GEORGIAN DIALECT CORPUS: LINGUISTIC AND ENCYCLOPEDIC INFORMATION IN ONLINE DICTIONARIES[1]

MARINA BERIDZE – DAVID NADARAIA – LIA BAKURADZE
Arnold Chikobava Institute of Linguistics, Tbilisi State University, Georgia

BERIDZE, Marina – NADARAIA, David – BAKURADZE, Lia: Georgian Dialect Corpus: Linguistic and Encyclopedic Information in Online Dictionaries. Journal of Linguistics, 2017, Vol. 68, No 2, pp. 109 – 121.

Abstract: The Georgian Dialect Corpus (GDC) has been created within the framework of the project "Linguistic Portrait of Georgia". It was the first attempt to create a structured corpus of Georgian dialects. The work of this project includes building the technical framework for a corpus, collecting the corpus (text) data of Georgian dialects including the lexicographic data (dictionaries), their linguistic processing, digitizing, developing annotation framework, making decision on the morphosyntactic annotation. Currently, the Georgian Dialect Corpus is a platform consisting of the dialect corpus, the text library, the lexicographical database/online dialect dictionaries. For the purposes of developing the lexicographical database and dialect dictionaries, we have created a new program – the Lexicographic Editor. It allows us to structure and improve the dictionaries with multiple linguistic and lexicographic information. The lexicographic concept of the GDC has been developed taking into consideration linguistic and social features of the Georgian dialects.

Keywords: corpus linguistics, corpus lexicography, dialect corpora

1 INTRODUCTION

The languages spoken in Georgia generally belong to the Kartvelian languages, also called South Caucasian languages, or Iberian languages, a family of languages including Georgian proper, Svan, and Zan (further split into Mingrelian and Laz) as well as dialects of Georgian. The Georgian Dialect Corpus covers 17 Georgian dialects out of which 3 are spoken outside the country. They are Fereydanian in Iran, Ingiloan in Azerbaijan and so-called "Chveneburebi" in Turkey. Also, the Laz language is mostly spoken in Turkey. Only two Laz speaking villages do exist in Georgia.

The linguistic research of the Georgian dialects dates back to the 1920s, however there were several preliminary works and descriptions of dialect vocabulary in the 19th century.

The Georgian dialects are classified according to their ethnogeography, region and particular linguistic features. The varied and diverse vocabulary plays an important role in identifying a particular dialect or sub-dialect.

In the 20th century, there were several large migrations in Georgia, such as the massive migrations during the Soviet period; environmental migrations; migrations

[1] This work was supported by Shota Rustaveli National Science Foundation (SRNSF) [grant number 217008].

DE GRUYTER OPEN
DOI 10.1515/jazcas-2017-0022

as a result of war conflicts in Abkhazia and Shida Kartli. All of these factors caused the distortion of the linguistic boundaries in these areas.

In the following sections we will present the lexicographic database, representation of linguistic annotation.

2 ONLINE DICTIONARIES IN GDC

2.1 The Lexicographic Database of the GDC

The lexicographic database of the GDC is a separate section of the corpus. This platform has functions such as collecting, processing and converting lexicographic data into dictionaries.

The lexicographic database includes:

- Digitized online dictionaries from earlier printed dictionaries
- Various lexicographic data collected from lexicographic fieldwork
- Lexicographic data published by various authors
- Lexicographic data extracted from the existing linguistic and/or ethnographic studies.

The lexicographic database grows continually, with new texts being added over time. The database covers over 10 dialects and lists about 60 000 entries. This database has been developed based on the traditional lexicographic principles and methods. The research team will follow this methods in compiling comprehensive online dictionary of other Georgian dialects. Overall, four online dictionaries has been published so far. The published dictionaries are: Fereydanian, Ingiloan, "Chveneburebi", and Laz dictionaries. New dictionaries with corresponding lexicographic data will be added to the corpus interface.

2.2 Georgian Dialect Lexicography and the Lexicographic Principles of GDC

The Georgian dialect vocabulary is represented in the sources as follows:

- Georgian monolingual and multilingual dictionaries
- Dialect dictionaries
- Data from monographs, research outputs, publications.

The quality and the content of the Georgian dialect dictionaries varies greatly. There are several comprehensive, academic dictionaries, such as dictionary of Ingiloan dialect [7] Kartlian dialect [8], Imeretian dialect [9], [10], Gachechiladze [10], Khevsurian dialect [11], Tushetian dialect [12], Adjarian dialect [13] and etc.

In some cases, there are several dictionaries available for one dialect. These dictionaries can vary in size, content and lexicographic principle, such as the dictionaries for Gurian, Imeretian and Ingiloan dialects.

There are not sufficient lexicographic data for some Georgian dialects. In this case, lexicographic information about those dialects has been collected from various monographs, publications, manuscripts, for example, for Lechkhumian, Pshavian and Mtiuletian dialects. Additionally, the comprehensive dialect dictionary comprising all Georgian dialects is also available [14].

In general, the existing dialect dictionaries do not follow a single lexicographic principle. In particular, the dictionary entries are presented differently in various

dictionaries, the structure of the entries and linguistic and encyclopedic information in each dictionaries varies massively.

The Dialect Vocabulary in the Dialect Dictionaries Represents the following:

- Dialect vocabulary used in the standard Georgian language
- Vocabulary that differs in word-form and meaning, collocation and/or phrases
- Vocabulary common to standard Georgian language, but having different meaning in dialects
- Some Archaic vocabulary that are only preserved in dialects
- Some common vocabulary for both dialects and standard Georgian language,
- Some foreign borrowings
- Morphemes and modal elements that are specific to a particular dialect
- Some proper nouns [15].

In general, the Georgian dialect dictionaries do not include (or very rarely) proper names, surnames, toponyms, regular phonetic variations, foreign borrowings, that has developed their own dialect forms.

As discussed above, the dictionary entries are structured differently in each dialect dictionary. There is no correspondence between the macro- and micro-structure of the dictionaries. In addition, there is no coherence in illustrating the lexicographic information. Some dictionaries present several phonetic variants as separate dictionary entries, or describe several grammatical variations. Thus, each dictionary uses its principle for presenting linguistic, encyclopedic information in the dictionaries.

The lexicographic concept and principle, we are using in our research, aims to reach the following goals:

- to edit, unify and re-structure the existing dictionaries,
- to add new dictionary entries from the dialect corpus (data collected from linguistic field works),
- to improve the dictionary by adding dictionary examples, linguistic and encyclopedic information.

3 LINGUISTIC INFORMATION IN GDC

3.1 The Lexicographic Editor of GDC

The lexicographic editor of GDC is a platform that allows documenting and re-structuring lexicographic information.

The lexicographic information in the editor can be integrated by directly adding texts or uploading Excel spreadsheets. The editor is easy to use. It does not require any special knowledge of a particular program(s).

The lexicographic editor has several functions and fields/tabs as follows:

1. Saving – there are several saving tabs, where information of a lexicographic source is being saved. This information allows us to create a new dictionary entry, to edit or correct it. As each dictionary entry has several lexicographic sources, there are several tabs accordingly.

2. to add a lemma;
3. to write up the description for the dictionary entry;
4. to add encyclopedic information;
5. to add information about foreign borrowings;
6. to add foreign borrowings with their corresponding spelling in original language;
7. to add information about the structure of the dictionary entry (enclitic, composition, compound grammatical forms, simple collocation etc.);
8. to add lexical information about the dictionary entry (neutral, specialised etc.);
9. to add information about the semantic group;
10. to add information about lexicographic sources;
11. to add a grammatical marker for the lemma;
12. to add the grammatical variation(s) of the lemma;
13. to add the derivation variation(s) of the lemma;
14. to add the phonetic variation(s) of the lemma;
15. to indicate synonymous words;
16. to link to other lexical item;
17. to add dictionary example;
18. to provide translation for the dictionary example or to add additional comment;
19. to add the source for the dictionary example.

As shown above, the lexicographic editor allows us to compile a dictionary entry, to edit it, make some changes there, and publish it.

All stages are carried out separately: processing, editing and publishing. At the processing stage, only two tabs are active, these are Lemma and Save tabs.

The first step is adding lemma with the corresponding lexicographic sources. After that, existing lexicographic information can be re-structured according to the specific fields. The final step is publishing the entry, when it becomes available to broader dictionary users.

The unpublished (unedited) dictionary entry can be searched in the lexicographical database with the information of the lexicographic source. However, the dictionary entry is not complete as it does not have additional linguistic and encyclopedic information. Also, lemma and its variants are not marked according to the GDC principles (Fig.1).

For the user purposes, unedited and edited dictionary entries can be distinguished by marking them with special symbols. Cf. Fig.1 and Fig.2.

3.2 Lemmas in GDC

In the GDC lemmas represent the following: a single word, collocation, multi-component units that are considered as an individual word. The collocations may include unlimited number of words, sometimes it can be a whole sentence (if a phrase). In the corpus platform, in the lemma field, a single word or collocation are written without indicating the variations.

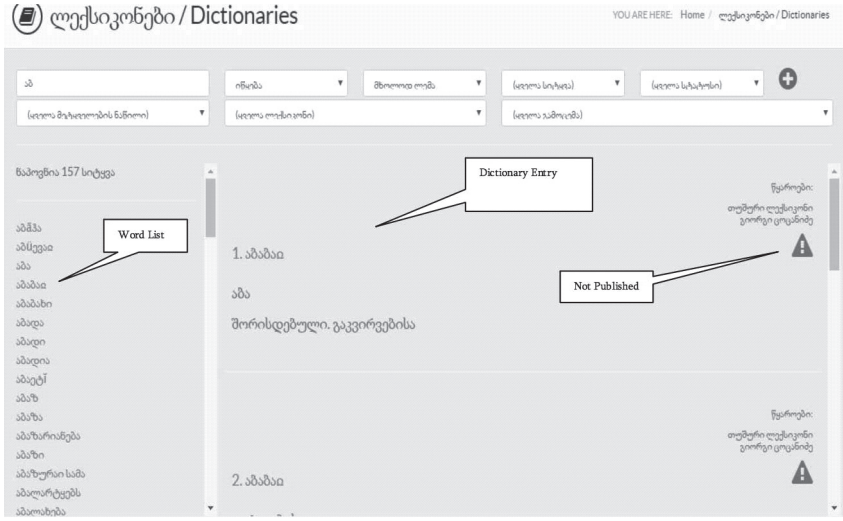

Fig.1. The progress of the Dictionary Entry Compilation

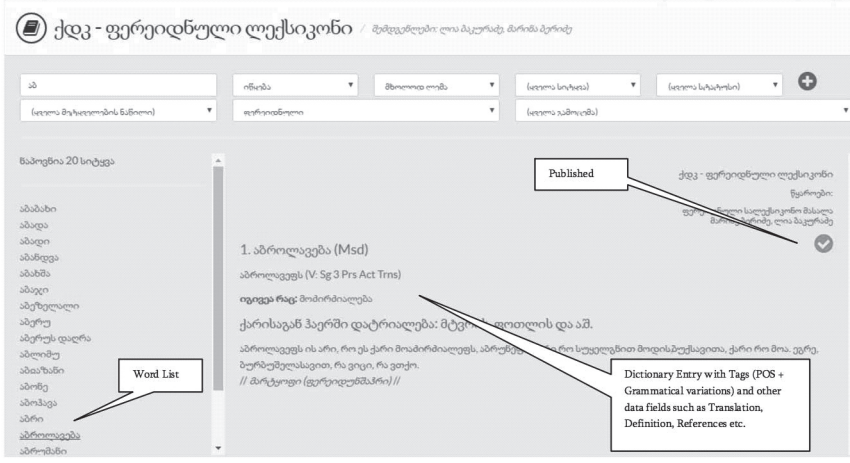

Fig.2. The Published Dictionary Entry in GDC

A lemma for nominals (noun, adjective, pronoun and numeral) are: nominative case, singular.

Lemmas for the verb are presented as follows: 3rd person of subject, resent tense, Singular. However in some cases (taking into consideration the lexicographic source) a Masdar can function as the lemma.

The dictionaries also include other parts-of-speech: adverbs, particles, conjugations, interjections, also some root or inflectional morphemes. These parts-of speech, unlike nominals and verbs, are rather easy to deal with in POS-tagging, as they do not change their forms.

Unlike traditional dictionaries, the dictionaries in the GDC include different types of proper names, such as ethnonyms, toponyms, hydronyms, anthroponyms, zoonyms, oeconyms etc., and foreign borrowings.

3.3 The Dictionary wordlist. A Lemma and Word.

The dictionary wordlist is a list of words that are represented in the corpus with corresponding lexicographic information. The wordlist is compiled as follows:

A lemma for a particular word is added in the lemma tab:

1. the lemma is linked to its phonetic variant(s), where the number of phonetic variant(s) is not limited.
2. the lemma is linked to its grammatical/derivational variant(s), where the number of variant(s) is not limited.

In addition, these variation(s), both lemmas and grammatical/derivational variation(s) have corresponding grammatical markers.

POS-tags are selected from the existing tagset that is based on hierarchical structure. The first line in this hierarchical structure is a list of parts-of-speech, also markers for the masdar, participle and other elements, such as the nominal root, the verbal root, the preverb etc. The second line of the hierarchy is a description of some basic grammatical features, also some derivational features.

Lemma – tagged with the first hierarchical marker.

Phonetic variant(s) – not tagged as the grammatical status of a phonetic variant is identical to lemma.

Grammatical/Derivational variant(s) – tagged according to its grammatical features.

Therefore, lemmas in the online dictionary are presented with several variants/ forms within one dictionary entry.

The query interface of the online dictionaries allows:

- to enter a search lemma and a word
- to search for individual words with their associated part-of-speech tags.

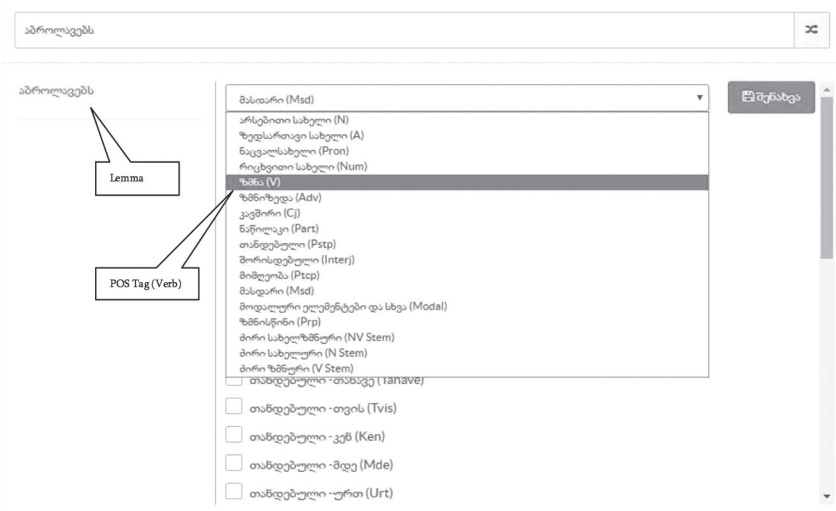

Fig. 3. Marking the Variants _ Example 1

Figure 3 shows the first stage of annotating grammatical and derivational variants, where the selected word-form has POS-tag assigned:

Lemma: abrolaveba (Msd) _ 'to cause spinning by wind'

Grammatical variant: abrolavebs (Verb _V)

After the mark-up, the information about the verbal categories are added, For instance, Person of Agreement, Number of agreement, Screeve etc. e.g.: abrolavebs _V Prs Sg 3

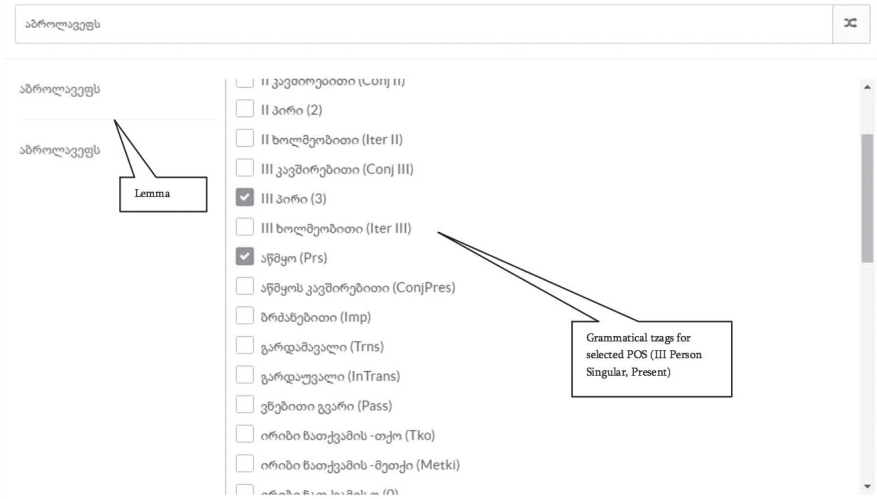

Fig. 4. Marking the Variants _ Example 2

3.4 Why to Include Separate Forms in the Dictionaries?

We have decided to include other variation(s) in the dictionaries for several reason. Firstly, following the dictionary principles of the GDC, a single word or collocation is represented in the lemma field. Secondly, it provides a comprehensive information about the word with their various phonetic variations. Additionally, it has somehow solved the problem of representing verbs in the dictionaries. In particular, there is no infinitive in Georgian and the masdar cannot represent the verb with its grammatical information in most cases. The Georgian monolingual dictionary [16] has established a rule for representing the verb in the dictionaries, where the verb is represented in 3rd person of subject, present tense, singular.

Our approach in representing a verb in the online dictionary has the following principles: no more than one word in the lemma tab, but the entry allows linking additional information such as phonetic and grammatical/derivational variations.

As an example, we will use the word adgomaj "stand up" from the Ingiloan dictionary. As for the grammatical and derivational variations, the words has eight forms. These forms are extracted from dictionary examples or texts from seven lexicographic sources. These variants are for example:

dgevis (V: SG 3 PRS) _ 'h/she stands up'; dgövi (V: SG 1 PRS) _ 'I stand up'; dgövita (V: PL 2 PRS) _ 'we stand up'; dgöodi (V: SG 1 IMPF)'I was standing up'; aadgomevar (V: SG 1 FUT)- 'I will stand up'; ovdek (V: SG 1 AOR)_ 'I stood up';

adgomi žanx (ADV) _ ‘time to stand up’; anadgom (PTCP) ‘standing’; aadgomela (PTCP) _ ‘to stand up’.

Based on the lexicographic sources, we made decision to introduce Masdar form as a dictionary entry (lemma) – adgomaj _ MSD. As for the grammatical variations, present tense, 3rd person of subject forms are also included – (dgevis _ V: SG 3 PRS).

One of the main principles in introducing a new variant to the word is that it should be well represented in various contexts in the lexicographic sources. As mentioned above, all these variations are included in the dictionary wordlist. Each variant can be used as a search word linked to the dictionary headword. See Fig.5.

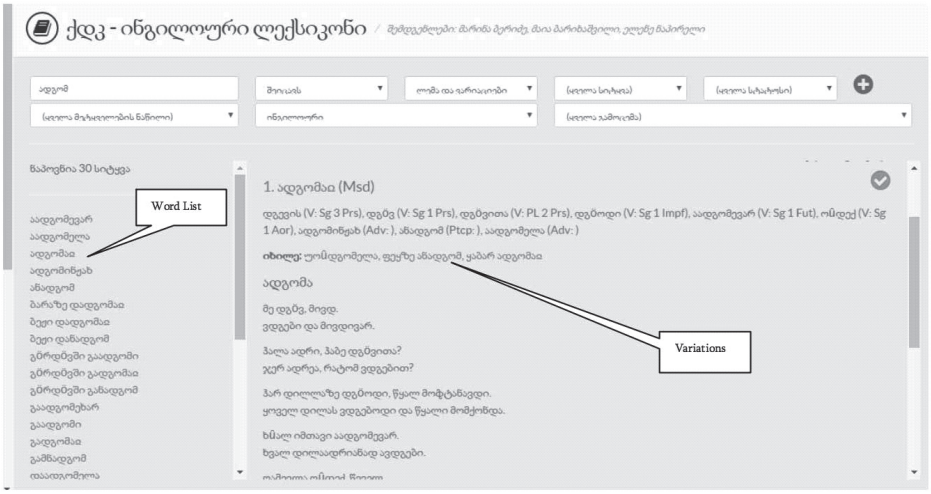

Fig. 5. Corpus Query Interface: Lemma and Variations

3.5 Definitions in GDC

Definitions in GDC are distributed in three different ways as follows:

- Simple variants (equivalents)
- Definition proper
- Encyclopedic information.

In the part Simple variants, one or several direct variants of dialect forms are introduced. In Definitions, information about the word that has a different meaning from the standard Georgian is given. The part Encyclopedic information includes different types of encyclopedic information that is entirely different from the definition.

It also contains other lexicographic information – the online dictionary has the function of adding lexicographic sources or dictionary examples. It allows to add unlimited number of dictionary examples and indicate our comments or translations.

The lexicographic sources can be selected from the reference database with predefined standard metadata. This function of the editor helps in improving the original dictionaries by adding the dictionary examples, indicating other existing lexicographic sources. The process implies combining several lexicographic

sources, including fieldwork data. The lexicographic sources are linked to a particular dictionary entry, but not to the whole dictionary. These enables the users to narrow down the query by indicating a particular lexicographic source or sources.

3.6 Additional Linguistic Information for the Dictionary Entry

As described above, the dictionary also offers information on the foreign borrowings and transliteration. This can be displayed in two ways. First, from the language selection tab by choosing the language. Second, the corresponding word in the original script can be chosen in a separate tab. This information is relevant in terms of sociolinguistics and ethnolinguistics, especially for the Georgian dialects that are spoken outside Georgia. These dialects have developed in foreign environment having no contact with other Georgian languages or dialects. As a result, there are many borrowings in these dialects.

Therefore, it is very relevant for our research to mark foreign borrowing in the Georgian dialects and indicate their Georgian (dialect) synonyms. It is an additional function for these online dictionaries. The dictionary editor can assign such properties as lexical/semantic groups and word structure.

A sub-section of the word (lemma) tagset is a lexical tagset, where both simple and compound words can be classified according to its lexical type. In lexical type, we cover the following: general vocabulary, literary vocabulary and specialized vocabulary. The literary vocabulary includes proverbs, expressions etc. The terms in specialised vocabulary have indicated the relevant field. A special attention is paid to the multi-element collocations.

Internal links. The synonymous words will be interlinked in the dictionary. The primary links connect synonymous words both within one dialect dictionary and in other dialect dictionaries.

3.7 Single Term Equals Single Sense

In the dialect dictionaries, word senses including ambiguous, figurative, specialized etc. are presented as a separate dictionary entry. For example, the dictionary entry mic'ai ('soil'/ 'land') in Ingiloan dialect dictionary is realized in 33 different ways. In our dictionary, all these sense are presented as separate dictionary entries that are connected with one another with the internal link.

Thus, the dictionary entry from the comprehensive lexicographic sources [7] are represented with a rather short description in our lexicographic database (mic'ai ('soil'/ 'land'). Also, the dictionary entry contains two examples. The rest of information from the lexicographic sources is distributed among 33 new dictionary entries compiled from these sources. In the new dictionary, the entries, lexicographic sources and examples are provided.

In some cases, information about the alternative sources is also added, such as information about other dictionaries, corpora, authors etc. In some cases an extra description about the entry provided when necessary. All new dictionary entries are interlinked with the original dictionary list (Fig. 7).

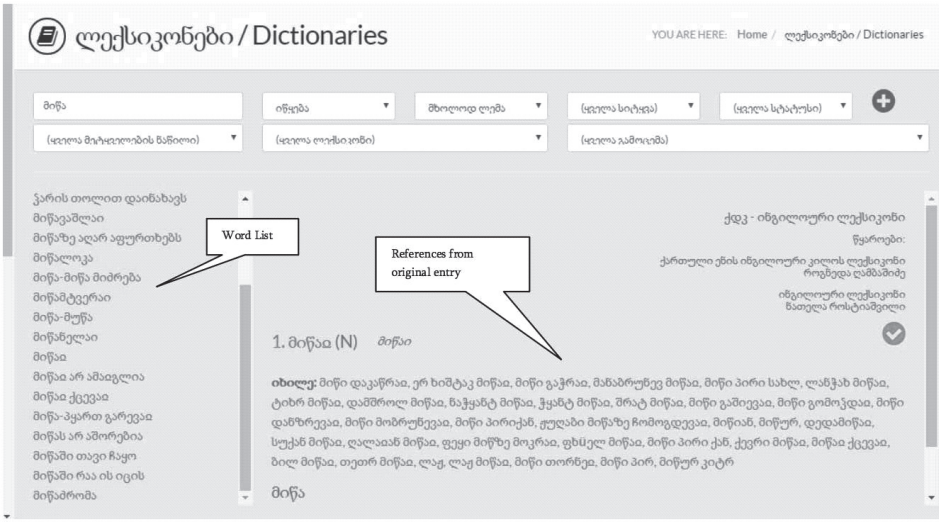

Fig. 6. Internal Links in the Dictionaries – Example 1

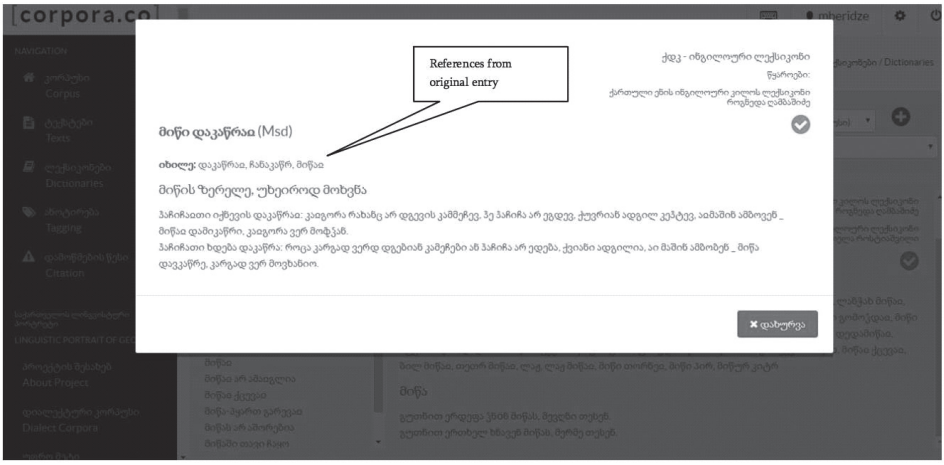

Fig. 7. Internal Links in the Dictionaries – Example 2

Figure 8 shows the link from the original dictionary entry to the new dictionary entry – miçi daḳaçraj ‘to plow soil unfruitfully’. In this example, a new dictionary entry provides the description of the lexicographic sources and the dictionary examples.

The internal links are bidirectional, Figure 8 shows the links of the new dictionary entry with the original entry, also with other dictionary units:

miçi daḳaçraj> miçaj

miçi daḳaçraj>daḳaçraj

Fig. 8. Internal Links in the Dictionaries – Example 3

3.8 Morphosyntactic and Semantic Markers

Morphosyntactic and Semantic Markers for the GDC have been developed taking into consideration the existing standards (TEI, EAGLES, The Leipzig Glossing Rules) and the Georgian National Corpus (GNC: `http://gnc.gov.ge`). Due to the specific features of the Georgian dialects, the markers were partially harmonised and modified, the additional markers were introduced when necessary.

Particularly, the most difficult tasks are related to the Georgian dialects that are spoken outside Georgia, such as the Fereydanian dialect in Iran. This dialect has been under direct influence of the Persian language having no contact with any Georgian languages or dialects for about four centuries. The new steps for cultural reintegration started by the end of 20th century. There are some new developments in the dialect and to capture these new or specific features, we have introduced new markers, as follows:

Pseudo _ incorrect forms from literary Georgian that are attested in the corpus data.

New _ Lexical parallel forms due the influence of the Georgian language.

Morphosyntactic and semantic annotation in the dictionary is carried out according to two hierarchical lists: In the first hierarchy, there are markers for lemma only, whereas in the second one, we have tags for morphological features. All these tags are selected and added to the relevant lexical unit and indicated in a separate tab.

3.9 The Dictionary Query Functions

The interface of the online dictionaries in the GDC allows for:

- Search for word, or a part of word
- Search for lemma or lemma variations
- Search for foreign borrowings, with indicating a specific language
- Search through a particular dialect
- Search according to a particular part-of-speech

- Search according to a particular grammatical features
- Search according to the status of the dictionary entry, e.g. in-progress etc.
- Search for completed dictionary entry
- Search according to the lexicographic source.

4 FUTURE PROSPECTS

The research team is currently working on the new project (funded by the Shota Rustaveli National Science Foundation) on the Lexicographic Database for Georgian dialects. The project aims to develop a comprehensive lexicographic database for the Georgian dialects and to build a dialect platform with the user interface. In this platform, we plan to integrate the corpus and lexicographic data of the Georgian dialects. This research will also examine the cartographic visualization of linguistic diversity of the Georgian dialects.

One of the plans of our research team is also to improve morphosyntactic and semantic annotation in the corpus. This will allow us to build different types of dictionaries, for example dictionaries only nominals, verbs, foreign borrowing dictionaries, phrase/collocation dictionaries etc.

The Georgian Dialect Corpus and its lexicographic database represents both synchronic and diachronic aspects of the dialects. We are currently developing the database about migrations through cartographic visualization. These will allow us to analyse linguistic data taking into account the linguistic area and migration.

The Georgian dialect corpus and visualization of linguistic diversity through cartography in the corpus will be one of the main linguistic resources. It can be introduced in teaching modules at universities. Also, it can be used in linguistic and interdisciplinary research.

References

[1] Beridze, M. et al. (2009). The Corpus of Georgian Dialects. In *Proceedings of the NLP, Corpus Linguistics, Corpus Based Grammar Research. Fifth International Conference Smolenice*, pages 25–35, Jazykovedný ústav Ľ. Štúra SAV, Bratislava, Slovakia.

[2] Beridze, M. et al. (2015). The Georgian Dialect Corpus: Problems and prospects. In *Proceedings of the conference on Historical Corpora Challenges and Perspectives*, pages 323–333, Frankfurt, Germany.

[3] Beridze, M. et al. (2011). Dictionary as a textual component of Corpus (Georgian Dialect Corpus). In *Proceedings of the conference on corpus linguistic*s, pages 92–97, St. Petersburg, Russia.

[4] Beridze, M. et al. (2014). Lexicographical concept of Georgian Dialect Corpus and problems of morphological analysis. In *Proceedings of the conference on Applied Linguistics in Science and Education*, pages 91–94, St. Peterburg, Russia.

[5] Beridze, M., Lortkipanidze, L., and Nadaraia, D. (2015). Dialect Dictionaries in the Georgian Dialect Corpus. In *Logic, Language, and Computation. 10th International Tbilisi Symposium on Logic, Language, and Computation, TbiLLC 2013*, pages 82–96, Springer, Berlin – Heidelberg, Germany.

[6] Beridze, M. et al. (2016). Lexicographic Potential of the Georgian Dialect Corpus. In *Proceedings of the XVII EURALEX International Congress, Lexicography and Linguistic Diversity*, pages 300–309, Ivane Lavakhishvili Tbilisi State University, Tbilisi, Georgia.

[7] Gambashidze, R. (1988). *Dictionary of Ingiloan dialect of Georgian Language*. Ganatleba, Tbilisi.

[8] Meskhishvili, M. et al. (1981). *Dictionary of Kartlian Diallect*. Metsniereba, Tbilisi.
[9] Dzotsenidze, K. (1974). *Upper Imeretian Dictionary*. Ivane Javakhishvili Tbilisi State University, Tbilisi.
[10] Gachechiladze, P. (1976). *Lexical material of the Imeretian dialect*. Metsniereba, Tbilisi.
[11] Chincharauli, Al. (2005). *Khevsurian Dictionary*. Kartuli Ena, Tbilisi.
[12] Tsotsanidze, G. (2012). *Dictionary of Tushian Dialect*. Sulakauri Publishing, Tbilisi.
[13] Nizharadze, S. (1975). *Adjarian dialect*. Sabchota Adjara, Batumi.
[14] Glonti, Al. (1975). *Dictionary of Georgian dialects*. Ganatleba, Tbilisi.
[15] Martirosov, A. (1985). The Main Issues of the Study of Georgian Dialect Vocabulary and Compilation of Dictionaries. *Ibero-Caucasian linguistics*, XXIII:139–148.
[16] *Explanatory dictionary of the Georgian language*. (1950–1964). 8 vols. Georgian Academy of Sciences [in Georgian]. Georgian Academy of Sciences, Tbilisi.
[17] Georgian National Corpus. Accessible at: `http://gnc.gov.ge.`
[18] Georgian Dialect Corpus. Accessible at: `http://corpora.co.`