

Manatee/Bonito – A Modular Corpus Manager

Pavel Rychlý

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

Abstract. A *corpus* is a large collection of texts in electronic form. *Corpus managers* are tools or sets of tools for coping with corpora. They can encode, query, and visualise texts. This paper describes widely used corpus manager Manatee that has many unique features: modular design, dynamic attributes, multi-values, multi-language support, almost unlimited corpus size and others.

The second part of the paper presents Bonito – the graphical user interface of the Manatee system. Other extensions of the system are also mentioned.

1 Introduction

Text Corpora play a crucial role in current linguistics. They provide empirical data for studies on how a language is used. Corpora are stored in computers but there are not many applications which can handle huge corpora of size in billions of tokens (word forms) which are being available in last years.

An ideal general-purpose corpus management tool should implement the following features:

text preparation – conversion from various formats, encodings, etc.;

metadata management – integration of the information about the source of data, authors, topics, genre, ...

tokenization – language-dependent determination of the elementary unit accessed, usually a word;

corpus annotation – potentially ambiguous, manual and automatic tagging on morphological, syntactic, semantic and pragmatic levels;

efficient corpus storage – the storage structures should enable fast retrieval of all stored data

concordancing – retrieving text snippets matching the user's query;

computation of statistics – searching for typical patterns in data, frequency distribution of various features, co-occurrence statistics, etc.

2 Manatee

The presented corpus management system Manatee is able to deal with extremely large corpora and is able to provide a platform for computing a wide

range of lexical statistics. It has all the above mentioned features and it is also language and tag-set/annotation independent.

The system is designed with the modular approach. There are an *indexing library* for compression, building and retrieving indexes; a *query evaluation module* with classes for different query operations, a *query parser* which transforms the queries into abstract syntactic trees, a set of *command line tools* for corpus building and maintenance, two *graphical user interfaces*.

The Manatee system is based on the text indexing library FinLib [1] that provides storage structures and retrieval procedures for corpus data based on an efficient implementation of inverted indexes, word compression, etc.

The system processes either the output of an external tokenizer or that of the simple internal one. There are two kinds of annotation that can be provided:

positional attributes adds linguistic information (like PoS tag, basic form) for each token,

structure annotation that denotes a structure in the text (e.g. a sentence, paragraph or document boundaries, noun or verb phrases).

There is also a special type of positional attributes – *dynamic attributes*. These are not store in the corpus explicitly, a function is declared in the corpus configuration. The function defines how to compute the value of the dynamic attribute for each particular token. There are three types of dynamic attribute usages:

1. transformation of tag values (e.g. displaying a full description of tag codes),
2. selection of partial information from an attribute (e.g. an attribute representing the gender is derived from a complex grammatical tag),
3. linking external information sources (e.g. linking a thesaurus database or morphological analyzer)

Selected attributes can be marked as ambiguous, it means the attribute can contain *multi-value* – a set of values. A user can ask for any of respective values to find such token.

The system is fully internationalized. The textual data can be stored in different character encodings including Unicode (UTF-8 [2]) and users can define local language settings for each corpus or even for each attribute. These settings are used during query evaluation and for locale sorting conventions.

Manatee implements a powerful query language. It enables searches given by restrictions on any positional attribute, on any meta information, or on any of their combinations. A given query can be further refined by means of positive or negative filters that are applied on the current result. Meta information can be used to create sub-corpora. The query language is an extension of the popular CQP [3].

3 Bonito

Bonito is a graphical user interface (GUI) of the 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.

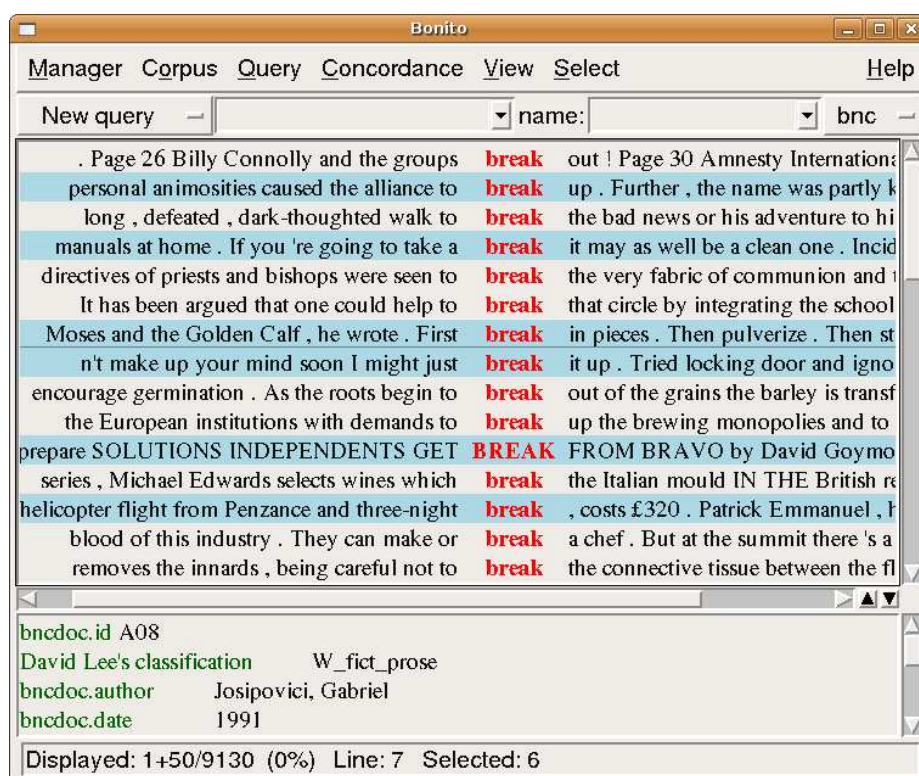


Fig. 1. Bonito application window

Bonito use the client/server architecture. Bonito is the client part, it is a standalone applications which runs on most systems (Unix + X Window, Windows 95+/NT/2000 ME/Vista, Macintosh and others with the Tcl/Tk support). The server part runs on a server (Unix, Windows, ...) where corpus data are stored. The client communicates to server over the Internet (TCP/IP) connection, the communication protocol is very lightweight, even a modem connection is sufficient for proper work. It is also possible to use "local" connection without Internet, in such case, client and server runs on a PC and corpus data are stored on a local disk.

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.

A concordance is one of the central objects of Bonito. Most of the Bonito's window area is formed by the concordance list where query results are displayed. An example of the Bonito window is Figure 1.

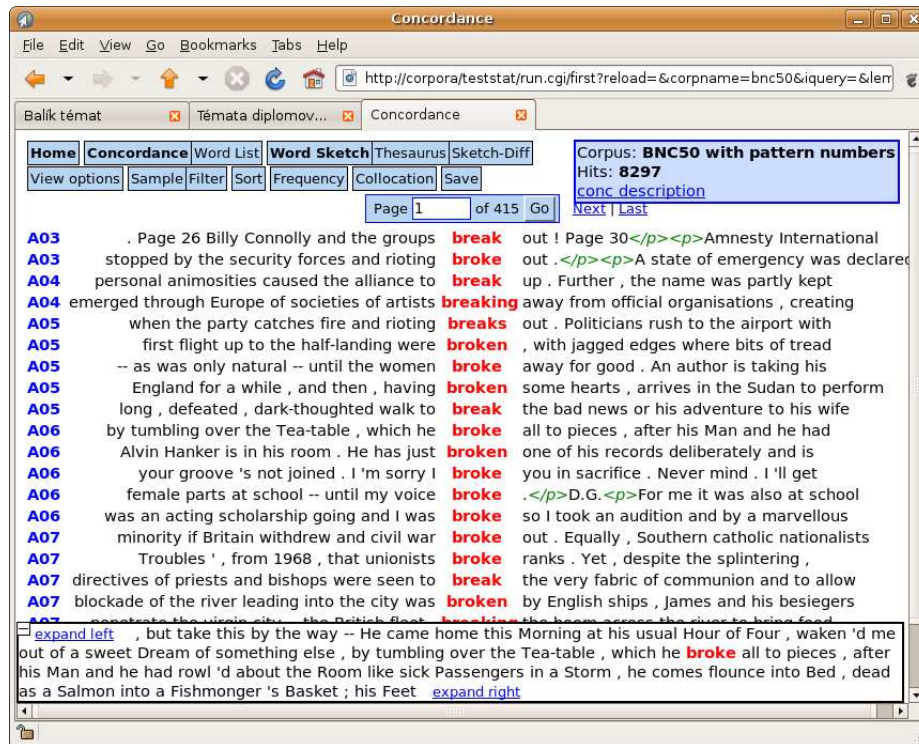


Fig. 2. Bonito2 concordance page

4 Bonito2

Bonito2 is a new GUI, it provides the same (or better) functionality as the older Bonito. The crucial difference is that it is a web application. Users do not need to install any client application, they use standard web browsers to access the interface. All regular web browsing techniques like cut&paste and bookmarks are available for users.

Web pages are generated by a CGI script on the web server. Standard web server authentication can be used to limit access. The known solutions of access and load control can be applied to the web server and the standard secured web protocol (https) can be used too. Finally, it is easy to connect Bonito2 to other sources and/or applications – the web pages can be read from other

applications, and links to external sources (a picture, a sound sample, a video) can be presented. Concordance view is displayed on Figure 2.

5 Sketch Engine

Because of the modular design of the system, it is easy to add more functions to the system or use the system as part of another application.

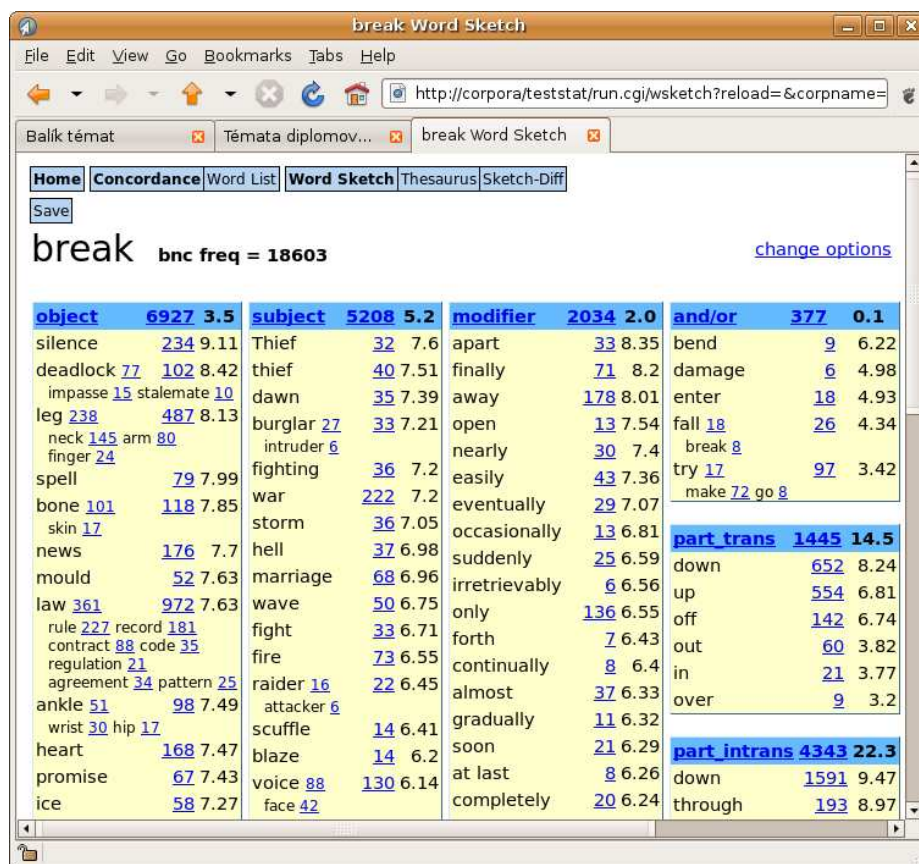


Fig. 3. Word Sketch of verb 'break'

The Sketch Engine [4] use the whole Manatee system and Bonito2 interface and provides word sketches, grammatical relations, and a distributional thesaurus as additions. A word sketch (see Figure 3) is a one-page, automatic, corpus-derived summary of a word's grammatical and collocational behaviour. Each word sketch contains direct links to concordances illustrating the listed collocation.

6 Conclusion

This paper describes main features of the Manatee corpus management system including graphical user interfaces Bonito and Bonito2. The system is in regular use in many research groups on universities around the world. There are also commercial companies (especially publishers) which use the system or its parts or extensions in day-to-day works.

Acknowledgements

This work has been partly supported by the Academy of Sciences of Czech Republic under the projects T200610406, T100300419 and by the Ministry of Education of CR within the Centre of basic research LC536 and National Research Programme 2C06009.

References

1. Rychlý, P.: Corpus managers and their effective implementation. Ph.D. thesis, Faculty of Informatics, Masaryk University (2000).
2. Yergeau, F.: RFC2279: UTF-8, a transformation format of ISO 10646. Internet RFCs (1998).
3. Schulze, B.M., Christ, O.: The CQP User's Manual. (1996).
4. Kilgarriff, A., Rychlý, P., Smrž, P., Tugwell, D.: The Sketch Engine. Proceedings of Euralex (2004) 105–116.