Bonito (version 1.4) – User's Guide

Pavel Rychly, TextForge

May, 2003

1 Introduction

Bonito is a graphical user interface (GUI) of a Manatee corpus manager. It enables queries to be formed and given to various corpora. The results are clearly displayed and can be changed in various ways. Statistics can also be computed on them.

1.1 Corpus

A Corpus is here defined as a sequence of so-called positions. Each position is made up of one word, number or punctuation mark, etc. The actual division into individual positions is performed automatically by external tools in most corpora and does not depend on Manatee or Bonito systems in any way. This may be different in various corpora.

Each position consists of a set of positional attributes. Each attribute contains a piece of simple word information (a word, a basic form, a part of speech, etc.). The position of an optional corpus always contains a minimum of one attribute with name *word*. The attribute contains an actual word at the given position. Different corpora contain different sets of attributes. Some corpora contain only the attribute mentioned above. Others contain a word, the basic form of the word (attribute *lemma*) and grammatical information (attribute *tag*). Some corpora contain grammatical information divided into more specific attributes.

The corpus may also contain various structure tags, such as sentence boundaries, paragraph or document boundaries. Certain types of tags may also contain additional information. Thus in many corpora the whole text is divided into documents by the structure tag <doc>. This structure tag usually contains the document source identifier.

In some annotated corpora grammatical information stored for each word is often denoted as a tag. This type of information, however, does not stand for the structure tags mentioned above. These tags (grammatical information) are stored in some of the positional attributes.

1.2 Corpus manager

The corpus query result is the so-called concordance list that creates all corpus positions corresponding with the query given. The concordance list is then displayed in KWIC (Key Word(s) In Context) format. The searched words are displayed with their contexts one below the other. The concordance list is sometimes abbreviated as concordance. The abbreviation KWIC represents the searched word or a word sequence.

2 Application layout

There is a menu in the upper part of the manager's main window and there is an entry box for making queries below the window. On the right at the end of the entry line there is a button with a current corpus name. It is possible to select a different corpus from the displayed list after clicking on the button and holding the mouse button.

Most of the window area is formed by the concordance list where query results are displayed. In the lower part there is the area of detail where a selected KWIC with a larger context is displayed. It is possible to zoom in the context interactively: **Up Arrow** enlarges the context to the left, **Down Arrow** to the right. There is a status line on the lowest border of the main window. It informs the user about the operation which is currently in progress. If there is no operation in progress, it informs the user about the lines displayed.

Information about the displayed lines is beside the word *Displayed* and it has one of the following forms:

- If all the result lines are displayed, *All* is written in the status line. The word is followed by a slash (/) and the number of lines (e.g. *All/13*).
- If only a part of the result is displayed, the information about the displayed lines seems more complicated at first (e.g. *141+50/363 (13%)*). The first number (before the + symbol) shows the number of the first displayed line. The number of displayed lines follows and there is the total number of lines behind the slash (/). In brackets there is the percentage of the displayed lines of the total number of all result lines. The example shown above can be read in the following way: from the 141st line 50 lines are displayed out of the total 363, which is 13%.

The status line beside the word *Line* contains the number of the current line and if a line is selected, the word *Selected* is followed by the number of selected lines.

It is possible to use the tabulator key for moving between the individual entry boxes of the query, the concordance list and the detail. The **F12** key switches between the concordance list and the entry box of the query. Space bar on a menu-button shows a menu with all possible values.

3 Main Menu

This chapter explains all of the menu items. ▹**Menu**▹**Item** is used in this text to denote a menu item *Item* from the menu *Menu*. For example menu item *Connection* from the *Manager* menu is denoted by ▹**Manager**▹**Connection**

3.1 Manager

Connection

Establishes connection to the server. It is not possible to work with corpora without a connection. It is possible to connect to either a remote server through the Internet/intranet or to launch the server locally. If the Internet connection is not set in the configuration, an automatic connection is performed once Bonito is launched. Otherwise this menu item is automatically selected and it is necessary to enter the user's name and password in the displayed window.

Internet connection

The connection is performed to a remote server, the address of which is in the entry box *Server address*. *User name* and *password* are required.

Run local

If you select this option, the server runs locally, that is, in the same computer as Bonito, and the whole manager and all corpora to be used have to be installed in this computer. The user's name and password are ignored. The access to the selected corpora is determined by the permissions of the corpus data files. For the server launch the command in the *Server command* entry box is used.

Server address

The address of the computer which the server is running on e.g. aurora.fi.muni.cz.

User name, Password

User's login and password. These can be different from user's login and/or password on the server. Password characters are not displayed in the entry box.

Server command

The command for the local server launch. (e.g. manateesrv "desam susanne" 2>errlog).

Change password

Changes the user's password. The old password entry is required and the new password entry is required twice. Password characters are not displayed in the entry boxes.

Refresh Result

Redraws the contents of the result box according to the assignment in ▹**Display**▹**Range**).

Options

Displays window with user's options for start-up or saving options for the next start-up. Default option values are set to suit most users. Advanced users may change the following options:

- *Help hits* – corresponds with the menu item ▹**Help**▹**Show** help hits
- *Browser command (for on-line help)* – this command performs the WWW browser start-up with the documentation page. It is used for displaying on-line help. (▹**Help**▹**Documentation**).
- *Maximum number of undo steps* – see menu ▹**Concordance**▹**Undo**
- *Number of queries in the history*
- *Number of "new" lines in Page-Up/Page-Down* – see menu ▹**View**▹**Page up** and ▹**View**▹**Page down**
- *Number of lines read per batch during display* – the concordance lines display runs continuously after the number of lines is assigned as the data from server are being received. If you have a fast (new) computer, the total display of many rows (more than 1000) may be accelerated. A higher value must be set. The acceleration shows itself only if there is a sufficient amount of data from the server. Thus, if you have a slow connection to the server (connection via modem), or if you entered a complicated query which has to be evaluated slowly, Bonito always displays only the rows the server provides—option modification will have no influence on the speed of the displaying process.
- *Internet server connection* – when checked, the radio button *Internet connection* will be selected in the login window. Otherwise, the option *Run local* will be selected.
- *Server address, Server port number* – the computer Internet address and the port number of the server on which the server part of the Manager is running.
- *Default corpus* – after connecting to the server the first available corpus is set. If you mostly work with another corpus, enter its name into this entry box and **Save**. When you connect next time, the assigned corpus will be set. The given name should match a corpus name exactly including character case.
- *Default context size* – see menu ▹**View**▹**Context**
- *Default display range* – see menu ▹**View**▹**Range**

If you use the **Apply** button, set values are valid only until the application exits. Button Save performs not only setting values for the current application – it also saves them into a configuration file and these values will be used when you next start the application. Some values (*Default* ...) are reasonable only for saving.

Change language

You can select the language of the user interface here. The selected language is the default one for next startups.

Exit

Application termination.

3.2 Corpus

Information summary

Displays the information summary of the selected corpus: its name and additional information from the corpus configuration file, size (number of positions), all attributes and structure tags. The size of vocabulary is displayed (number of entries) for each attribute; that is, the number of different words (types), number of different lemmata, tags etc. – according to the attribute. For each structure tag the number of its occurrences in the corpus is displayed.

Statistics

Basic statistical values are computed for given attribute values: the number of occurrences in the whole corpus, number of occurrences of the whole bigram, the mutual information (MI) score and T-score.

Bigram occurrences are computed according to a given span (window). Default values *from 1 to 1* mean that the second word is to directly follow the first one. Values *from 1 to 5* means the size of a span is 5, i.e. there can be up to 4 positions between the first word and the second one. A minus sign means the inverted order of the word occurrence. Thus *from -1 to -1* means that the second word is to directly precede the first one. *From -5 to 5* then means that there can be a maximum distance of 5 positions between the words and that the words can be in an arbitrary order.

Create subcorpus

Creates a subcorpus from the selected corpus according to given conditions. Subcorpora can be used only for queries. Any statistics computation uses the whole corpus.

Base corpus

The name the corpus from which a subcorpus will be created. This is the selected corpus from the main Bonito window.

Subcorpus name

After a successful subcorpus creation a new corpus will be added to the list of available corpora. The name of the new corpus will be in the form *base corpus*:*subcorpus name*. For example, if the current corpus is *susanne* and we create a subcorpus named *press* there will be *susanne:press* in the corpus list.

Structure

The name of a structure tag (structure attribute) used for subcorpus. In the subcorpus there will be only positions included in the structure of this name which fulfils the given condition. Subcorpora are typically created from the top structures which represents documents or texts.

Conditions

This condition restricts attribute(s) of the structure to certain values. For example, in the Susanne corpus the *press reportage* genre is in texts with names beginning by *A*. Texts names are stored in the *file* attribute of the <*doc*> structure in Bonito. Then the condition for a ‘press reportage’ subcorpus is:

```
file = "A.."
```

Delete subcorpus

Displays a list of all available subcorpora and allows to delete a selected subcorpus.

Word list

Displays all words from a selected attribute matching a given pattern.

Default attribute

Displays a list of all attributes of the current corpus and allows to change the default attribute which is used for queries without an attribute name.

3.3 Query

Evaluate

Performs an evaluation of the given query (or P-filter, N-filter, or collocation, see Chapter 4). The evaluation itself is performed on the server: Bonito only receives the result. When the function is activated, a **Stop** button appears to the right of the corpus name. This button enables the current action to be canceled.

List of named queries

Displays the list of named queries. A user can select a query for an evaluation or delete a selected query from the list.

List of templates

Displays the list of templates (see Chapter 5 – Templates). A user can select a template for an evaluation, delete a selected template from the list, add a new template or change the text or the description of a selected template.

Graphical construction

In the pop up window it is possible to prepare a query using a graphic interface without using query syntax.

New template

Adds a new template to the list of templates.

Name

name of the template – it will be used in queries

Template

the template text itself

Description

optional template description

Import templates

After a template file selection (e.g. from another user) it adds extra templates to the list of templates. The default templates file extension is `tpl`.

Export templates

Exports templates to the file.

Import named queries

After a query file selection (e.g. from another user) it adds new named queries to the list of named queries. The default queries file extension is `qry`.

A history file (`history.qry`) can be also imported, in this case only named queries are used.

Export named queries

Exports named queries to the given file.

3.4 Concordance

Summary

Displays a concordance summary: the corpus name, the concordance size and a list of actions (beginning by the initial query) used for the concordance creation.

Save to file

Saves the selected rows of the concordance list to a file.

Character encoding

Choose the encoding of the saved lines (value ‘-’ means the server’s corpus native encoding).

Header

Selection of the header format for saving information about a query.

Displayed lines only

Saves currently displayed lines only.

All lines

Saves all lines of the concordance list.

Line numbering

If checked the individual lines will be numbered in the output file.

Align KWIC

If checked the key words will be aligned one below the other.

The **Context** button enables changing the context for saving. This context is implicitly the same as the context for displaying (see ▹**View**▹**Context**). After pressing the **Save** button it is necessary to name the file being saved.

Print

Prints the selected lines of the concordance to a printer. The same information is required as for the ▹**Concordance**▹**Save to file**.

Delete selected

Deletes selected lines from the concordance list (see ▹**Select** menu: how to select/highlight a line). The number of selected lines is displayed in the status line in the bottom border of the main window. Depending on the current range of displayed lines, it is possible that some of the selected lines are not displayed.

Reduce

Reduce the number of lines in the concordance list. Specify which lines are to remain in the result and how many lines, percent or hundredth of a percent from the initial number of lines are to remain.

Simple Sort

Sorts the concordance list according to KWIC, left or right context and selected options.

Number of positions to sort

how many positions to sort

Sort

Specifies which positions from the line will be compared during the sort. Assume *Number of positions to sort* = 3 in the following schema. Number 1 means the most important position (sorts on this position and resolves ties by sorting higher levels), number 3 means the least important position.

```
                ........ <KWIC KWIC> ............
 left context      3 2 1 <          >
 KWIC from left          <1 2 3     >
 KWIC from right         <     3 2 1>
 right context           <          > 1 2 3
```

Ignore case

Ignores case when sorting.

Backward

Individual words will be sorted from the last character to the first one.

Attribute

A positional attribute by which the sort will be performed. It is possible to choose from the positional attributes of the corpus.

Generic Sort

Sorts the concordance list according to one or more given conditions. Every condition determines one position according to which the individual lines will be compared.

Make unique

If checked, there remains only one line in the result for all lines whose sort intervals match.

Add

Adds another sort condition.

Delete

Deletes the selected sort condition. Selection is made with mouse.

OK

Executes the sort.

Close

Closes the window without executing any sort.

Every sort condition contains:

Sorting position

The number of the position which will be compared. Negative numbers mean positions before the chosen boundary, positive numbers mean positions after it. (See also ▹**Corpus**▹**Statistics**.)

from:

Determines the boundary beginning in the same way as in the filter or collocation specifying (see Chapter 4 – Queries)

Sort attribute

Selection from corpus positional attributes.

Ignore case

Ignores case when sorting.

Backward

Individual words will be sorted from the last character to the first.

Count sort

Sorts concordance lines according to an average frequency of words in the given context. After the sorting, the first lines in the concordance contain the most frequent words.

Line group sort

Sorts concordance lines according to group numbers. Groups are sorted in ascending order (the first group has number 1), at the end of the concordance there are lines without group assignment.

Undo

Undoes the last change. It displays the previous concordance. A user can cancel the last change (reduction, deletion, P/N-filtering) or sorting. Undo can be used repeatedly. The maximum number of undo steps can be changed in ▹**Manager**▹**Settings**: the default value is 5.

Redo

Redo the last undone change.

Assign name

Gives a name to the current concordance list. If you would like to use the results of a query again you can assign a name to a concordance and go directly to this concordance without a repeated query evaluation. Named concordances are directly accessible from ▹**Concordance**▹**Named**.

Delete named

The list of all the named concordances is displayed and the requested concordance can be deleted by selecting from the list and pressing the **Delete** button.

3.4.1 Concordance->Statistics

Frequency distribution

Counts the frequency of words or other attributes, or their sequences in the requested positions.

Limit

Only sequences with frequency higher than the entered limit will be included in the result. The default limit of 0 means that all values will be counted. For concordance lists with a large number of lines, the full result can mean a large amount of data being passed from the server and sorted.

Displaying them can take a long time if more than several thousand lines are to be displayed, depending on the computer performance.

Every condition contains:

Attribute

The attribute name (selected from the corpus positional attributes)

Position

The position which will be compared.

The list works in the same way as the Generic sort. After pressing the **OK** button, the computation is executed and the result window is displayed.

It is possible to further change the way the results are displayed. This can be done using the following controls:

Limit

Lines with a frequency less or equal to the entered limit are not displayed in the result. The number of displayed lines is always counted and shown alongside.

There are three possibilities for each entered position to be chosen:

Show

Words will be displayed in the normal way.

Sum

Words will be displayed and their subtotal will be displayed. For the last position, the options **Show** and **Sum** are identical because the sum is always counted for that position.

Hide

Words will not be counted or displayed at all.

After every change of limit or display settings a new result display is performed. The width of the individual columns can be adjusted by dragging the column header with the left mouse button.

Collocations

Computes the most important collocations in the given context according to the following parameters:

Attribute

The attribute name: selected from positional attributes

In the range from, to

The initial or terminal context position. Positive values are counted to the right from the end of KWIC, negative values are counted from the beginning of KWIC to the left.

Minimum frequency in corpus

Statistics will be counted only for the words whose total frequency in the corpus is higher than the entered frequency.

Minimum frequency in given range

Statistics will be counted only for the words whose total frequency in the given context is higher than the entered frequency.

Maximum number of displayed lines

If there are more lines in the result, only the given number of the highest scored is displayed.

Sort according to frequency

Determines the type of sort according to which the result lines will be displayed. This applies only to the selection of the most frequent lines (see previous parameter): displayed lines can be then sorted according to an arbitrary column.

The absolute frequency sort is similar to the T-score and the relative frequency sort is identical with the MI-score (see details below).

The result is displayed in table form. The table can be saved to file by pressing the **Save** button. The table can be sorted according to an arbitrary column if you right mouse click on the header of the required column.

The meaning of the values in the individual columns follows:

First column

The first column is named according to the name of the counted attribute (e.g. *word*). It contains the values of the given attribute for which the statistics were counted.

MI-score

The mutual information of a word and a concordance.

T-score

The T-score of a word and a concordance

Rel. f

The relative frequency of a word, i.e. the percentage of all the occurrences of the word in the given context.

Abs. f

The absolute frequency of a word, i.e. how often a word occurred in the given context.

A right mouse button click on a word in the first column displays a menu containing two items: *P-filter* and *N-filter*. An activation of one of these items apply the appropriate filter on the current concordance.

Distribution overview

A window is displayed in which the number of lines of result (frequency) and the so-called reduced frequency can be seen. Further it shows a graphical representation of the distribution of the individual result lines within the whole corpus. Axis *X* shows the individual corpus positions, and *Y* shows the number of occurrences in the given position in corpus.

If the individual lines of a concordance list are evenly distributed within the whole corpus, the individual lines in the graph are of the same length and are displayed evenly along the whole window length. If, conversely, most lines are from “one” part of the corpus (e.g. from one document) there are distinctly more longer lines in one part of the window.

It is possible to “jump” to the selected part of the corpus by clicking on a line under the mouse cursor.

3.5 View

References

In this window you select which references should be displayed for the individual rows. If you choose *Token number*, a token number of the KWIC beginning is displayed. If you select the name of a tag (for example *doc*) the order number of the KWIC respective tag is displayed (for example *doc#2* means the KWIC is the second document from the beginning of the corpus). If you select the name of a tag attribute (for example *doc.file*), values of this attribute will be displayed, e.g. *doc.file=A03* means that the given KWIC is in a document where the attribute *file* has the value *A03*). References are displayed in green at the beginning of each row.

Attributes

In this window the user can select which attributes (e.g. lemma, tag) will be displayed.

Only in KWIC

Selected attributes will be displayed only for KWIC positions.

For all positions

Selected attributes will be displayed for KWIC positions and also for all positions in the displayed context.

Structures

Specifies which structure tags (structure attributes) will be displayed. If a structure tag contains attributes (e.g. sentence identifiers *id* of <*s*> tag), it is possible to tick the selected attribute. Then the values of this attribute will be displayed inside the relevant tag (e.g. <*s id=12/3*>).

Context

Specifies in which context the words will be displayed. A character, a position or an arbitrary structure tag can be a unit for both the left and the right side. If a character is a unit, whole words are displayed in such a way that at least the entered number of characters is displayed.

Range

Determines which lines will be displayed and the number of displayed lines. If the entered number of lines is 0, all rows are displayed.

Page up/Page down

The previous/following page is displayed. The number of rows per page can be changed (▷**Manager**▷**Options**), the default value is 20.

Jump to

Displays the specified range of lines (see ▷**View**▷**Range**) starting with the specified line. The number of the line from which the lines are displayed is shown in the status line before a ‘+’ sign.

3.6 Select

Selected lines can be stored in the clipboard or deleted from the concordance list (see ▷**Concordance**▷**Delete selected**). Selected lines are displayed against a blue background. The number of selected lines is shown in the status line at the bottom of the main window.

Lines can be selected with the mouse or the keyboard. The left mouse button or the space bar selects an unselected line or cancels the selection of a selected line. Shift+left mouse button selects all the lines between the given line and the current line.

All

Selects all lines

None

Cancels any selection

Invert

Selects non-selected lines and vice versa.

Insert to clipboard

Puts the selected lines in the clipboard for copying to other applications.

Export to CED

Passes selected lines to the corpus editor, CED. This function is available only on the UNIX platform.

3.7 Help

About

Shows a window with brief application information and the Manatee version number.

License

Shows a license window.

Show help hits

If selected and the mouse cursor stops for a while a short description is displayed for a GUI widget (button, entry-box etc.) under the mouse cursor.

Documentation

Launches on-line documentation in a web browser.

4 Queries

The user can enter a real query in the query language or a template (see below) into the query entry-box. A query consists of at least two parts: query type (selected from the menu in the upper-left corner) and query text or template (put into the first entry box).

A query type is one of the following:

New Query

creates a new concordance list for further modifications (filtering, reduction, sorting, ...). **Ctrl-Q** also selects this type.

P-filter

Positive filter – only matching lines remain in the concordance list. **Ctrl-P** selects P-filter.

N-filter

Negative filter – matching lines are removed from the concordance list. **Ctrl-N** selects N-filter.

collocation

Matching positions are highlighted in the given interval of each line.

P/N-filter and collocation query types need an interval within which to search matching positions for each line. The interval is given by its first and last positions.

First/Last

For collocations specifies which occurrence in the given interval should be considered a collocation.

From, To:

The start and the end of an interval for P/N filter, or for the search of collocation.

The number specifies a count of units. Positive numbers mean the count of units to the right, after the specified reference position. Negative numbers mean the count of units to the left, before the specified reference position. Number 0 means exactly the given position. A unit is, according to the selection, either a position or a structure tag. The last selection button *From* specifies the reference position from which the given number of units will be counted. A value can be chosen from the following:

<KWIC

– from the beginning of the words found

KWIC>

– from the end of the words found

<n.coll.

– from the beginning of the n-th collocation

n.coll.>

– from the end of the n-th collocation

Note: Every entered query is saved in the query history. It is not saved however, if the query itself and its name are identical with the previous query. It is possible to use cursor up/down arrows in the entry box of the query to go through the history.

If a query name is entered into the entry box following the query, the query is saved into the list of named queries.

The list of named queries, the list of templates and history are automatically saved when the application is closed. (▷**Manager**▷**Exit**)

5 Templates

The template is a kind of query which simplifies repeated queries of the same type. This means that a complicated query is created only once as a template. During various executions only values for the template need to be changed.

5.1 New template

A template consists of a name, a template text and an optional description. A name is a unique identifier of a template and is used when a query is invoked. All the three items are entered in ▷**Query**▷**New query**. The template text and the description can be later modified in the list of templates (▷**Query**▷**List of templates**).

In the template, text variables can be used. When the template is used, variables are substituted with real parameters. These variables consist of `$` character followed by a number. The first parameter will be then substituted for all occurrences of variable `$1`, the second parameter for all occurrences of variable `$2`, etc.

For example, a template for all word forms of a regular English verb could look like this:

```
[word="$1" | word="$1s" | word="$1ing" | word="$1ed"]
```

5.2 Using of templates

When a template is invoked, it is written into the same row as a usual query. It differs from a usual query in this way: the first character from the left is an exclamation mark (!), the template name follows, then a colon (:) and parameters divided by a space. If the name of the above template was *verb*, the query assignment would look like this:

```
!verb: help
```

6 Concordance

The concordance list displays the rows from the query result as set in ▷**View**▷**Range**. The individual KWIC are displayed one below the other. For navigating through the list the following keys and mouse actions are used:

Up/Down Arrows

Move one line up/down.

Page-Up/Page-Down

Move one page up/down.

Space

Toggles the current row selection.

Enter (Return), double-click

For the current row a larger context is displayed in the detail window.

Ctrl+Enter (Ctrl+Return), right-button-click

For the current row a full reference is displayed in the detail window.

Ctrl-Page-Up/Page-Down (arrows in the right bottom corner of the list)

New (not displayed) lines are displayed before/after displayed lines. If, in ▷**View**▷**Range**, the *Random* range of the displayed lines is selected, the user can change the range of the displayed lines with the Up or Down arrows to *First* or to *Last*.

F12

Switches the cursor (moves a focus) from the entry box of the query and the concordance list.

6.1 Groups

It is possible to assign a group number to each concordance line. Group numbers are from 1 to 99. Groups from 1 to 9 can be assigned by the respective keys. Groups from 10 to 99 are assigned by three key presses: first *Control-E* and then two keys with the respective digits. The *0* key cancels any group assignment.