Software Demonstration: **GCQP – Multiplatform Graphical User Interface to the CQP corpus manager**

Pavel RYCHLÝ, Brno, Czech Republic

Abstract

This software demonstration presents a new user interface to the corpus manager CQP. The key feature of this system is the client/server architecture. The server is a wrapper of CQP, it runs on a UNIX system and uses CQP commands for the corpus query evaluation. There are also some new features like the generic sort of concordance lines, the multilevel frequency distribution or the collocations identification based on MI-score a and T-score.

The client (GCQP) is a graphical user interface and it is possible to run it on UNIX with X Window System, MS Windows 95/98/NT or Macintosh systems. It communicates with the server over Internet. GCQP makes all features of CQP and server enhancements available to users in a friendly environment.

1 Introduction

1.1 IMS Corpus Workbench

Text corpora represent one of the main sources of information for computational linguists and lexicographers. The IMS Corpus Workbench is a set of tools for the management of large text corpora and retrieval of information from these corpora.

Within the workbench, a corpus is represented as a sequence of *positions*. Every position is divided into a set of *positional attributes* and each attribute contains a character information. One of the positional attributes (*word*) represents the particular word form at each position, other attributes are used for corpus annotation (lemma, grammatical tag, etc.). We can also store some *structural tags* in the corpus. They are used for structural annotation of the texts in the corpus: document, paragraph and/or sentence boundaries, annotation of headings, etc. There are no restrictions on the number of positional attributes and structural tags per corpus.

The principal tool of the workbench is CQP (which stands for Corpus Query Processor) [Schulze and Christ, 1996], which evaluates given queries and returns the result on the screen or to another output that can be used in further processing. CQP defines a very powerful query language. To name just the basic features, users can use Boolean combinations of regular expressions over attribute values, regular expressions over positions, boundaries of structures, labels of positions and its references, built-in functions, subcorpora queries. The CQP query language (although it may be quite difficult for beginners or non-expert users) is mostly the reason why to prefer IMS Corpus Workbench among available corpus managers.

For more comfortable interaction or presentation there is XKWIC [Christ, 1995] – a graphical user interface running in X Window System – execution of which is unfortunately limited to a few clones of UNIX systems with X Window and the initial configuration is not easy. This is the fundamental drawback of the whole system: the user friendly environment is available only for users with UNIX workstation and direct connection (by local network) to the computer with the installed workbench.

1.2 New Graphical User Interface

The new graphical user interface (GCQP) was developed for the Institute of the Czech National Corpus in Prague and the Natural Language Processing Laboratory at our faculty. GCQP provides an interface to the CQP, all its functions, and adds several new features. New system is based on the client/server architecture. The server is a wrapper of CQP, it runs on a UNIX system and the client part is platform independent.

In the following sections we describe the system in more details.

2 Hardware and Software Requirements

The server part of the system (cqsd) is written mainly in C++. Only a small subsystem of users' accounts management is written in Perl. Because the server depends on CQP (it calls cqp command for query evaluation, CQP C-API library (libcu) is used for the physical corpus access) it is possible to run it only on systems supported by CQP. Today, there are two variants of UNIX the CQP runs on: Solaris and Linux. The server is provided for both platforms.

Hardware requirements for the server are the same as for the CQP: Sun server or workstation with a SPARC processor, or a regular PC with Intel Pentium, at least 32MB of memory are recommended. Disk space requirements depend on the size of installed corpus or corpora and their annotation (the number of different positional attributes, the number of structural tags). A small corpora like Brown corpus (1 mil. tokens, [Francis and Kučera, 1979]) need only a few tens of MB of disk space, a large annotated corpus with 100 mil. tokens can take several GB of disk space. Very large corpora also need more physical memory for fast query evaluation.

GCQP itself (the client part of the system) is written in pure Tcl/Tk [Ousterhout, 1994], which ensures the platform independence. It runs on UNIX with X Window System, MS Windows 95/98/NT or Macintosh systems. There are no special hardware requirements: wherever you can run any of the mentioned operation system you can run GCQP even for very large corpora.

The communication between GCQP and the server is based on the standard TCP/IP (Internet) protocol. Thus, both sides need a connection to the Internet or an intranet. The protocol is optimized for the amount of transmitted data, therefore even a slow modem connection is sufficient for a proper usage of GCQP. It is also possible to use the system on a stand-alone workstation where both client and server runs.

3 Client/Server Design

There was one required feature in the client/server design of the system: to enable users of the corpus to use their low-end computers with their familiar operation systems. This led us to the selection of Tcl/Tk as a programing language for the client and the decision that all time-consuming tasks should be performed on the server. The client is responsible only for a navigation of users during query construction and presentation of the results received from server.

For each client connection to the server computer a new server process is started. All these processes share corpus data. A simple multitasking environment was implemented, so that several demanding tasks on the server side, like a query evaluation or a computation of frequency distributions or collocations, are processed at the background. It means that client is not blocked during such tasks.

3.1 Communication Protocol

The communication protocol between client and server is designed in the UNIX and Internet style. The whole communication is character based. A client transmits commands, the server responds by results or control data. There are more than 40 build-in commands and a special one which enables an execution of external programs. It is possible to extend server capabilities in this way without recompilation of the server.

The set of implemented commands is not limited to the graphical user interface of a corpus manager. For example, there is a small Tcl script `gcqp-count` which evaluates queries given to a corpus and prints the number of lines in the result of each query. This program can be run in batch mode without any user interaction.

3.2 User Access Restrictions

Users have login names and passwords and they have to authenticate at the beginning of each session (GCQP startup). There is a possibility to restrict user access to corpora in several ways:

- Groups of users (or each user) have different lists of available corpora.
- Users can access server only from selected computers (or sites).
- A “hardcut” (the maximum number of lines in a concordance list) can be assigned to a user. A corpus administrator can use this option to reduce server load.

4 New Functions

In this section we focus our attention on the new functions of GCQPwhich are not available in CQP.

4.1 Queries

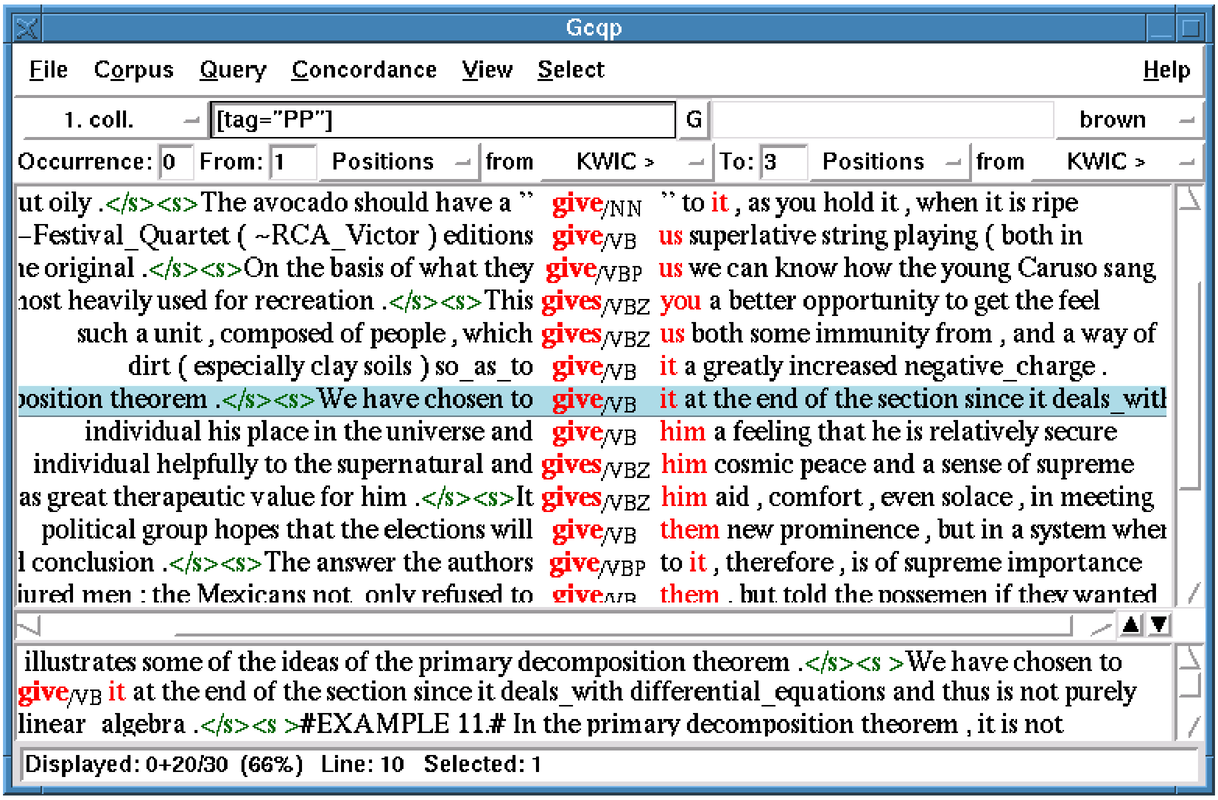
Because queries are evaluated through `cqp` commands, all forms of the CQP query language are supported in GCQP. There are two small enhancements that save much users’ time and many users like GCQP mostly for this feature:

- The most simple (but also the most frequent) type of queries – query on a single word form – can be entered without quotes.

- All queries are stored in the history and access to this history is provided through a single key-press. The query history is automatically saved on exit and restored on a new start-up of GCQP. The number of queries in the history is configurable.

P/N filter A positive/negative filters can be used to reduce number of lines in a concordance. Lines matching given condition in a given span will remain in concordance list (P-filter) or will be deleted from the concordance list (N-filter). Conditions are regular CQP queries and spans are entered in the terms of positions or structural tags.

Collocations Users can find and highlight words in context (within given span). It works like `set collocate` in CQP but there could be more than one collocation (there is no limit to the number of collocations in a concordance list) and the collocation can occupy more than one position (it could be a *sequence* of words). The following Figure 1 shows the main window of GCQP with highlighted collocations.



1. coll. [tag="PP"] G brown

Occurrence: 0 From: 1 Positions from KWIC > To: 3 Positions from KWIC >

```
ut oily .</s><s>The avocado should have a "  give/NN  " to it , as you hold it , when it is ripe
–Festival_Quartet ( ~RCA_Victor ) editions  give/VB  us superlative string playing ( both in
e original .</s><s>On the basis of what they  give/VBP  us we can know how the young Caruso sang
ost heavily used for recreation .</s><s>This  gives/VBZ  you a better opportunity to get the feel
     such a unit , composed of people , which  gives/VBZ  us both some immunity from , and a way of
          dirt ( especially clay soils ) so_as_to  give/VB  it a greatly increased negative_charge .
osition theorem .</s><s>We have chosen to  give/VB  it at the end of the section since it deals_witl
      individual his place in the universe and  give/VB  him a feeling that he is relatively secure
  individual helpfully to the supernatural and  gives/VBZ  him cosmic peace and a sense of supreme
as great therapeutic value for him .</s><s>It  gives/VBZ  him aid , comfort , even solace , in meeting
   political group hopes that the elections will  give/VB  them new prominence , but in a system wher
l conclusion .</s><s>The answer the authors  give/VBP  to it , therefore , is of supreme importance
iured men : the Mexicans not only refused to  give/VB  them , but told the possemen if they wanted
```

```
illustrates some of the ideas of the primary decomposition theorem .</s><s >We have chosen to
give/VB it at the end of the section since it deals_with differential_equations and thus is not purely
linear  algebra .</s><s >#EXAMPLE 11.# In the primary decomposition theorem , it is not
```

Displayed: 0+20/30 (66%) Line: 10 Selected: 1

Figure 1: The main window of GCQP.

Frequently asked queries A list of favourite queries for immediate use is stored. Users do not need to browse their query history or enter a complex query. They just select an item from the prepared annotated list. This is also a useful tool for corpus presentations.

Templates Users with greater experience may use selected query types regularly. Such queries can be rather complex with only small alternations. It is possible to locate a similar query in the history or in the list of the favourite queries and edit it, but there is a simpler solution – query templates. Users can define a template of the query with special "variables" inserted. Then they can enter only the name of the template and parameters which are automatically substituted for variables.

For example, we can define a template for query of two words, first word given by a word form, second given by a grammatical tag. The second word can occur in different distances from the first one. In the following template (named `wt`) `$AAA`, `$BBB` and `$CCC` represent variables:

```
wt:     MU (meet "$AAA" [tag="$BBB"] 1 $CCC)
```

Then we can search for words *give* or *gives* (regular expression `gives?`) followed by a personal pronoun (`PP` tag) in the distance at most 3 words:

```
!wt: gives? PP 3
```

The Figure 1 shows the result of this query.

4.2 Concordance Lists

GCQP supports all common operations with concordance lists: save the list into a text file, print the list on a printer, delete selected lines, reduce the number of lines in the list (according to the given percentage or the number of lines), simple sort (according to keywords, left or right context). There is a possibility of more generic sort with more than one sorting rule. GCQP also offers the computation of the generic frequency distribution with many levels of grouping and subtotals.

Undo Each modification of a concordance list can be taken back. The number of undo levels is configurable.

Collocation candidates GCQP can produce a list of words with the highest MI-score or T-score in a given span. The table of results contains also the relative and absolute frequency for each word. An example for the word *night* in Brown corpus is displayed in Table 1.

Distribution overview A graphical overview displays in what parts of the corpus the concordance lines occur. An example of such overview for the word *surface* in Brown corpus is given in Figure 2.

4.3 Other Features

The tool is language independent (any set of tags can be used) similarly as with CQP. However two Czech specific features were implemented: Concordance lines can be sorted according to Czech sort rules and the output (save/print) is provided in selectable character encoding. It is

Word	MI-score	T-score	Rel. freq [%]	Abs. freq
tomorrow	9.467	2.233	33.33	5
Sunday	8.66	1.995	19.05	4
Saturday	8.66	1.995	19.05	4
previous	8.529	1.995	17.39	4
Monday	8.178	1.726	13.64	3
last	7.895	3.59	11.21	13
at	4.541	2.871	1.096	9
that	3.203	2.674	0.4337	9
The	2.413	1.624	0.2508	4
the	1.322	2.324	0.1177	15

Table 1: Collocation candidates of the word *night* in Brown corpus.

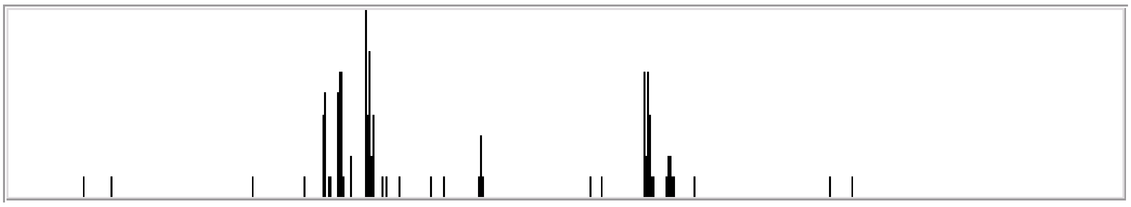


Figure 2: Distribution overview for the word *surface* in Brown corpus.

possible to add other character encoding tables and change sorting tables depending on the selected language.

At the begining, there was only Czech version of GCQP. Now, there is also English variant and a new language localization is very easy, it can be achieved by the translation of two configuration files.

References

[Christ, 1995] Christ, O. (1995). *The XKWIC User Manual.*

[Francis and Kučera, 1979] Francis, W. N. and Kučera, H. (1979). *Brown Corpus Manual.* Brown University, Providence, Rhode Island, revised and amplified edition.

[Ousterhout, 1994] Ousterhout, J. K. (1994). *Tcl and Tk Toolkit.* Addison-Wesley.

[Schulze and Christ, 1996] Schulze, B. M. and Christ, O. (1996). *The CQP User's Manual.*