From DOC Files to a Modern Online Dictionary

Tinatin Margalitadze, George Keretchashvili

Lexicographic Centre at Ivane Javakhishvili Tbilisi State University
Address: 1 Chavchavadze av. Tbilisi, Georgia
E-mail: tinatin@margaliti.ge, contact@dictionary.ge

Abstract

The aim of this paper is to describe the process of development of software for the Comprehensive English-Georgian Online Dictionary, posted on the Internet in 2010. The Dictionary engine is built on PHP/MySQL platform and combines three major branches: user interface, administrative interface and billing system, thus making it an integrated and dynamic resource. User functionalities include: bidirectional search; auto suggestions; auto corrections; online payments, etc. The administrative interface of the Dictionary holds a number of administrative functionalities, such as: dictionary vocabulary management functionalities; generation and conversion tools necessary for editors; user registration management functionalities, etc.
The Online Dictionary databases were generated from the DOC files which contained raw text data: words, grammatical characteristics of words, pronunciations and descriptions, altogether and separated by spaces just as in any sentence. After thorough analysis and testing, a special converter was written that would automatically analyze and separate raw data input into separate rows and fields. Our experience of transformation of the DOC files into a modern online resource may be interesting for the e-lexicography community. This paper will also discuss some other applications which are under development at the Lexicographic Centre.

Keywords: data transformation; online dictionary development; control panel.

1. History of the Dictionary

Work on the Comprehensive English-Georgian Dictionary (CEGD) began in the 1960s in the Department of English Philology of Tbilisi State University. In the 1980s, a small team of editors embarked on a thorough revision of the dictionary material and launched publication of the dictionary in fascicles (1995–2012). Currently printed and published are 14 fascicles of the English-Georgian dictionary (www.margaliti.ge), which cover 2,380 pages of the printed dictionary. The online version of the dictionary, posted on the Internet in 2010, is based on the aforementioned fascicles (www.dict.ge). The CEGD comprises 110,000 entries, covering several hundred thousand English meanings, collocations, phrasal verbs, idioms, and terms from different fields (T. Margalitadze, 2012).

One of the important issues faced by the editors of the CEGD has been 'linguistic and cultural anisomorphism' (Hartmann and James 1998: 51) between the English and Georgian languages, resulting in semantic asymmetry of seemingly similar words of these languages. Semantic asymmetry is even wider between genetically unrelated

and structurally different languages, as is the case with the Georgian and English languages. English-Georgian lexicography is not exceptional in this respect, as it is the central problem of bilingual lexicography at large. This issue, and the treatment of equivalence in the CEGD, was presented at the XV International Congress of EURALEX in Oslo (T. Margalitadze, 2012).

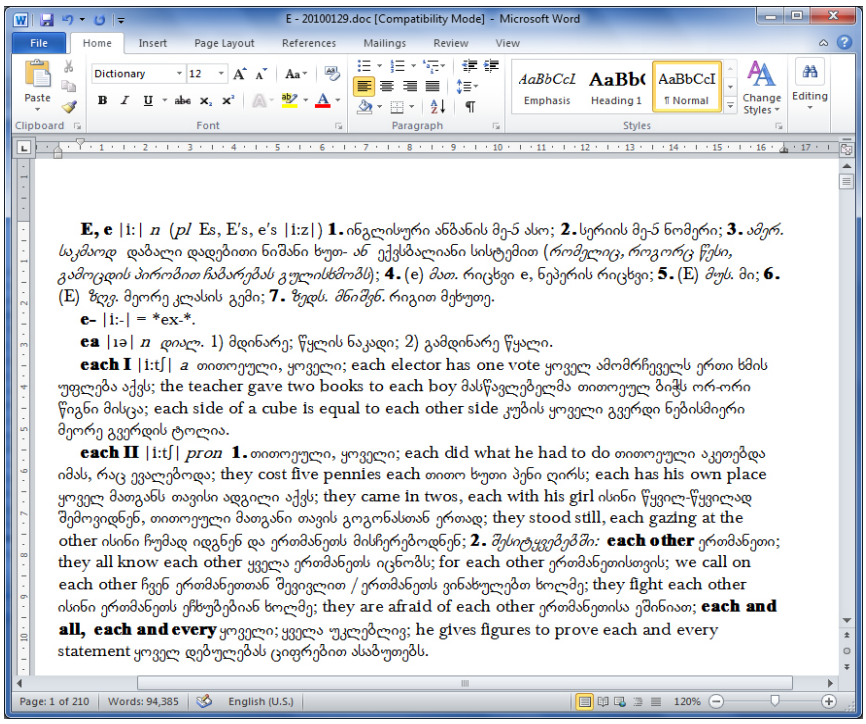

Figure 1: Text represented in MS Word document

The CEGD was not created in a Dictionary Writing System (DWS). In the 1990s, half of the dictionary, the compiled and edited entries, existed on cards (letters A–L). In 1993, the Lexicographic Centre started digitalization of the dictionary material and the first fascicle, the letter A, appeared in 1995. Back in the 1990s, there was not even a proper Georgian font with extended character support and a special font ("Dictionary", see Figure 1) was created for the project. It is probably worth noting that the configuration of the Dictionary font was based on the Russian script, "Cyrillic", changed into the Latin script several years later.

Dictionary cards were digitalized into the DOC files and in subsequent years the work continued in MS Word.

2. Data Transformation

As mentioned above, digitalized dictionary material, as well as the entries created later, existed in a formatted text edited by text processors like MS Word (see Figure 1).

The DOC files contained raw text data that included words, grammatical characteristics of words, and pronunciations and descriptions, altogether and separated only by spaces just as any sentence. The text was represented in a special, non-Unicode encoding and was slightly formatted (see Figure 1).

```
<p class=MsoNormal style='text-indent:14.2pt'><b><span style='font-family:DictionaryBold'>E,↓
e </span></b><span style='font-family:Dictionary'>|i:|<i> n  </i>(<i>pl  </i>Es,↓
E’s, e’s |i:z|) </span><b><span style='font-family:DictionaryBold'>1.</span></b><span↓
style='font-family:Dictionary'> ÈÌÂÊÈÑÓÞÈ ÀÌÁÀÌÈÑ ËÄ-5 ÀÑÍ;</span><b><span↓
style='font-family:DictionaryBold'> 2</span></b><b><span style='font-family:↓
DictionaryBold'>.</span></b><span style='font-family:Dictionary'> ÑÄÐÈÈÑ ËÄ-5↓
ÌÍÉËÄÐÈ; </span><b><span style='font-family:DictionaryBold'>3</span></b><b><span↓
style='font-family:DictionaryBold'>.</span></b><span style='font-family:Dictionary'>↓
<i>ÀËÄÐ. </i><i>ÑÀÉËÀÍÃ</i>  ÃÀÁÀÊÈ ÃÀÃÀÁÈÇÈ ÌÈØÀÌÈ ÞÓÇ- <i>ÀÌ  </i>ÄÕÅÑÁÀÊÈÀÌÈ↓
ÑÈÑÒÄËÈÇ (<i>ÐÍËÄÊÈÚ, ÐÍÂÍÐÚ ÜÄÑÈ, ÂÀËÍÚÃÈÑ ÎÈÐÍÁÈÇ ÙÁÁÀÐÄÁÁÀÑ ÂÓÊÈÑÞËÍÁÑ</i>); </span><b><span↓
style='font-family:DictionaryBold'>4</span></b><b><span style='font-family:↓
DictionaryBold'>.</span></b><span style='font-family:Dictionary'> (e)<i> ËÀÇ.</i>↓
ÐÈÚÞÅÈ e, ÌÄÎÄÐÈÑ ÐÈÚÞÅÈ; </span><b><span style='font-family:DictionaryBold'>5</span></b><b><span↓
style='font-family:DictionaryBold'>.</span></b><span style='font-family:Dictionary'>↓
(E) <i>ËÓÑ</i>. ËÈ; </span><b><span style='font-family:DictionaryBold'>6</span></b><b><span↓
style='font-family:DictionaryBold'>.</span></b><span style='font-family:Dictionary'>↓
(E) <i>ÆÖÅ</i>. ËÄÍÐÄ ÉÊÀÑÈÑ ÂÄËÈ; </span><b><span style='font-family:DictionaryBold'>7.</span></b><i><span↓
style='font-family:Dictionary'> ÆÄÃÑ. ËÌÈØÅÌ.</span></i><span style='font-family:↓
Dictionary'> ÐÊÂÈÇ ËÄÞÓÇÄ.</span></p>↓
```

Figure 2: Data represented in HTML format

"E, e"	"n"	"i:"	"(pl Es, E\'s, e\'s [i:z]) <p>1. ინგლისური ანბანის მე-5 ასო; <p>2. სერიის მე-5 ნომერი; <p>3. ამერ. საკმაოდ დაბალი დადებითი ნიშანი ხუთ- ან ექვსბალიანი სისტემით (რომელიც, როგორც წესი, გამოცდის პირობით ჩაბარებას გულისხმობს); <p>4. (e) მათ. რიცხვი e, ნეპერის რიცხვი; <p>5. (E) ფუს. მი; <p>6. (E) ზღვ. მეორე კლასის გემი; <p>7. ზედს. მნიშვნ. რიგით მეხუთე."
"e-"	""	"i:-"	"= <r>ex-</r>."
"ea"	"n"	"iə"	"დიალ. 1) მდინარე; წყლის ნაკადი; <p>2) გამდინარე წყალი."

Figure 3: Data ready for saving as Comma Separated Values (CSV) format

After thorough analysis and testing, a special converter was written that would automatically analyze and separate raw data input into separate rows and fields. Before dictionary data can be used for the database, the following procedures should be performed:

- A DOC file is prepared by replacing a couple of special symbols presented in the texts by other special symbols in order to be further interpreted as required;
- Then the file is converted into an HTML file, thus converting the initial text

into the data that can be parsed by converter (see Figure 2);

- The HTML file is slightly cleaned manually and submitted for conversion;
- Converter runs through the file structure and indicates errors if found;
- After the errors have been corrected, the converter parses the file and makes all necessary conversions that might include more than 20 conversions for each word set;
- Then the data is split into different fields, and special formatting is applied which outputs it in the CSV (Comma Separated Values) format (see Figures 3 and 4);
- After the CSV file is generated it can be imported into any database.

Even after inserting data into the database, several scripts are run over newly-inserted records in order to achieve the database consistency and to provide efficient search results. Final data can be later directly edited through the Dictionary Control Panel.

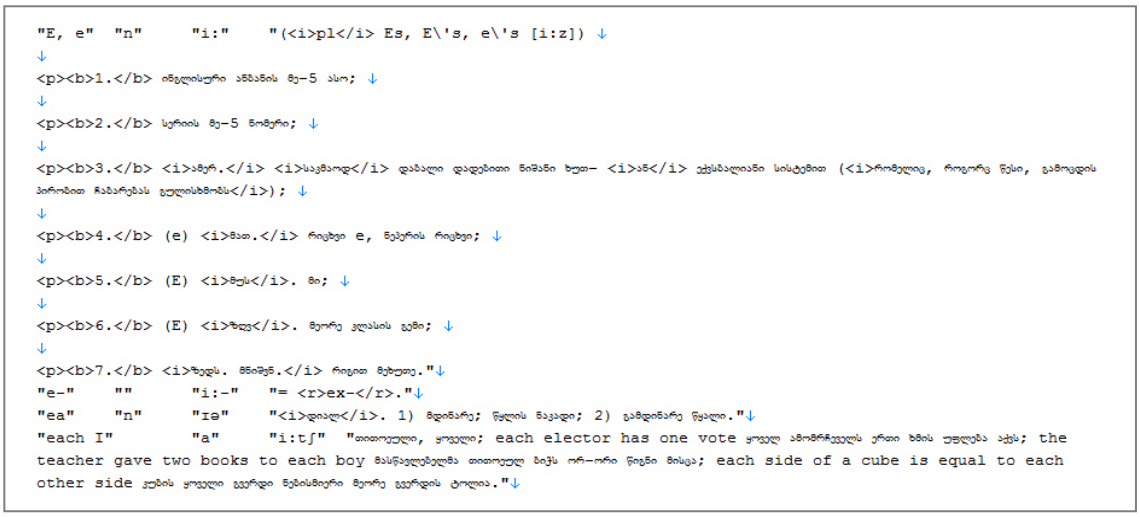

Figure 4: Data in Comma Separated Values (CSV) format

3. Online Dictionary

The Comprehensive English-Georgian Online Dictionary (CEGD) is a unique, hand-written web based application easily accessible from any Internet-enabled device, such as computers, cell phones, tablets etc. The Dictionary engine is built on PHP/MySQL platform and combines three major branches: user interface, administrative interface and billing system, thus making it an integrated and dynamic resource (see Figure 5).

During the first year of the operation some new functionalities were added to the program: the user interface became bilingual, a drop-down bilingual suggestion feature was added to the search box, an auto correction/suggestion system was

implemented to correct typos, search backend was improved, entry layouts were improved for easier reading, colors and tooltips were implemented for abridgements, video tutorials were added to the user guide, etc. There is an online feedback form available to provide support for users with technical or other issues.

Figure 5: CEGD

Both the database and the engine of the CEGD are in the process of constant upgrading and improvement in order to provide the users with an up-to-date, user-friendly, safe and perfect product.

3.1 User functionalities

The bilingual user interface front- and backends hold two categories of functionalities. One combines generic system screens and functionalities like user registrations, profile editing functions, safe logins, password resets, news etc.

The other part of the system is responsible for bidirectional search (the engine includes the search functions that make it possible to look up both English and Georgian words and phrases despite the fact that the dictionary vocabulary database is one way: English to Georgian only); auto suggestions; the search engine also includes auto suggestions on spelling errors, etc. While viewing any particular word and its translation, next, previous and several nearby wordlists appear for easier navigation; words can be listed and navigated by letters, etc.

Though all the interfaces and functionalities were designed to be intuitive and easy to use, the CEGD is supplied with a user's guide with detailed textual and video instructions on how to search for English words, collocations, phrasal verbs, and idioms, as well as Georgian words and phrases. The user guide also explains the structure and organization of the entries and other details.

3.2 Administrative functionalities

Administrative screens and functions are designated for editors, managers, technical administrators and other personnel who support online dictionary operations.

The following functionalities are available through the CEGD Control Panel:

- Dictionary vocabulary management, including the functions of viewing and editing the dictionary vocabulary (see Figure 6), as well as the function of adding new entries;
- Generation and conversion tools necessary for editors (see Figure 7);
- User registration management;
- Statistics including registrations, search logs, etc.

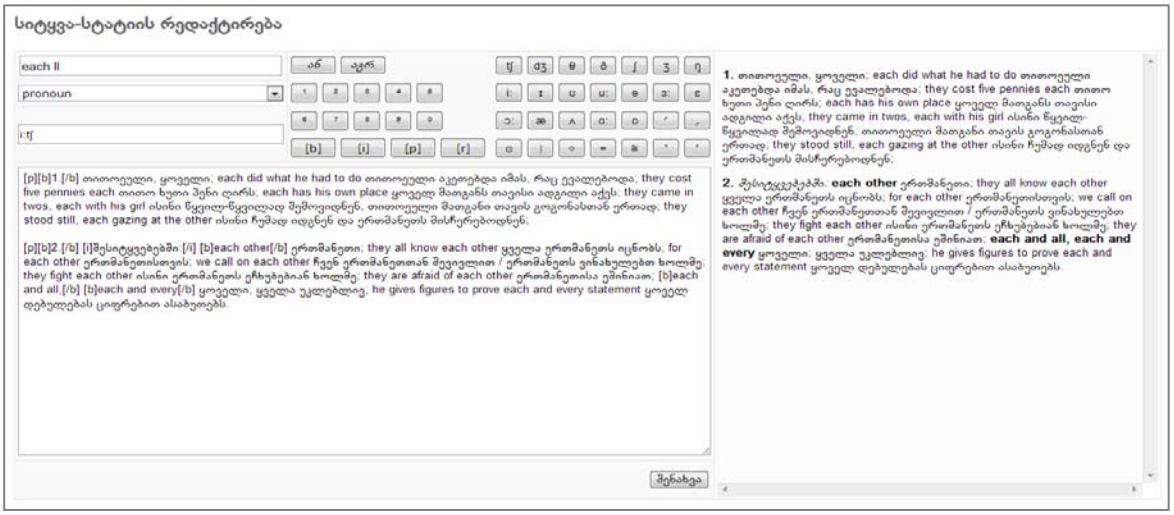

Figure 6: Editing an entry from CEGD Control Panel

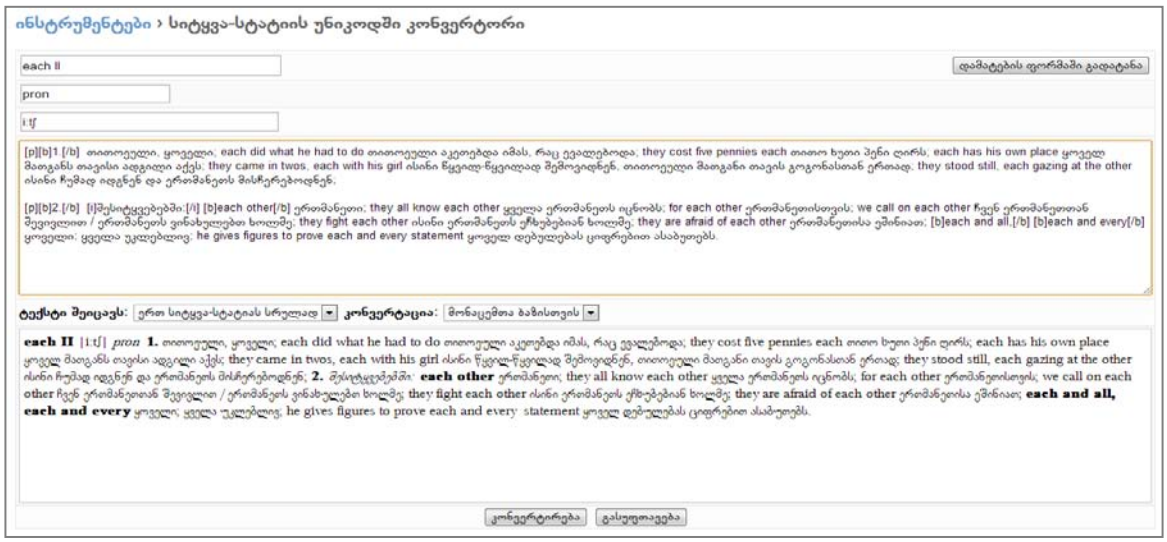

Figure 7: Converter for editors

4. Specialized Dictionaries

After its successful launch, as a result of expertise gained over more than a year of operation of the CEGD, and based on accumulated experience including user feedback, a number of improvements were applied to the core engine of the CEGD: backend search functions and database extension tables were redesigned and rewritten to provide improved performance. Frontend and search result pages were also slightly modified and a clean, light version of the core engine was used for smaller specialized dictionaries of the Lexicographic Centre, namely for the "English-Georgian Military Dictionary" (http://mil.dict.ge), compiled in 2009 at the request of the Georgian Ministry of Defense and posted on the Internet in 2011 (see Figure 8), and the "English-Georgian Biology Dictionary" (http://bio.dict.ge), the current project of the Lexicographic Centre, financed by Shota Rustaveli National Science Foundation of Georgia.

Figure 8: English-Georgian Military Dictionary

Light versions of the Online Dictionary Application operate in the same way as the CEGD system.

5. Future software projects

Currently the Lexicographic Centre is working on the development and improvement of its web applications, and also on the development of new software tools and solutions for the projects of the Centre.

5.1 Desktop Application

Web applications are very common nowadays in this country. However, there are cases where web application is not the right solution and the user prefers a locally installed desktop application. This fact led to our decision to develop a desktop application for online dictionaries. Work on the first version of the electronic dictionary, i.e. the desktop application is already completed and is being tested (see Figure 9).

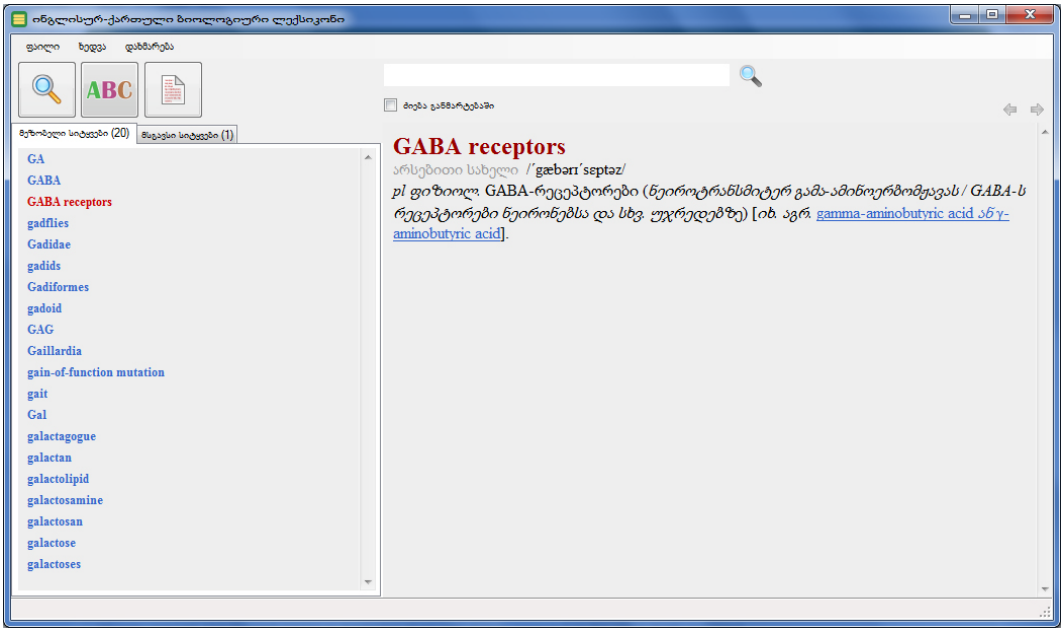

Figure 9: Desktop application (Pre-alpha version)

Desktop application includes predominantly the same functionalities as online dictionaries of the Lexicographic Centre, described above, but will be easier to access and use even in the case of slow or bad Internet connection. Unlike web applications, desktop applications will be integrated into the user's computer and will enable the addition of the functionality of directly translating words from many other applications (like word processing applications) by simple clicks or using keyboard combinations. Being offline does not mean being outdated: the desktop application database will have the functionality of being updated from the Internet, as the Lexicographic Centre regularly releases new updates of its online databases.

5.2 Dictionary writing system

As mentioned above, dictionary creating processes were conducted in the Lexicographic Centre with very limited technical resources, which required much effort to work on the data in the past. Nowadays, modern technologies offer more options and possibilities to maximize results and add more functionality and manipulation options to the dictionary data. As it was becoming more and more difficult and uncomfortable to handle Word files, the Lexicographic Centre has

launched the development of a Dictionary writing system. When this project is completed and the existing dictionaries are integrated into it, this will allow the Lexicographic Centre to add to its products synonyms, antonyms, and pictures, to apply different fonts and colors, as well as adding other functions essential to modern dictionary databases.

5.3 Mobile Application

Modern mobile devices like smartphones and tablets are becoming more and more popular in this country and are essential for students and business people, etc. In order to bring comfort and simplicity to those users, the creation of special applications are planned in order to meet mobile device requirements.

5.4 Lightweight interface of online dictionaries

Some mobile users prefer websites instead of downloading and installing applications on smartphones or tablets. As mobile devices are usually smaller in size and have limited interaction options compared to personal computers, the creation of lightweight interfaces, specially designed for mobile use are planned at the Lexicographic Centre.

6. References

Comprehensive English-Georgian Online Dictionary. (2010). T. Margalitadze (Editor-in-Chief) et al. Lexicographic Centre at Tbilisi State University, Tbilisi. http://www.dict.ge

English-Georgian Online Biology Dictionary. (2012). T. Margalitadze (Editor) et al. Lexicographic Centre at Tbilisi State University, Tbilisi. bio.dict.ge.

English-Georgian Online Military Dictionary. (2011). T. Margalitadze (Editor) et al. Lexicographic Centre at Tbilisi State University, Tbilisi. mil.dict.ge.

Hartmann, R. R. K. and G. James. (1998). Dictionary of Lexicography. London: Routledge.

Margalitadze T. (2012). The Comprehensive English-Georgian Online Dictionary: Methods, Principles, Modern Technologies. Proceedings of the XV EURALEX International Congress. Oslo, Norway. http://www.euralex.org/elx_proceedings/Euralex2012/pp764-770 Margalitadze.pdf