

Corpus Query System Bonito

Recent Development

Vojtěch Kovář

Faculty of Informatics, Masaryk University Brno
Botanická 68a, 602 00 Brno, Czech Republic
xkovar3@fi.muni.cz

RASLAN 2007

Outline

- 1 Motivation
- 2 CQS Manatee/Bonito
- 3 Saving outputs in XML
- 4 Localization Mechanism
- 5 Conclusions and Future Directions

Motivation

- Big text corpora
 - very useful source of linguistic information
 - used for research, language teaching, lexicography ...
- Corpus query systems (CQS)
 - enable people to work with large text data
 - comfortable searching and querying corpora

Motivation

- Big text corpora
 - very useful source of linguistic information
 - used for research, language teaching, lexicography ...
- Corpus query systems (CQS)
 - enable people to work with large text data
 - comfortable searching and querying corpora

Motivation

- Big text corpora
 - very useful source of linguistic information
 - used for research, language teaching, lexicography ...
- Corpus query systems (CQS)
 - enable people to work with large text data
 - comfortable searching and querying corpora

Motivation

- Big text corpora
 - very useful source of linguistic information
 - used for research, language teaching, lexicography ...
- Corpus query systems (CQS)
 - enable people to work with large text data
 - comfortable searching and querying corpora

Motivation

- Big text corpora
 - very useful source of linguistic information
 - used for research, language teaching, lexicography ...
- Corpus query systems (CQS)
 - enable people to work with large text data
 - comfortable searching and querying corpora

Motivation

- Big text corpora
 - very useful source of linguistic information
 - used for research, language teaching, lexicography ...
- Corpus query systems (CQS)
 - enable people to work with large text data
 - comfortable searching and querying corpora

Motivation

■ CQS Manatee/Bonito

- developed at Masaryk university, Czech Republic
- fast searching mechanism, powerful query language
- user-friendly interface
- wide variety of languages
- extending by adding new functions according to users needs

■ Newly added functions

- saving outputs in XML
- localization mechanism

Motivation

■ CQS Manatee/Bonito

- developed at Masaryk university, Czech Republic
- fast searching mechanism, powerful query language
- user-friendly interface
- wide variety of languages
- extending by adding new functions according to users needs

■ Newly added functions

- saving outputs in XML
- localization mechanism

Motivation

■ CQS Manatee/Bonito

- developed at Masaryk university, Czech Republic
- fast searching mechanism, powerful query language
- user-friendly interface
- wide variety of languages
- extending by adding new functions according to users needs

■ Newly added functions

- saving outputs in XML
- localization mechanism

Motivation

■ CQS Manatee/Bonito

- developed at Masaryk university, Czech Republic
- fast searching mechanism, powerful query language
- user-friendly interface
- wide variety of languages
- extending by adding new functions according to users needs

■ Newly added functions

- saving outputs in XML
- localization mechanism

Motivation

■ CQS Manatee/Bonito

- developed at Masaryk university, Czech Republic
- fast searching mechanism, powerful query language
- user-friendly interface
- wide variety of languages
- extending by adding new functions according to users needs

■ Newly added functions

- saving outputs in XML
- localization mechanism

Motivation

■ CQS Manatee/Bonito

- developed at Masaryk university, Czech Republic
- fast searching mechanism, powerful query language
- user-friendly interface
- wide variety of languages
- extending by adding new functions according to users needs

■ Newly added functions

- saving outputs in XML
- localization mechanism

Motivation

■ CQS Manatee/Bonito

- developed at Masaryk university, Czech Republic
- fast searching mechanism, powerful query language
- user-friendly interface
- wide variety of languages
- extending by adding new functions according to users needs

■ Newly added functions

- saving outputs in XML
- localization mechanism

Motivation

■ CQS Manatee/Bonito

- developed at Masaryk university, Czech Republic
- fast searching mechanism, powerful query language
- user-friendly interface
- wide variety of languages
- extending by adding new functions according to users needs

■ Newly added functions

- saving outputs in XML
- localization mechanism

Motivation

■ CQS Manatee/Bonito

- developed at Masaryk university, Czech Republic
- fast searching mechanism, powerful query language
- user-friendly interface
- wide variety of languages
- extending by adding new functions according to users needs

■ Newly added functions

- saving outputs in XML
- localization mechanism

Manatee/Bonito System

■ Manatee

- low-level corpus processing
- fast searching even in very big corpora (billions of words)

■ Bonito

- graphical user interface
- works through standard web browser
→ platform independent
- written in Python language
- templating engine for generating web pages

Manatee/Bonito System

■ Manatee

- low-level corpus processing
- fast searching even in very big corpora (billions of words)

■ Bonito

- graphical user interface
- works through standard web browser
→ platform independent
- written in Python language
- templating engine for generating web pages

Manatee/Bonito System

■ Manatee

- low-level corpus processing
- fast searching even in very big corpora (billions of words)

■ Bonito

- graphical user interface
- works through standard web browser
→ platform independent
- written in Python language
- templating engine for generating web pages

Manatee/Bonito System

■ Manatee

- low-level corpus processing
- fast searching even in very big corpora (billions of words)

■ Bonito

- graphical user interface
- works through standard web browser
→ platform independent
- written in Python language
- templating engine for generating web pages

Manatee/Bonito System

■ Manatee

- low-level corpus processing
- fast searching even in very big corpora (billions of words)

■ Bonito

- graphical user interface
- works through standard web browser
 - platform independent
- written in Python language
- templating engine for generating web pages



Manatee/Bonito System

■ Manatee

- low-level corpus processing
- fast searching even in very big corpora (billions of words)

■ Bonito

- graphical user interface
- works through standard web browser
 - platform independent
- written in Python language
- templating engine for generating web pages



Manatee/Bonito System

■ Manatee

- low-level corpus processing
- fast searching even in very big corpora (billions of words)

■ Bonito

- graphical user interface
- works through standard web browser
 - platform independent
- written in Python language
- templating engine for generating web pages



Manatee/Bonito System

■ Manatee

- low-level corpus processing
- fast searching even in very big corpora (billions of words)

■ Bonito

- graphical user interface
- works through standard web browser
 - platform independent
- written in Python language
- templating engine for generating web pages



[Home](#)
[Concordance](#)
[Word List](#)
[Word Sketch](#)
[Thesaurus](#)
[Sketch-Diff](#)
[View options](#)
[Sample](#)
[Filter](#)
[Sort](#)
[Frequency](#)
[Collocation](#)
[Save](#)

Page of 39 Go [Next](#) | [Last](#)

Corpus: **British National Corpus**
 Hits: 769
[conc description](#)

A0D Chanders and Gardners upon the Feast of **Corpus** Christie ,);</p><p>The three Shepherds
A0K ethnography , are often missing from the finite **corpus** of empirical observation . This transformational
A0K outside the usual limits of the body or **corpus** of collected material . In my own case
A1A to convey this knowledge , whether as a **corpus** or a skill . The conveying need not , indeed
A1B that can be pointed to as constituting the **corpus** or the canon of Pound 's criticism . The
A68 made the suggestion that he should sleep in **Corpus** but take his meals in his old college of
A68 wet blanket . In those days the Fellows of **Corpus** were rather proud of the briskness of their
ABS mounted .</p><p>Greatness trickled from the **corpus** of his image , his career now like a gunshot

Saving outputs in XML format

- HTML outputs in Bonito 2
 - suitable for viewing the search results
 - bad for their saving and further processing
- Two saving options added
 - Text
 - XML

Saving outputs in XML format

- HTML outputs in Bonito 2
 - suitable for viewing the search results
 - bad for their saving and further processing
- Two saving options added
 - Text
 - XML

Saving outputs in XML format

- HTML outputs in Bonito 2
 - suitable for viewing the search results
 - bad for their saving and further processing
- Two saving options added
 - Text
 - XML

Saving outputs in XML format

- HTML outputs in Bonito 2
 - suitable for viewing the search results
 - bad for their saving and further processing
- Two saving options added
 - Text
 - XML

Saving outputs in XML format

- HTML outputs in Bonito 2
 - suitable for viewing the search results
 - bad for their saving and further processing
- Two saving options added
 - Text
 - XML

Saving outputs in XML format

- HTML outputs in Bonito 2
 - suitable for viewing the search results
 - bad for their saving and further processing
- Two saving options added
 - Text
 - XML

Saving outputs in XML format

- Text and XML saving implementation
 - many possible outputs
 - new template for each possible output
- XML structure
 - different for each kind of output
 - self-explaining XML tags
 - good structuralization of the data

Saving outputs in XML format

- Text and XML saving implementation
 - many possible outputs
 - new template for each possible output
- XML structure
 - different for each kind of output
 - self-explaining XML tags
 - good structuralization of the data

Saving outputs in XML format

- Text and XML saving implementation
 - many possible outputs
 - new template for each possible output
- XML structure
 - different for each kind of output
 - self-explaining XML tags
 - good structuralization of the data

Saving outputs in XML format

- Text and XML saving implementation
 - many possible outputs
 - new template for each possible output
- XML structure
 - different for each kind of output
 - self-explaining XML tags
 - good structuralization of the data

Saving outputs in XML format

- Text and XML saving implementation
 - many possible outputs
 - new template for each possible output
- XML structure
 - different for each kind of output
 - self-explaining XML tags
 - good structuralization of the data

Saving outputs in XML format

- Text and XML saving implementation
 - many possible outputs
 - new template for each possible output
- XML structure
 - different for each kind of output
 - self-explaining XML tags
 - good structuralization of the data

Saving outputs in XML format

- Text and XML saving implementation
 - many possible outputs
 - new template for each possible output
- XML structure
 - different for each kind of output
 - self-explaining XML tags
 - good structuralization of the data

Save Concordance

Save concordance as: Text XML

Save pages: All
 Only page:

Include heading:

Number lines:

Align KWIC:

Maximum number of lines:


```
<concordance>
  <heading>
    <corpus>bnc</corpus>
    <hits>769</hits>
    <query>lc,[word="corpus"|lemma="corpus"] 769 </query>
  </heading>
  <lines>
    <line>
      <ref>A0D</ref>
      <left_context>Chandlers and Gardners upon the Feast of </left_context>
      <kwic> Corpus </kwic>
      <right_context> Christie , ) : The three Shepherds </right_context>
    </line>
    <line>
      <ref>A0K</ref>
      <left_context>ethnography , are often missing from the finite </left_context>
      <kwic> corpus </kwic>
      <right_context> of empirical observation . This transformational </right_context>
    </line>
  </lines>
</concordance>
```

Localization Mechanism

■ Motivation

- worldwide use of the system

■ Templating engine

- old templating engine not suitable for the localization
- switching to the Cheetah Templating Engine
- translating all templates

■ Translation

- using gettext tools

Localization Mechanism

■ Motivation

- worldwide use of the system

■ Templating engine

- old templating engine not suitable for the localization
- switching to the Cheetah Templating Engine
- translating all templates

■ Translation

- using gettext tools

Localization Mechanism

■ Motivation

- worldwide use of the system

■ Templating engine

- old templating engine not suitable for the localization
- switching to the Cheetah Templating Engine
- translating all templates

■ Translation

- using gettext tools

Localization Mechanism

■ Motivation

- worldwide use of the system

■ Templating engine

- old templating engine not suitable for the localization
- switching to the Cheetah Templating Engine
- translating all templates

■ Translation

- using gettext tools

Localization Mechanism

■ Motivation

- worldwide use of the system

■ Templating engine

- old templating engine not suitable for the localization
- switching to the Cheetah Templating Engine
- translating all templates

■ Translation

- using gettext tools

Localization Mechanism

■ Motivation

- worldwide use of the system

■ Templating engine

- old templating engine not suitable for the localization
- switching to the Cheetah Templating Engine
- translating all templates

■ Translation

- using `gettext` tools

Localization Mechanism

■ Motivation

- worldwide use of the system

■ Templating engine

- old templating engine not suitable for the localization
- switching to the Cheetah Templating Engine
- translating all templates

■ Translation

- using `gettext` tools

Localization Mechanism

■ Motivation

- worldwide use of the system

■ Templating engine

- old templating engine not suitable for the localization
- switching to the Cheetah Templating Engine
- translating all templates

■ Translation

- using gettext tools

Localization Mechanism

- User interface language selection
 - preferred language in web browser
 - setting the UI language according to the user
- Input and output encoding
 - corpora in different encodings
 - → using UTF 8
 - recoding all inputs and outputs from/to UTF 8

Localization Mechanism

- User interface language selection
 - preferred language in web browser
 - setting the UI language according to the user
- Input and output encoding
 - corpora in different encodings
 - → using UTF 8
 - recoding all inputs and outputs from/to UTF 8

Localization Mechanism

- User interface language selection
 - preferred language in web browser
 - setting the UI language according to the user
- Input and output encoding
 - corpora in different encodings
 - → using UTF 8
 - recoding all inputs and outputs from/to UTF 8

Localization Mechanism

- User interface language selection
 - preferred language in web browser
 - setting the UI language according to the user
- Input and output encoding
 - corpora in different encodings
 - → using UTF 8
 - recoding all inputs and outputs from/to UTF 8

Localization Mechanism

- User interface language selection
 - preferred language in web browser
 - setting the UI language according to the user
- Input and output encoding
 - corpora in different encodings
 - → using UTF 8
 - recoding all inputs and outputs from/to UTF 8

Localization Mechanism

- User interface language selection
 - preferred language in web browser
 - setting the UI language according to the user
- Input and output encoding
 - corpora in different encodings
 - → using UTF 8
 - recoding all inputs and outputs from/to UTF 8

Localization Mechanism

- User interface language selection
 - preferred language in web browser
 - setting the UI language according to the user
- Input and output encoding
 - corpora in different encodings
 - → using UTF 8
 - recoding all inputs and outputs from/to UTF 8

[Hlavní strana](#)
[Konkordance](#)
[Seznamy slov](#)
[Word Sketch](#)
[Tezaurus](#)
[Sketch-Diff](#)

Korpus:

Dotaz:

Další možnosti

Lemma: Slovní druh:

Fráze:

Slovní tvar: Slovní druh: Rozlišovat velikost písmen:

CQL:

Implicitní atribut: [Přehled značek](#)

Kontext

Typ dotazu: z následujících.

	Levý kontext	Pravý kontext
Velikost kontextu:	<input type="text" value="5"/> tokenů.	<input type="text" value="5"/> tokenů.
Lemma:	<input type="text"/>	<input type="text"/>
Slovní druh: (Ctrl+click pro vicenásobný výběr)	<input type="text" value="adjective"/> <input type="text" value="adverb"/> <input type="text" value="conjunction"/> <input type="text" value="determiner"/>	<input type="text" value="adjective"/> <input type="text" value="adverb"/> <input type="text" value="conjunction"/> <input type="text" value="determiner"/>

Typy textů

Hlavní strana	Konkordance	Seznamy slov
---------------	-------------	--------------

Možnosti zobrazení	Příklad	Filtr	Třídění	Frekvence	Kolokace	Uložit
--------------------	---------	-------	---------	-----------	----------	--------

Korpus: Persian web corpus
Výskytů: 1384
popis konkordance

<http://amin177.blogfa.com/><http://amin177.blogfa.com/><http://amin177.blogfa.com/><http://amin177.blogfa.com/><http://amin177.blogfa.com/><http://amin177.blogfa.com/><http://amin177.blogfa.com/><http://amin177.blogfa.com/>

[První](#) | [Předchozí](#) | Strana ze 70 Go | [Další](#) | [Poslední](#)

بارکته‌ی تاریک فرض کردند ، فکر کنید به لاج از زمین **نیست** ، نه اینکه جاش ختایه ، نیست ، عن حدی نیازی به
 کنید به لاج از زمین نیست ، نه اینکه جاش ختایه ، **نیست** ، عن حدی نیازی به نور چراغ روی سرم نثارم ولم لازمه
 نیگه لازمه و بدون نور هیچ چیز دیده‌ای در کار **نیست** ، با نور تازه همیشه بقتضی چیزی رو دید و این هواج
 این حیانه یاد از یاد اقسام ، بختری که چندان زیبا **نیست** ولم هواج راد رقتن به قول گویندو کنیلهای بزرگش
 ابترت ایران گفتن شده | به نظرم این قضیه اصلا تضادس **نیست** و بچه ها حد خوان به آرمایشی بکنن بیین گفتن کاهل
 شهرستانه که احیان گذرشون به این ویلاگ حد افه بد **نیست** بکن وشع اولورا چه جوریه ! اینم بگم که از کتفه
 : یک نوسه همین | ارف از تو به جا بود وایکن چه کنم **نیست** + نوشته شده در یکشنبه پانزدهم مرداد 1385 ساعت 1
 داستان و عقابه و به سری چیزای نیگه که ختس معلوم **نیست** چه حدشن بخترین ! از دیروز جمعه 9 اردیبهشت 85 ، ۱۰

Conclusions and Future Directions

Conclusions

- We have presented two of recently added features in the Manatee/Bonito system

Future Development

- Adding more features to the system according to the users needs

Conclusions and Future Directions

Conclusions

- We have presented two of recently added features in the Manatee/Bonito system

Future Development

- Adding more features to the system according to the users needs

Conclusions and Future Directions

Conclusions

- We have presented two of recently added features in the Manatee/Bonito system

Future Development

- Adding more features to the system according to the users needs