Lexicographic Potential of the Georgian Dialect Corpus

Marina Beridze, David Nadaraia
Ivane Javakhishvili Tbilisi State University,
Arn. Chikobava Institute of Linguistics
e-mail: marine.beridze@gmail.com

Abstract

The project *Linguistic Portrait of Georgia* envisages various aspects of documentation of Georgian linguistic reality by means of corpus methodologies. This title is an umbrella for three large-scale projects within the framework of which *The Georgian Dialect Corpus – GDC* (http://corpora.co) was developed.
Presently, the architecture and text base of the corpus have been designed, being permanently developed and updated. Besides, the lexicographic base of the corpus is organized, agglomerating data from printed dialect dictionaries.
The lexical stock of the corpus is presented based on text, lexicographic and encyclopaedic data.
The total quantity of tokens in the corpus is estimated to be up to 2 000 000, while the lexicographic base has 60 000 items (lemmas with entries) by now; this quantity is considerably increased owing to phonetic and grammatical variations, frequently associated with a single lexical item.
Keywords: Georgian Dialect Corpus; lexicographic base; encyclopaedic data

1 Georgian Dialect Corpus: General Characteristics

The Georgian Dialect Corpus (GDC) was developed as a means for documentation and study of Georgian linguistic diversity (Beridze, Nadaraia 2009: 25; Beridze, Lordkipanidze, Nadaraia: 2015-1, 323). Consisting of several components, this is a reference resource, which, with respect to its interdisciplinary objectives, considerably differs from other similar corpora. The principal difference is that GDC has corpus and library modes of text access; it integrates the lexicographic base, a prospective universal lexicographic base of the Georgian language.
Currently, GDC consists of the following user platforms:

- text corpus proper
- text library
- online dictionaries of GDC

1.1 Text corpus of GDC

Comprehensive dialect text body was prepared and uploaded to the corpus, including: dialect texts published after the 1920s; materials (manuscript archive, audio archive) of various dialectological expeditions in the 20th century; materials (hundreds of hour-long digital video and audio archives) collected during fieldwork organized within the framework of our projects both in Georgia and internationally (Iran, Turkey, Azerbaijan). Extensive activities have been carried out for the sake of philological processing, unification, and corpus integration of the texts.
GDC is equipped with the text meta-annotation system, represented by several blocks of features:

a) Block of linguistic data (language, dialect, sub-dialect)
b) Block of text features (topic, chronotope, place of recording, publication, etc.)
c) Block of data about a recorder and editor of texts

d) Block of data about a narrator and his/her family, which, alongside with ordinary biographical and social data, comprises information about migration processes.

The design of the corpus enables to integrate data of languages (primarily, of Kartvelian languages) other than Georgian (currently, Laz text data are being integrated).
By means of the lexicographic editor, the design of the corpus enables to integrate non-textual lexical data (Beridze, Nadaraia, Lordkipanidze 2015-2: 82). Presently, the user interface of the corpus allows of the following kinds of queries:

- by a complete word or its fragment (word beginning, word ending, any part of word)
- by part of speech
- by a grammatical marker
- by language, dialect, place of recording, narrator, thematic feature, chronotope, and other meta-features.

Combined query results according to linguistic markers and meta-features in GDC allow to:

- view word-list
- view contexts
- view full text.

(Examples are given in the Figures 1, 2, 3 below)

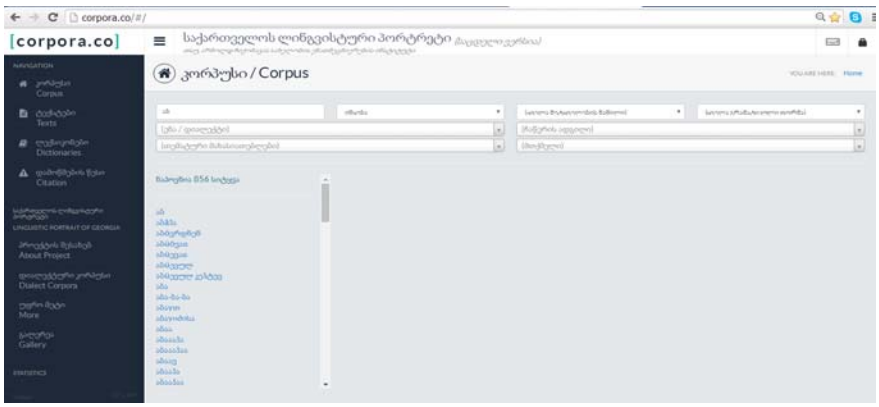

Figure 1: Word-list.

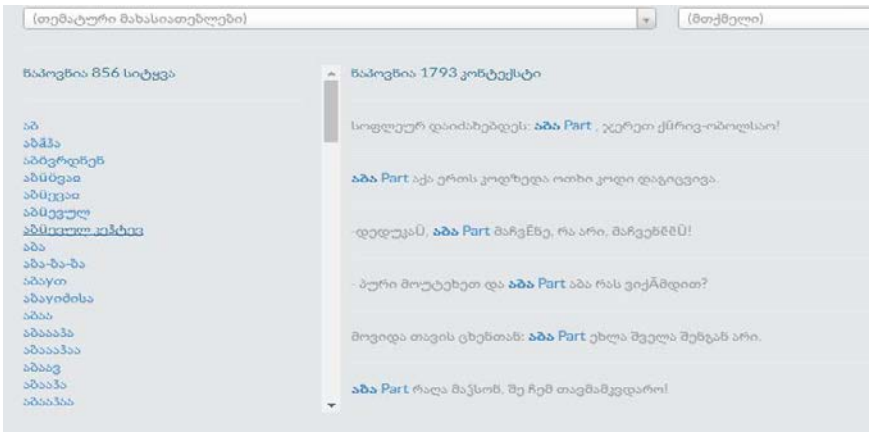

Figure 2: Contexts.

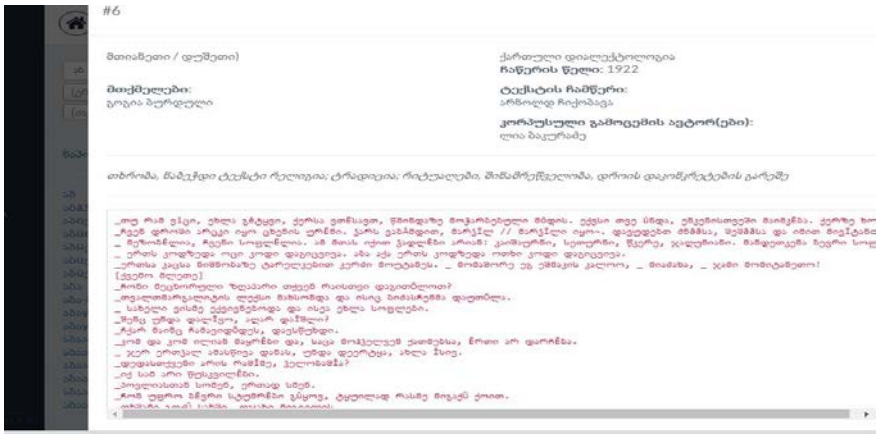

Figure 3: Text.

Currently, the problem of part-of-speech tagging is almost solved in the corpus, and individual dialects are marked with morpho-syntactic features. Besides, new texts are added to the corpus. Presently, the Georgian Dialect Corpus can be characterized in the following way: a special, representative, dynamic corpus, being permanently appended with new texts and developed technologically.
GDC is prospectively aimed at becoming a tool for documentation and study of Georgia's linguistic diversity and a principal educational resource.

1.2 Morphological annotation in GDC

The concept of the semi-automated lemmatization and morphological annotation of the corpus implied applying of the standard language parser and dialect dictionaries in the process (Lordkipanidze, Beridze, Nadaraia: 2015, 82-96).
A standard language parser-based system was developed, by means of which various lists of dialectal words were created:

- total recognition lists
- ambiguity lists
- lists of unrecognized words

The lists allow of:

- completion of "total recognition" lists produced as a result of the application of certain rules developed after the analysis of the lists of unrecognized words
- deep morphological markup (correct assignment of grammatical categories except for parts of speech) for up to 20%-40% of words (in various dialects)
- manual disambiguation and lemmatization in the disambiguation editor according to lists of ambiguous and unrecognized words (Beridze, Nadaraia, Lordkipanidze 2015-2).

An annotation and disambiguation editor is functioning within the corpus by means of which it is possible:

- to access necessary information by means of advanced search – to receive lists for testing by way of configuring word fragment, dialect, part of speech, grammatical form, status of word (annotated, ambiguous, non-annotated), and other features
- to view all contexts of selected words for testing or disambiguation
- to edit a lemma

- to edit a grammatical marker
- to ensure automated “selection” of ambiguous markers for contexts
- to save changes
- to ensure automated “assignment” of selected (revised) markers in contexts.

(Examples are given in the Figures 4, 5, 6 below)

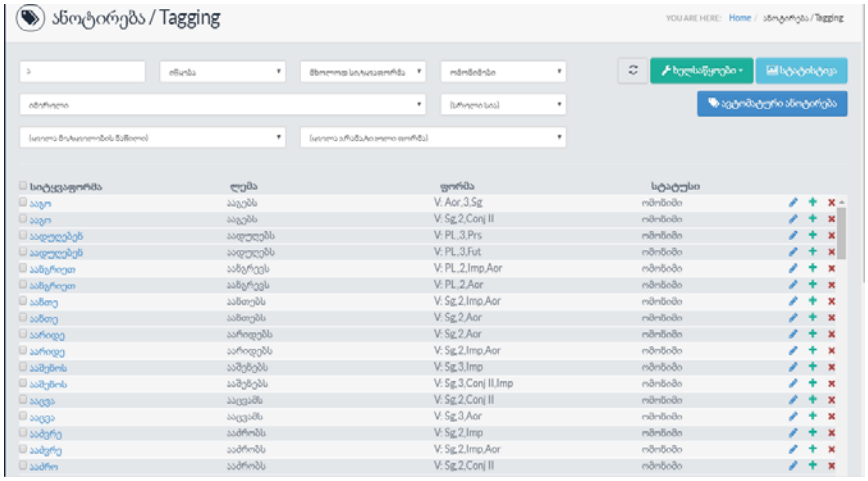

Figure 4: Testing annotation.

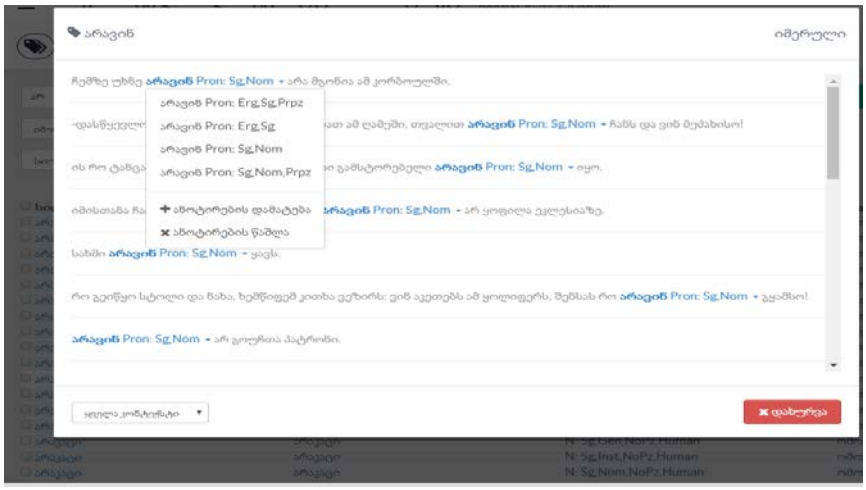

Figure 5: Testing annotation.

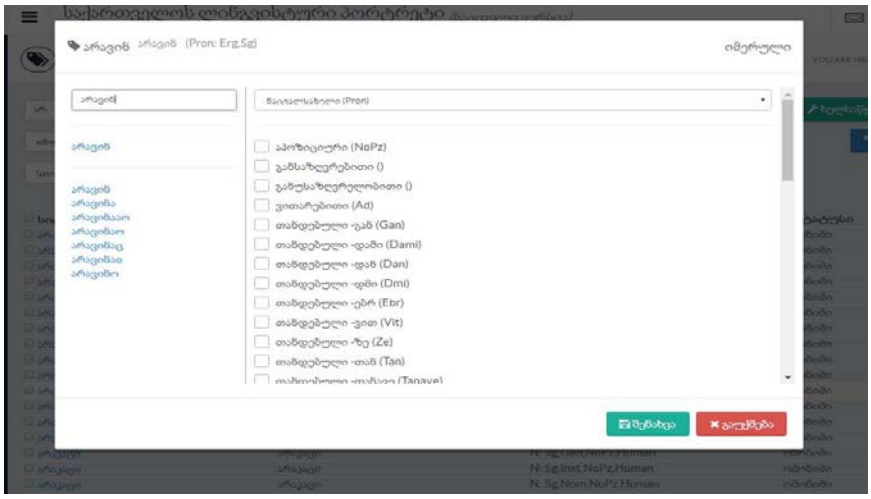

Figure 6: Editing annotation.

1.3 Text library

This component of GDC was developed in order to enhance its interdisciplinary function. Currently, the Georgian Dialect Corpus is a bulky collection of texts, in fact an oral historical and cultural sketch of the 20th century Georgia. Deep interdisciplinary study of these texts would be difficult by means of corpus query, since access to text in a corpus is only available at one of the (final) stages of linguistic investigation.

The text library is designed so that, at the very first step of query, a search result is a text. Search within the text library has been structured in accordance with the following features:

- linguistic features (language, dialect, sub-dialect)
- geographical feature (place of recording)
- bibliographical feature (publication, collection, etc.)
- personal feature (narrator, author)
- thematic feature (genre, topic, etc.)
- chronological feature (time period the text refers to)

Addition of the text library component increased the potential of interdisciplinary application of the corpus, exposed opportunities of its use as an educational, research, publishing, etc. tool.

2. Prerequisites for a Universal Dictionary of GDC

The idea of compiling a universal dictionary of GDC emerged at the stage of the corpus development, when the lexicographic base was created and, alongside with the technological objectives (to support morphological annotation), the opportunity of its processing and transformation into structured online dictionaries became realistic.

A universal online dictionary is aimed at incorporating the vocabulary of the Kartvelian languages into an integral lexicographic system. We designed and developed an instrument allowing the automatic generation of various types of dictionaries based on the universal dictionary.

2.1 Lexicographic base and online dictionaries of GDC

The lexicographic base of GDC is an individual platform built into the structure of the corpus. The base includes data from printed lexicographic sources (dialect dictionaries or dictionary materials). Lexical, linguistic and encyclopaedic data are arranged in the base in order to allow their use in various functions:

- in the function of a tool of representativeness and corpus annotation
- in the function of the generation of online dictionaries.

A grammatical marker is assigned to each lemma of a printed source integrated into the lexicographic base. The word-list is integrated into the common word-list, and, during query, a dictionary item will be reflected in the concordance in the same way as a token of the text data. A dictionary item is presented in the common concordance together with an entry in the same way as a textform is accompanied with its context.

Such inclusion of dictionaries in the corpus is aimed at increasing the degree of representativeness, also facilitating the partial solution of the problem of morphological annotation. Currently, there are up to 60 000 entries in the lexicographic base.

There are frequent instances when a word, included in the lexicographic base, does not occur in the text body of GDC and it is only documented by its dictionary version.

Currently, we have two fundamental objectives: to complete morphological annotation of the corpus and to develop and extend the universal lexicographic base.

The novelty, introduced by the extension of the lexicographic component in our project, is the inclusion of a mechanism of the creation of corpus-based dictionaries in the corpus design. This is neither a dictionary in the corpus as a text component (Beridze, Nadaraia 2011: 92), nor a dictionary in the corpus as a markup tool (Lordkipanidze, Beridze, Nadaraia, 2015: 82-96), but rather the crossroads of the "encounter" of a dictionary and a corpus, whereby a new dictionary is created, thus integrating entire lexicographic knowledge from a text and an old dictionary.

To sum up the activities carried out for the sake of developing the lexicographic component, it can be stated that we are standing at this "crossroads" where the available lexicographic knowledge (paper dictionaries) are entirely digitalized, integrated into the corpus as a source of morphological information and of relevant textual illustration; some of those dictionaries became a basis for new electronic dictionaries (Fereidanian, Chveneburebi, Ingiloan, Laz). Now it is necessary to:

- enrich the dictionaries with corpus sources
- create new electronic dictionaries.

Gradually, all of these dictionaries will be processed with the lexicographic principle of GDC: a dictionary entry will be constructed in accordance with the lexicographic editor of the corpus, various kinds of linguistic and encyclopaedic information will be added, textual data will be included in various structural fields (phonetic and word-formation variants, illustrations, etc.) of the lexicographic base, in order to create and publish online dictionaries (Beridze, Lordkipanidze, Nadaraia 2014: 91).

2.2 Online dictionaries of GDC

Four electronic dictionaries were compiled within the framework of the corpus – Laz, Fereidanian, Chveneburebi (speech of ethnic Georgians living in Turkey), Ingiloan. These dictionaries are based upon and take into consideration printed dictionaries and/or individual lexicographic publications, completely storing the lexicographic information contained in a printed source; however, it is not a diplomatic publication – an electronic version of a printed dictionary.

Dictionary entries of a printed dictionary are structured in accordance with the requirements of "the dictionary editor" of GDC. Some information (variants, synonyms, similar forms) appear as internal links; entries are frequently split into several entries, etc. Alongside with integrating information from available dictionaries, the principal source for the enrichment of the dictionaries of GDC is the text data of the corpus. Based on the data:

- new entries are added
- distinct meanings are added
- block of illustrations is being completed
- synonymic and homonymic items will be identified
- phonetic, grammatical and word-formation variations are added
- linguistic information is added: lemma is assigned grammatical group marker, variations are assigned inflection markers
- information about a word's foreign origin is added.

(Examples are given in the Figures 7, 8, 9 below)

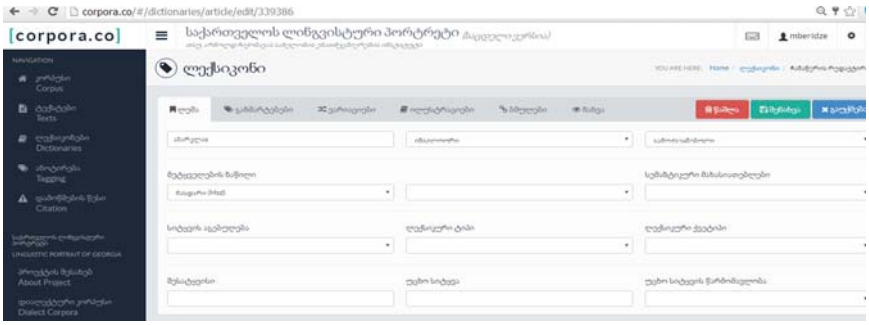

Figure 7: Editing new entries.

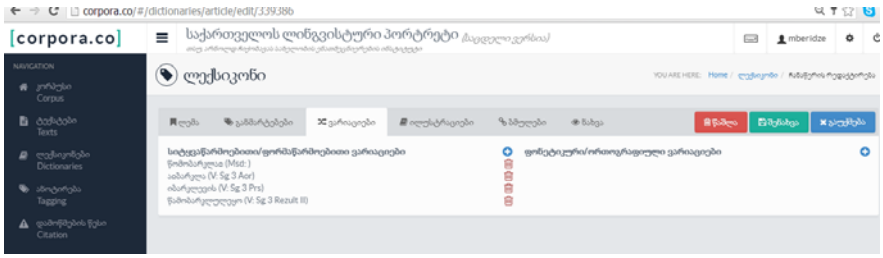

Figure 8: Editing variations.

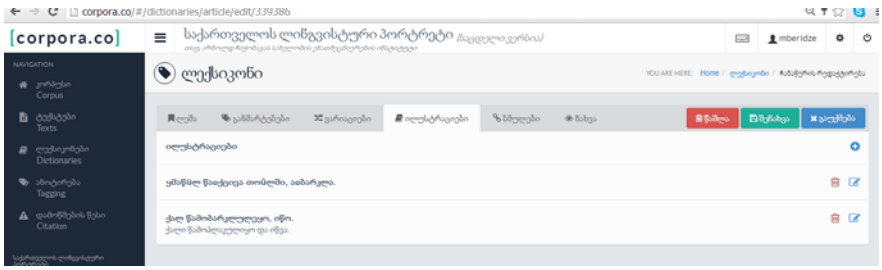

Figure 9: Illustrations.

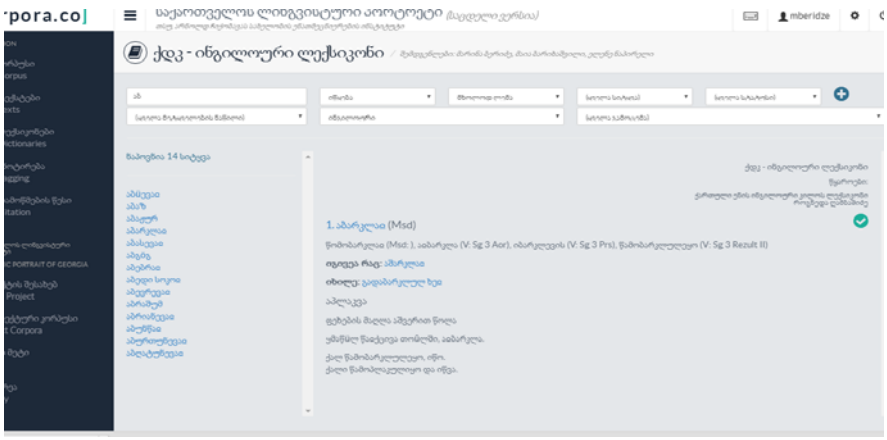

Figure 10: Ingiloan Dictionary.

The lexicographic editor of GDC stores dictionary entry as a database which, in its turn, is associated with the set of various features (bibliographic, linguistic, encyclopaedic, etc.) as individual lists.
Currently the user interface of GDC presents Georgian dialect dictionaries in the following way:

- as already stated, a dictionary entry is presented together with the text wordlist during search in the corpus
- during search in dictionaries, it presents a dictionary lemma and dictionary entry, arranged in accordance with the lexicographic principle of GDC (published dictionaries).
- during search in dictionaries, filtered search – in all dictionaries – presents lemmas of both published and unpublished dictionaries (other lexicographic information will become available after their publication).

2.4 Presenting linguistic information in the dictionaries of GDC

In GDC, linguistic information is presented by way of both linguistic annotation of textual word-list and the marking of various structural fields of a dictionary.
The lexicographic base of GDC is arranged in the way that linguistic information has been "distributed" among various structural fields. A dictionary has an element of deep and surface annotation as it is in the corpus, and, in addition, the structure of a dictionary entry allows of the unlimited addition of other kind of information.
In the lexicographic editor, linguistic information is distributed in the following way:
Lemma – a POS marker is assigned.

Grammatical / Word-formation Variation – markers of grammatical categories are assigned: case and number to nouns, screeve, number, person – to verbs. Besides, appropriate postpositions, particles, extension markers, other (clitic) will be assigned.
A phonetic variation will be assigned linguistic features. We mean that it differs from a lemma only phonetically and they have common linguistic features.
Alongside with morphological marking, a function of marking according to a semantic group, word structure, lexical type, word origin is arranged in the lexicographic editor.
Markup according to word (lemma) structure: In this field, we identified the following features: simple word, enclitic, free collocation, complex grammatical form, set phrase.
The issues of multi-componential linguistic items are very sensitive in corpus linguistics.
As we know, during the machine processing of texts, normally, a textform or a unity of markers "separated by spaces" is assumed as an initial item. However, it has also been noted that such an orthographic criterion cannot be a reliable basis in identifying of tokens (Mel'čuk 1997: 198-199). This problem is particularly salient during automated or semi-automated annotation when, owing to such approach, a considerable portion of linguistic information may be lost.
With a view to the aforementioned, alongside with a token – a word (separated by spaces), we assumed a minimum text unit to be multi-componential items the sub-division of which determines the structure of the markup mechanism.
One of the possible solutions in markup of multi-componential items is to compile lists of multi-componential items.
The decision to introduce multi-componential items as lemmas in the online dictionaries of GDC and, in addition, to assign appropriate markers is a basis for the compilation of such descriptive list. Later, they can facilitate markup of similar items throughout the corpus.
A lexical type list is a subordinate one to the classification list of word (lemma) structure, allowing of the classification of items of various structures in accordance with a lexical type.
The following are identified within lexical type features: neutral vocabulary, formulaic vocabulary, and special vocabulary.
The tree groups are subject to sub-division:

- Formulaic vocabulary: proverbs, catch phrases, various formulae (curse, pray, revile, etc.)
- Terms: usage domains will be assigned.
- Within neutral vocabulary, we identify proper nouns, phonosemantic words, other semantic groups.

The lexicographic editor is arranged in the way as to enable a user, during search, to generate a desired dictionary; for instance, dictionaries of pronouns, terms, uninflected words, cursing formulae, proper nouns, names of rivers, surnames, etc. (Figure 11).

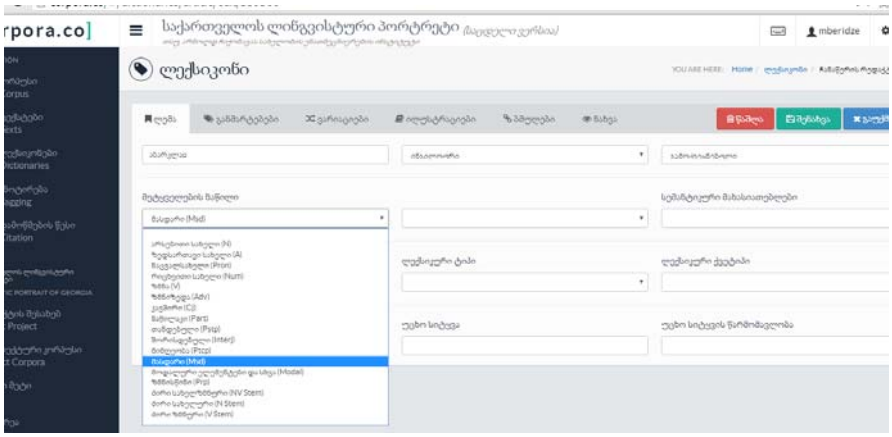

Figure 11: Editing linguistic information.

3 Conclusion

The next step in the development of the Georgian Dialect Corpus is the extension and upgrade of the universal lexicographic base, and, relying on the Georgian Dialect Corpus, the building of a technological and information resource for the generatoin of special dictionaries (of idioms, onomasitc, borrowings, grammatical, terminological, thematic, etc.).

Accomplishment of this task is closely and immediately associated with the development of the lexicographic component of GDC.

Our further activities are aimed at developing a universal and unique text and lexicographic reference system not only with an extensive and flexible query system but also with an opportunity to self-generate dictionaries of various types.

4 References

Beridze Marina, Nadaraia David, Lordkipanidze Liana (2015-1). The Georgian Dialect Corpus: problems and prospects, Corpus Linguistics and interdisciplinary Perspectives on Language Historical Corpora. Challenges and Perspectives, (CLIP), vol.5 Tübingen (Narr).

Beridze Marina, Nadaraia David, Lordkipanidze Liana (2015-2). Dialect dictionaries and morphological annotation in the Georgian Dialect Corpus Logic, Language, and Computation; Springer.

Beridze Marina, Nadaraia David, Lordkipanidze Liana (2014). Lexicographic Conception of Georgian Dialect Corpus and Problems of its Morphological Annotation, Applied Linguistics in Research, and Education, Proceedings of the 7th International Biannual Conference, Saint-Petersburg.

Beridze Marina, Nadaraia David (2011). Dictionary as a textual component of Corpus (Georgian Dialect Corpus), Corpus Linguistics – 2011. Proceedings, St. Petersburg, 2011.

Beridze Marina, Nadaraia David (2009). The Corpus of Georgian Dialects, NLP, Corpus Linguistics, Corpus Based Grammar Research, Tribune.

Mel'čuk I. A. (1997). Курс общей морфологии. Том I. М.: Языки русской культуры.