů na znak
 - UTF-8
 - 1 až 4 byty na znak
 - UTF-16
 - 2 až 4 byty na znak

Kódování znaků

- 8 bitů \approx 256 znaků
 - ASCII – základ 7 bitů
 - kódování pro češtinu
 - ISO-Latin-2, Windows-1250, 852
- Unicode
 - 32bitů na znak
 - UTF-8
 - 1 až 4 byty na znak
 - UTF-16
 - 2 až 4 byty na znak

Kódování znaků

- 8 bitů \approx 256 znaků
 - ASCII – základ 7 bitů
 - kódování pro češtinu
 - ISO-Latin-2, Windows-1250, 852
- Unicode
 - 32bitů na znak
 - UTF-8
 - 1 až 4 byty na znak
 - UTF-16
 - 2 až 4 byty na znak

Kódování znaků

- 8 bitů \approx 256 znaků
 - ASCII – základ 7 bitů
 - kódování pro češtinu
 - ISO-Latin-2, Windows-1250, 852
- Unicode
 - 32bitů na znak
 - UTF-8
 - 1 až 4 byty na znak
 - UTF-16
 - 2 až 4 byty na znak

Kódování metainformací

- escape-sekvence

- speciální znak mění význam následujících znaků
 - \n, \t, & , <tag>

- SGML

- Standard Generalised Markup Language
- ISO 8879:1986(E)

- XML

- Extensible Markup Language
- W3C, 1998

Kódování metainformací

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

Kódování metainformací

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

Kódování metainformací

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

Kódování metainformací

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

Kódování metainformací

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

Kódování metainformací

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

Kódování metainformací

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

Kódování metainformací

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

XML

- 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
 - * >;, <;, &;, ´;

XML

- 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
 - >;, <;, &;, ´;

XML

- 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
 - >;, <;, &;, ´;

XML

- 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
 - >;, <;, &;, ´;

XML

- 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
 - >; <; &; ´;

XML

- 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
 - >; <; &; ´;

XML

- 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
 - >; <; &; ´;

XML

- 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
 - >, <, &, ´

XML

- 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
 - >, <, &, é

Standardy pro ukládání textů

- SGML/XML
- TEI
 - Text Encoding Initiative (1994)
 - TEI Guidelines for Electronic Text Encoding and Interchange
- CES, XCES
 - Corpus Encoding Standard

Standardy pro ukládání textů

- SGML/XML
- TEI
 - Text Encoding Initiative (1994)
 - TEI Guidelines for Electronic Text Encoding and Interchange
- CES, XCES
 - Corpus Encoding Standard

Standardy pro ukládání textů

- SGML/XML
- TEI
 - Text Encoding Initiative (1994)
 - TEI Guidelines for Electronic Text Encoding and Interchange
- CES, XCES
 - Corpus Encoding Standard

Standardy pro ukládání textů

- SGML/XML
- TEI
 - Text Encoding Initiative (1994)
 - TEI Guidelines for Electronic Text Encoding and Interchange
- CES, XCES
 - Corpus Encoding Standard

Standardy pro ukládání textů

- SGML/XML
- TEI
 - Text Encoding Initiative (1994)
 - TEI Guidelines for Electronic Text Encoding and Interchange
- CES, XCES
 - Corpus Encoding Standard

Standardy pro ukládání textů

- SGML/XML
- TEI
 - Text Encoding Initiative (1994)
 - TEI Guidelines for Electronic Text Encoding and Interchange
- CES, XCES
 - Corpus Encoding Standard

Obsah korpusu

Co je v korpusu uloženo?

- text
- metainformace
- struktura dokumentu
 - odstavce, nadpisy, verše, věty
- značkování
 - informace o slovech/pozicích
 - morfologie, základní tvary, syntaktické vazby, ...

Obsah korpusu

Co je v korpusu uloženo?

- text
- metainformace
- struktura dokumentu
 - odstavce, nadpisy, verše, věty
- značkování
 - informace o slovech/pozicích
 - morfologie, základní tvary, syntaktické vazby, ...

Obsah korpusu

Co je v korpusu uloženo?

- text
- metainformace
- struktura dokumentu
 - odstavce, nadpisy, verše, věty
- značkování
 - informace o slovech/pozicích
 - morfologie, základní tvary, syntaktické vazby, ...

Obsah korpusu

Co je v korpusu uloženo?

- text
- metainformace
- struktura dokumentu
 - odstavce, nadpisy, verše, věty
- značkování
 - informace o slovech/pozicích
 - morfologie, základní tvary, syntaktické vazby, ...

Obsah korpusu

Co je v korpusu uloženo?

- text
- metainformace
- struktura dokumentu
 - odstavce, nadpisy, verše, věty
- značkování
 - informace o slovech/pozicích
 - morfologie, základní tvary, syntaktické vazby, ...

Obsah korpusu

Co je v korpusu uloženo?

- text
- metainformace
- struktura dokumentu
 - odstavce, nadpisy, verše, věty
- značkování
 - informace o slovech/pozicích
 - morfologie, základní tvary, syntaktické vazby, ...

Obsah korpusu

Co je v korpusu uloženo?

- text
- metainformace
- struktura dokumentu
 - odstavce, nadpisy, verše, věty
- značkování
 - informace o slovech/pozicích
 - morfologie, základní tvary, syntaktické vazby, ...

Tokenizace

Rozdělení textu do pozic

- může silně ovlivnit výsledky dotazování, četnosti i značkování
- token (pozice) = základní prvek korpusu
- většinou slovo, číslo, interpunkce
 - * bude-li, don't – 4 možnosti:
 1. [don't]
 2. [don] | 't|
 3. [don] | '| | t|
 4. [do] | n't| – v BNC
 - * zkratky (s tečkama?)
 - * datumy
 - * desetinná čísla, ...

Tokenizace

Rozdělení textu do pozic

- může silně ovlivnit výsledky dotazování, četnosti i značkování
- token (pozice) = základní prvek korpusu
- většinou slovo, číslo, interpunkce
 - * bude-li, don't – 4 možnosti:
 1. [don't]
 2. [don] [t̚]
 3. [don] ['] [t̚]
 4. [do] [n't] – v BNC
 - * zkratky (s tečkama?)
 - * datumy
 - * desetinná čísla, ...

Tokenizace

Rozdělení textu do pozic

- může silně ovlivnit výsledky dotazování, četnosti i značkování
- token (pozice) = základní prvek korpusu
- většinou slovo, číslo, interpunkce
 - *bude-li, don't* – 4 možnosti:
 1. |don't|
 2. |don| '|t|
 3. |don| '|' |t|
 4. |do| |n't| – v BNC
 - zkratky (s tečkama?)
 - datumy
 - desetinná čísla, ...

Tokenizace

Rozdělení textu do pozic

- může silně ovlivnit výsledky dotazování, četnosti i značkování
- token (pozice) = základní prvek korpusu
- většinou slovo, číslo, interpunkce
 - **bude-li, don't** – 4 možnosti:
 1. |don't|
 2. |don| '|t|
 3. |don| '|' |t|
 4. |do| |n't| – v BNC
 - zkratky (s tečkama?)
 - datumy
 - desetinná čísla, ...

Tokenizace

Rozdělení textu do pozic

- může silně ovlivnit výsledky dotazování, četnosti i značkování
- token (pozice) = základní prvek korpusu
- většinou slovo, číslo, interpunkce
 - **bude-li, don't** – 4 možnosti:
 1. |don't|
 2. |don| '|t|
 3. |don| '|' |t|
 4. |do| |n't| – v BNC
 - zkratky (s tečkama?)
 - datumy
 - desetinná čísla, ...

Tokenizace

Rozdělení textu do pozic

- může silně ovlivnit výsledky dotazování, četnosti i značkování
- token (pozice) = základní prvek korpusu
- většinou slovo, číslo, interpunkce
 - **bude-li, don't** – 4 možnosti:
 1. |don't|
 2. |don| '|t|
 3. |don| '|' |t|
 4. |do| |n't| – v BNC
 - zkratky (s tečkama?)
 - datumy
 - desetinná čísla, ...

Tokenizace

Rozdělení textu do pozic

- může silně ovlivnit výsledky dotazování, četnosti i značkování
- token (pozice) = základní prvek korpusu
- většinou slovo, číslo, interpunkce
 - **bude-li, don't** – 4 možnosti:
 1. |don't|
 2. |don| '|t|
 3. |don| '|' |t|
 4. |do| |n't| – v BNC
 - zkratky (s tečkama?)
 - datumy
 - desetinná čísla, ...

Vertikální text

- jednoduchý formát i jeho zpracování
 - každý token na samostatném řádku
 - struktury formou XML značek
 - značkování odděleno tabulátorem (různé atributy k dané pozici)
- podrobnosti na:
 - <http://nlp.fi.muni.cz/>
 - → Informace pro současné a potenciální spolupracovníky
 - → Textové korpusy
 - → Popis vertikálů

Vertikální text

- jednoduchý formát i jeho zpracování
 - každý token na samostatném řádku
 - struktury formou XML značek
 - značkování odděleno tabulátorem (různé atributy k dané pozici)
- podrobnosti na:
 - <http://nlp.fi.muni.cz/>
 - → Informace pro současné a potenciální spolupracovníky
 - → Textové korpusy
 - → Popis vertikálů

Vertikální text

- jednoduchý formát i jeho zpracování
 - každý token na samostatném řádku
 - struktury formou XML značek
 - značkování odděleno tabulátorem (různé atributy k dané pozici)
- podrobnosti na:
 - <http://nlp.fi.muni.cz/>
 - → Informace pro současné a potenciální spolupracovníky
 - → Textové korpusy
 - → Popis vertikálů

Vertikální text

- jednoduchý formát i jeho zpracování
 - každý token na samostatném řádku
 - struktury formou XML značek
 - značkování odděleno tabulátorem (různé atributy k dané pozici)
- podrobnosti na:
 - <http://nlp.fi.muni.cz/>
 - → Informace pro současné a potenciální spolupracovníky
 - → Textové korpusy
 - → Popis vertikálů

Vertikální text

- jednoduchý formát i jeho zpracování
 - každý token na samostatném řádku
 - struktury formou XML značek
 - značkování odděleno tabulátorem (různé atributy k dané pozici)
- podrobnosti na:
 - <http://nlp.fi.muni.cz/>
 - → Informace pro současné a potenciální spolupracovníky
 - → Textové korpusy
 - → Popis vertikálů

Vertikální text

- jednoduchý formát i jeho zpracování
 - každý token na samostatném řádku
 - struktury formou XML značek
 - značkování odděleno tabulátorem (různé atributy k dané pozici)
- podrobnosti na:
 - <http://nlp.fi.muni.cz/>
 - → Informace pro současné a potenciální spolupracovníky
 - → Textové korpusy
 - → Popis vertikálů

Vertikální text

- jednoduchý formát i jeho zpracování
 - každý token na samostatném řádku
 - struktury formou XML značek
 - značkování odděleno tabulátorem (různé atributy k dané pozici)
- podrobnosti na:
 - <http://nlp.fi.muni.cz/>
 - → Informace pro současné a potenciální spolupracovníky
 - → Textové korpusy
 - → Popis vertikálů

Vertikální text

- jednoduchý formát i jeho zpracování
 - každý token na samostatném řádku
 - struktury formou XML značek
 - značkování odděleno tabulátorem (různé atributy k dané pozici)
- podrobnosti na:
 - <http://nlp.fi.muni.cz/>
 - → Informace pro současné a potenciální spolupracovníky
 - → Textové korpusy
 - → Popis vertikálů

Vertikální text

- jednoduchý formát i jeho zpracování
 - každý token na samostatném řádku
 - struktury formou XML značek
 - značkování odděleno tabulátorem (různé atributy k dané pozici)
- podrobnosti na:
 - <http://nlp.fi.muni.cz/>
 - → Informace pro současné a potenciální spolupracovníky
 - → Textové korpusy
 - → Popis vertikálů

Zpracování textů na UNIXu

- coreutils

- cat, head, tail, wc, sort, uniq, comm
- cut, paste, join, tr

- grep

- awk

- sed / perl

Zpracování textů na UNIXu

- coreutils

- cat, head, tail, wc, sort, uniq, comm
 - cut, paste, join, tr

- grep

- awk

- sed / perl

Zpracování textů na UNIXu

- coreutils

- cat, head, tail, wc, sort, uniq, comm
- cut, paste, join, tr

- grep

- awk

- sed / perl

Zpracování textů na UNIXu

- coreutils
 - cat, head, tail, wc, sort, uniq, comm
 - cut, paste, join, tr
- grep
- awk
- sed / perl

Zpracování textů na UNIXu

- coreutils
 - cat, head, tail, wc, sort, uniq, comm
 - cut, paste, join, tr
- grep
- awk
- sed / perl

Zpracování textů na UNIXu

- coreutils
 - cat, head, tail, wc, sort, uniq, comm
 - cut, paste, join, tr
- grep
- awk
- sed / perl

Příklady použití coreutils

- slovník z vertikálního textu

```
cut -f 1 -s desam.vert |sort |uniq -c \
|sort -rn >desam.dict
```

- jednoduchá tokenizace

```
tr -cs 'a-zA-Z0-9' '\n' <GPL >GPL.vert
cat GPL.vert |sort |uniq -c |sort -rn >GPL.dict
```

- všechny bigramy

```
tail -n +2 GPL.vert |paste GPL.vert - |sort |uniq -c
|sort -rn
```

Příklady použití coreutils

- slovník z vertikálního textu

```
cut -f 1 -s desam.vert |sort |uniq -c \
|sort -rn >desam.dict
```

- jednoduchá tokenizace

```
tr -cs 'a-zA-Z0-9' '\n' <GPL >GPL.vert
cat GPL.vert |sort |uniq -c |sort -rn >GPL.dict
```

- všechny bigramy

```
tail -n +2 GPL.vert |paste GPL.vert - |sort |uniq -c
|sort -rn
```

Příklady použití coreutils

- slovník z vertikálního textu

```
cut -f 1 -s desam.vert |sort |uniq -c \
|sort -rn >desam.dict
```

- jednoduchá tokenizace

```
tr -cs 'a-zA-Z0-9' '\n' <GPL >GPL.vert
cat GPL.vert |sort |uniq -c |sort -rn >GPL.dict
```

- všechny bigramy

```
tail -n +2 GPL.vert |paste GPL.vert - |sort |uniq -c
|sort -rn
```

Korpusové manažery

nástroje na zpracování korpusů

- uložení textu
- editace/příprava textu
- značkování
- rozdělení do pozic (tokenizace)
- vyhledávání (konkordance)
- statistiky

Korpusové manažery

nástroje na zpracování korpusů

- uložení textu
- editace/příprava textu
- značkování
- rozdělení do pozic (tokenizace)
- vyhledávání (konkordance)
- statistiky

Korpusové manažery

nástroje na zpracování korpusů

- uložení textu
- editace/příprava textu
- značkování
 - rozdělení do pozic (tokenizace)
 - vyhledávání (konkordance)
 - statistiky

Korpusové manažery

nástroje na zpracování korpusů

- uložení textu
- editace/příprava textu
- značkování
- rozdělení do pozic (tokenizace)
- vyhledávání (konkordance)
- statistiky

Korpusové manažery

nástroje na zpracování korpusů

- uložení textu
- editace/příprava textu
- značkování
- rozdělení do pozic (tokenizace)
- vyhledávání (konkordance)
- statistiky

Korpusové manažery

nástroje na zpracování korpusů

- uložení textu
- editace/příprava textu
- značkování
- rozdělení do pozic (tokenizace)
- vyhledávání (konkordance)
- statistiky

Systém Manatee

- korpusový manažer
- přímo podporuje
 - * uložení textu
 - * vyhledávání (konkordance)
 - * statistiky
- externí nástroje
 - * značkování
 - * rozdělení do pozic

Systém Manatee

- korpusový manažer
- přímo podporuje
 - uložení textu
 - vyhledávání (konkordance)
 - statistiky
- externí nástroje
 - značkování
 - rozdělení do pozic

Systém Manatee

- korpusový manažer
- přímo podporuje
 - uložení textu
 - vyhledávání (konkordance)
 - statistiky
- externí nástroje
 - značkování
 - rozdělení do pozic

Systém Manatee

- korpusový manažer
- přímo podporuje
 - uložení textu
 - vyhledávání (konkordance)
 - statistiky
- externí nástroje
 - značkování
 - rozdělení do pozic

Systém Manatee

- korpusový manažer
- přímo podporuje
 - uložení textu
 - vyhledávání (konkordance)
 - statistiky
- externí nástroje
 - značkování
 - rozdělení do pozic

Systém Manatee

- korpusový manažer
- přímo podporuje
 - uložení textu
 - vyhledávání (konkordance)
 - statistiky
- externí nástroje
 - značkování
 - rozdělení do pozic

Systém Manatee

- korpusový manažer
- přímo podporuje
 - uložení textu
 - vyhledávání (konkordance)
 - statistiky
- externí nástroje
 - značkování
 - rozdělení do pozic

Systém Manatee

- korpusový manažer
- přímo podporuje
 - uložení textu
 - vyhledávání (konkordance)
 - statistiky
- externí nástroje
 - značkování
 - rozdělení do pozic

Systém Manatee

hlavní zaměření

- velké korpusy
- rozsáhlé značkování
 - * morfologické, syntaktické, metainformace
- návaznost na další aplikace/nástroje
 - * korpusový editor (CED), tvorba slovníků
- univerzálnost
 - * různé jazyky, kódování, systémy značek

Systém Manatee

hlavní zaměření

- velké korpusy
- rozsáhlé značkování
 - morfologické, syntaktické, metainformace
- návaznost na další aplikace/nástroje
 - korpusový editor (CED), tvorba slovníků
- univerzálnost
 - různé jazyky, kódování, systémy značek

Systém Manatee

hlavní zaměření

- velké korpusy
- rozsáhlé značkování
 - morfologické, syntaktické, metainformace
- návaznost na další aplikace/nástroje
 - korpusový editor (CED), tvorba slovníků
- univerzálnost
 - různé jazyky, kódování, systémy značek

Systém Manatee

hlavní zaměření

- velké korpusy
- rozsáhlé značkování
 - morfologické, syntaktické, metainformace
- návaznost na další aplikace/nástroje
 - korpusový editor (CED), tvorba slovníků
- univerzálnost
 - různé jazyky, kódování, systémy značek

Systém Manatee

hlavní zaměření

- velké korpusy
- rozsáhlé značkování
 - morfologické, syntaktické, metainformace
- návaznost na další aplikace/nástroje
 - korpusový editor (CED), tvorba slovníků
- univerzálnost
 - různé jazyky, kódování, systémy značek

Systém Manatee

hlavní zaměření

- velké korpusy
- rozsáhlé značkování
 - morfologické, syntaktické, metainformace
- návaznost na další aplikace/nástroje
 - korpusový editor (CED), tvorba slovníků
- univerzálnost
 - různé jazyky, kódování, systémy značek

Systém Manatee

hlavní zaměření

- velké korpusy
- rozsáhlé značkování
 - morfologické, syntaktické, metainformace
- návaznost na další aplikace/nástroje
 - korpusový editor (CED), tvorba slovníků
- univerzálnost
 - různé jazyky, kódování, systémy značek

Klíčové vlastnosti

- modulární systém
- přístup z různých rozhraní
 - grafické uživatelské rozhraní (Bonito)
 - aplikační programové rozhraní (API)
 - příkazový řádek
- rozsáhlá data
 - až 2 mld. pozic
 - neomezeně atributů a metainformací
- rychlosť
 - vyhledávání, statistiky

Klíčové vlastnosti

- modulární systém
- přístup z různých rozhraní
 - grafické uživatelské rozhraní (Bonito)
 - aplikační programové rozhraní (API)
 - příkazový řádek
- rozsáhlá data
 - až 2 mld. pozic
 - neomezeně atributů a metainformací
- rychlosť
 - vyhledávání, statistiky

Klíčové vlastnosti

- modulární systém
- přístup z různých rozhraní
 - grafické uživatelské rozhraní (Bonito)
 - aplikační programové rozhraní (API)
 - příkazový řádek
- rozsáhlá data
 - až 2 mld. pozic
 - neomezeně atributů a metainformací
- rychlosť
 - vyhledávání, statistiky

Klíčové vlastnosti

- modulární systém
- přístup z různých rozhraní
 - grafické uživatelské rozhraní (Bonito)
 - aplikační programové rozhraní (API)
 - příkazový řádek
- rozsáhlá data
 - až 2 mld. pozic
 - neomezeně atributů a metainformací
- rychlosť
 - vyhledávání, statistiky

Klíčové vlastnosti

- modulární systém
- přístup z různých rozhraní
 - grafické uživatelské rozhraní (Bonito)
 - aplikační programové rozhraní (API)
 - příkazový řádek
- rozsáhlá data
 - až 2 mld. pozic
 - neomezeně atributů a metainformací
- rychlosť
 - vyhledávání, statistiky

Klíčové vlastnosti

- modulární systém
- přístup z různých rozhraní
 - grafické uživatelské rozhraní (Bonito)
 - aplikační programové rozhraní (API)
 - příkazový řádek
- rozsáhlá data
 - až 2 mld. pozic
 - neomezeně atributů a metainformací
- rychlosť
 - vyhledávání, statistiky

Klíčové vlastnosti

- modulární systém
- přístup z různých rozhraní
 - grafické uživatelské rozhraní (Bonito)
 - aplikační programové rozhraní (API)
 - příkazový řádek
- rozsáhlá data
 - až 2 mld. pozic
 - neomezeně atributů a metainformací
- rychlosť
 - vyhledávání, statistiky

Klíčové vlastnosti

- modulární systém
- přístup z různých rozhraní
 - grafické uživatelské rozhraní (Bonito)
 - aplikační programové rozhraní (API)
 - příkazový řádek
- rozsáhlá data
 - až 2 mld. pozic
 - neomezeně atributů a metainformací
- rychlosť
 - vyhledávání, statistiky

Klíčové vlastnosti

- modulární systém
- přístup z různých rozhraní
 - grafické uživatelské rozhraní (Bonito)
 - aplikační programové rozhraní (API)
 - příkazový řádek
- rozsáhlá data
 - až 2 mld. pozic
 - neomezeně atributů a metainformací
- rychlosť
 - vyhledávání, statistiky

Klíčové vlastnosti

- modulární systém
- přístup z různých rozhraní
 - grafické uživatelské rozhraní (Bonito)
 - aplikační programové rozhraní (API)
 - příkazový řádek
- rozsáhlá data
 - až 2 mld. pozic
 - neomezeně atributů a metainformací
- rychlosť
 - vyhledávání, statistiky

Klíčové vlastnosti

- **multihodnoty**

- zpracování víceznačných značkování

- **dynamické atributy**

- vyhledávání a statistiky na počítaných datech

- **subkorpusy**

- **silný dotazovací jazyk**

- dotazy na všechny atributy, metainformace

- pozitivní/negativní filtry

- regulérní výrazy + booleovské operátory

Klíčové vlastnosti

- multihodnoty
 - zpracování víceznačných značkování
- dynamické atributy
 - vyhledávání a statistiky na počítaných datech
- subkorpusy
- silný dotazovací jazyk
 - dotazy na všechny atributy, metainformace
 - pozitivní/negativní filtry
 - regulérní výrazy + booleovské operátory

Klíčové vlastnosti

- multihodnoty
 - zpracování víceznačných značkování
- dynamické atributy
 - vyhledávání a statistiky na počítaných datech
- subkorpusy
- silný dotazovací jazyk
 - dotazy na všechny atributy, metainformace
 - pozitivní/negativní filtry
 - regulérní výrazy + booleovské operátory

Klíčové vlastnosti

- multihodnoty
 - zpracování víceznačných značkování
- dynamické atributy
 - vyhledávání a statistiky na počítaných datech
- subkorpusy
- silný dotazovací jazyk
 - dotazy na všechny atributy, metainformace
 - pozitivní/negativní filtry
 - regulérní výrazy + booleovské operátory

Klíčové vlastnosti

- multihodnoty
 - zpracování víceznačných značkování
- dynamické atributy
 - vyhledávání a statistiky na počítaných datech
- subkorpusy
- silný dotazovací jazyk
 - dotazy na všechny atributy, metainformace
 - pozitivní/negativní filtry
 - regulérní výrazy + booleovské operátory

Klíčové vlastnosti

- multihodnoty
 - zpracování víceznačných značkování
- dynamické atributy
 - vyhledávání a statistiky na počítaných datech
- subkorpusy
- silný dotazovací jazyk
 - dotazy na všechny atributy, metainformace
 - pozitivní/negativní filtry
 - regulérní výrazy + booleovské operátory

Klíčové vlastnosti

- multihodnoty
 - zpracování víceznačných značkování
- dynamické atributy
 - vyhledávání a statistiky na počítaných datech
- subkorpusy
- silný dotazovací jazyk
 - dotazy na všechny atributy, metainformace
 - pozitivní/negativní filtry
 - regulérní výrazy + booleovské operátory

Klíčové vlastnosti

- multihodnoty
 - zpracování víceznačných značkování
- dynamické atributy
 - vyhledávání a statistiky na počítaných datech
- subkorpusy
- silný dotazovací jazyk
 - dotazy na všechny atributy, metainformace
 - pozitivní/negativní filtry
 - regulérní výrazy + booleovské operátory

Klíčové vlastnosti

- multihodnoty
 - zpracování víceznačných značkování
- dynamické atributy
 - vyhledávání a statistiky na počítaných datech
- subkorpusy
- silný dotazovací jazyk
 - dotazy na všechny atributy, metainformace
 - pozitivní/negativní filtry
 - regulérní výrazy + booleovské operátory

Klíčové vlastnosti

- frekvenční distribuce
 - víceúrovňová
 - všechny atributy a metainformace
- kolokace
 - různé statistické funkce

Klíčové vlastnosti

- frekvenční distribuce
 - víceúrovňová
 - všechny atributy a metainformace
- kolokace
 - různé statistické funkce

Klíčové vlastnosti

- frekvenční distribuce
 - víceúrovňová
 - všechny atributy a metainformace
- kolokace
 - různé statistické funkce

Klíčové vlastnosti

- frekvenční distribuce
 - víceúrovňová
 - všechny atributy a metainformace
- kolokace
 - různé statistické funkce

Klíčové vlastnosti

- frekvenční distribuce
 - víceúrovňová
 - všechny atributy a metainformace
- kolokace
 - různé statistické funkce