**Manana Tandaschwili**
Goethe University Frankfurt, Germany
**Ramaz Khalvashi, Gregory Kakhiani, Mzia Khakhutaishvili, Nana Tsetskhladze**
Batumi State University, Georgia
**Rusudan Papiashvili**
Georgian Technical University, Georgia

# Modern Program Packages (FLEx and Elan) of the Language Data Management and the Prospects of Their Usage

## ABSTRACT

Documenting the languages and cultures and managing the data is firm basis of the future interdisciplinary researches. This gains even more importance for endangered and unexplored languages. The paper deals with the issues connected to the usage of Modern Program Packages (Flex and Elan) vividly and professionally for interdisciplinary researches, advantages and unique abilities of them. The work provides the information about the principles of Doculinguistics and about the modern standards of documentation. The issue of fitting Georgian, as an agglutinative Language to the international standards so that its nature is truly revealed, appears to be a challenging task.

The cooperative scientific project "Linguocultural Digiarchive" was implemented by Batumi Shota Rustaveli State University and Frankfurt Goethe University. It was created using the modern framework standards of documentation and archiving in Elan and FLEx formats.

FLEX is able to show the essential peculiarities of the language and to show the ways of fragmentation and the functions of morphemes: various existing wordforms, kind of gramma, semantic and pragmatic categories.

FLEx as a modern package of managing the language data gives an unique opportunity to show the diversity of language materials and to synchronically and diachronically study a wide range of issues; to check the scientific hypothesis; to create new lexicons; to describe the paradigm of declension and conjugation; to create the grammar framework; to give automatic analysis of the data, to create a united standard.

***Kay words:*** *Language documentation, FLEx, Elan, Batumi Liguocultural DigiArchive* [1]

## Introduction

In the last thirty years it has become essential to document unexplored and endangered languages. This is even more important for the modern world, because it needs to have a reliable knowledge of Linguocultural problems. Because of the "Big Languages", language forms and categories that are found in unexplored languages have not yet been analyzed linguistically.

[1] This project has been made possible by financial support from the Shota Rustaveli National Science Foundation (№DI 2016-43). All ideas expressed herewith are those of the author, and may not represent the opinion of the Foundation itself.

Consequently, the knowledge of the system of General lingiuistics appears to be imperfect and incomplete. Reinterpretation of language as linguocultural phenomenon became the basis of new scientific trend – doculingistics.

In a document- "The Atlas of the World's Languages in Danger of Disappearance and Georgia"[1], includes the Georgian Language among those languages, which may be replaced by some dominant languages by the end of the 21st century. Accordingly, the specialized literature underlines the importance of the following issues: creation of digitalization, consolidation of documented materials standardization, availability and creation of national corpus (Gippert, Tandashvili, 2012; Tandashvili, 2016; Tandashvili, Phurtskhvanidze, 2013; Lomia, Gersamia, 2012).

This article aims at revealing the prospects of using Modern Program Packages (Flex and Elan) vividly and professionally for interdisciplinary researches. It also depicts the advantages and unique abilities of them. The article discusses the issue of fitting Georgian, as an agglutinative Language to the international standards so that its nature is truly revealed, that appears to be a challenging task.

The creation of National Corpus promoted the accumulation of the experience in documentation. It meant recording the forms in authentic, natural situations (Himmelmann, 1998: pp.161–195); It is about the multipurpose of the records (Gippert, Himmelmann, Mosel, 2006) and about the major strategies of language documentation (recording, processing, preservation and dissemination of the primary data) etc.

## Discussion

The cooperative scientific project "Linguocultural Digiarchive"[2] was impelemented by Batumi Shota Rustaveli State University and

[1] http://www.ice.ge/kartuliena/pages/unesco/atlas_g.pdf

[2] http://digiarchive.bsu.edu.ge

Frankfurt Goethe University. These universities were mainly responsible for introducing the main principles of doculinguistics and creating the modern framework stardards of documentation and archiving in Elan and FLEx formats.

The project aimed at finding and development the verbal materials (biographies, household and agricultural details, the cultural and historical facts preserved in memory, religious rituals and customs, ethnological materials) by using new methodological and technological basis, creating a certain framework for digital documenting (Tandashvili, Khalvashi, Beridze, Khakhutaishvili, Tsetskhladze, 2017), that consists of several stages: finding the resources (recording the material), recording (registration), digitalizing, preserving (archiving) and further caring (protecting) of the data.

Four types of resources are being prepared: A- Archived audio and video materials (MP3 and AVI formats); B - Archived audio and video materials with transcribed texts (in Elan format); C - Multimedia annotation of the Archived video materials: transcribed, glossed and interlinised; (Processed in FLEx); D- Digitally documented and archived video material with multimedia annotation and English Translation (Tandashvili, and all, 2017).

Transcribed material was processed in FLEx[3], which is a multimedia platform for Data Management, text differentiation and analysis. The package was created by international scientific society in SIL international[4].

Together with FLEx, professional instrument Elan[5] was used for multimedia annotation, transcription, glossing and interlining. Annotated video, audio and multimedia files were united in EAF[6]

It should be taken into consideration, that finding, structuring

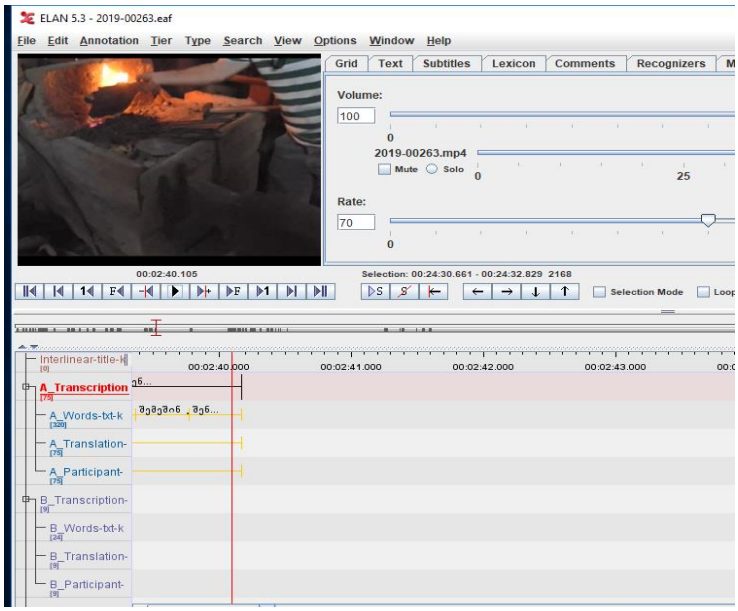
[3] https://software.sil.org/fieldworks/download/
[4] http://www.sil.org/
[5] https://tla.mpi.nl/tools/tla-tools/elan/download/
[6] http://www.file-extension.org/de/extensions/eaf)

and preserving the resources using modern program packages (FLEx and Elan) in the framework of this project appears to be unprecendented and not yet fully used to its maximum capacity.

The main advantage of ELAN is that it can incorporate video material and text resources in one file. Nereby it is also able to synchronize transcribed text with video and audios signals. Consequently, this kind of diverse (social, cultural, political, economical, spiritual, religious etc.) material is credible and all the resources resulting from this work are interdisciplinary. It will become the basis of several linguocultural reseaches as it comprises different traditions that can be a very valuable information for different branches of science. It also makes it possible to analyze some issues of the verbal speech, gestures and mimes.

After the video material is processed, Elan text is segmented into sentences. Some fields have been created for structuring the video files: 1. transcription-txt. kat – for transcrpting the sentences; 2.Translation-gls.en – for English Translation; 3. Words-txt.kat – For lemmes; 4. Participant-note.kat –for additional information concerning infromants (surname; name, age etc).

This is how the final file looks: (See picture 1)




Picture.1. Structure of fields in ELAN

This kind of file is ready for being processed in FLEx. The advantage of the given program is the standardization and automatization of the language data and better interface. It met all the five key requirements necessary for field research: 1. On the cost of processing interlinear text, the lexicon is automatically generated and then, completed. 2. all the operations are conducted inside the FLEx and a user is not obliged to use the other program. 3. The morphological model is pre-processed and it is integrated in the program as grammatical characteristics. 4. The principle of working in the program is not complicated as it is bused on the principles of preceding programs. 5. The software is easily perceivable and usable (Black, Simons, 2006).

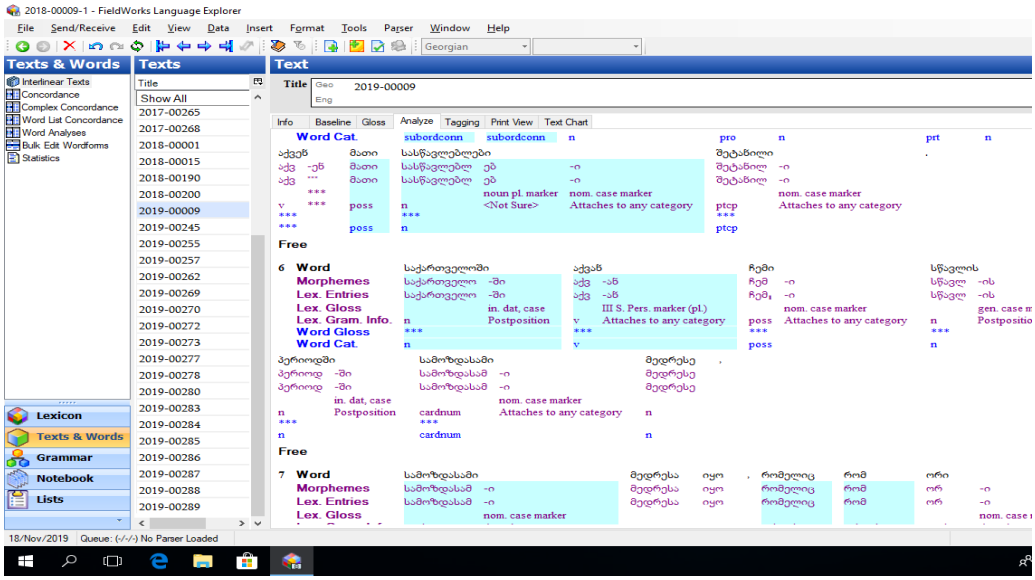
FLEx also has an editor of metadata (the following things are indicated: the time and place of recording; thematics; the age, education, migration of the teller; the type of conversations (monogue and dialogue)) etc. The materials are prepared in IMDI[7], CMD[8] ლა XML[9] formats and the multitude of formats is necessary for the exchange of data with different programs.

Flex, as a very elaborate system of language data management. The program includes ten classes and 88 fields for describing the vocabulary. For morphological analysis, there are 60 classes and 185 fields. FLEx is a unique instrument for creating a lexicon. After exporting a file from ELAN to FLEx, the annotated and internationalized words that are necessary for integrating the data are depicted in the Lexicon. This operation goes beyond the function of ELAN. The information about certain words is collected, such as: anthrophologic category, genre (monologue, behavioral text, narrative, also the sources, researchers, localization, tellers etc). The lexicon also contains wordlist with grammar indexes and definitions. Thus, making it possible to analyze the materials synchronically and diachronically.

There are different windows in Texts & words instrument. The window - **Info** gives comprehensive information about certain words; In **Baseline**, one can find a full text; **Gloss** gives the translation of the words. In **Analyse,** grammatical

[7] https://tla.mpi.nl/imdi-metadata/
[8] https://portal.clarin.nl/node/4061
[9] https://www.w3.org/XML/

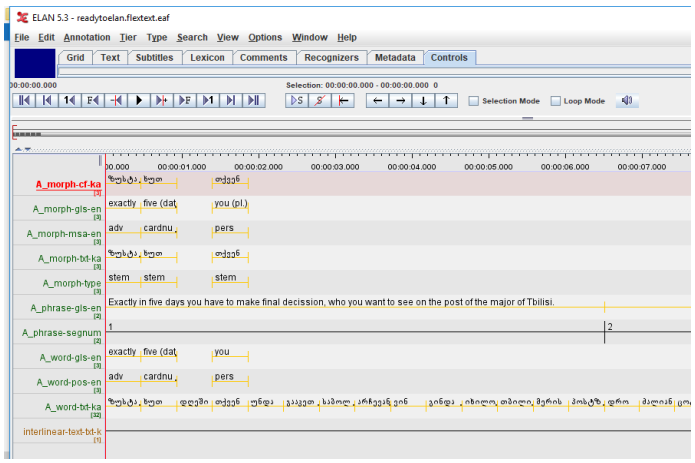
analysis of the words is possible. The stem, the root, affixes are identified and classified using **Parser**, the grammatical analyser. **Print view** gives an opportunity to preview the document before printing. **Tagging** implies tagging the text. **Text chart** is an instrument dealing with text discourse (picture 2).




Picture 2. Instrument **Texsts & words in** FLEX

List is a multifunctional instrument: academic domains, anthropological categories, the level of education, genres, dialectological information, semantic domains (social behaviour, language and thinking, daily life, activity, physical activity) and they have their subdomains.

The processed information in FLEX is again imported in ELAN (picture 3).




Picture **3:** Language data processed in ELAN

Segmentation system is very flexible and convenient.

As it can be seen at the picture, FlEX is invaluable in a way that in the instrument "Grammar", one can find parts of speech with definitions and it is possible to add affix pattern (the name, description of the pattern) and also, subcategories.

Using segmentation and glossing, FLEX is able to show the essential peculiarities of the language and to show the ways of fragmentation and the functions of morphemes: various existing wordforms and kind of grammat, semantic and pragmatic categories are expressed in the language using the morphemes.

For example, suffix "*-eb"* is a marker of the plural number (***khe-eb-i*** (trees)) and in the verb it is the thematic suffix (**ak'et-eb** (you do)). **"o"** is the third person marker in imperfect screeve (***ts'aiġ-o*** *tsigni* (He tooks the book)); **"o"** is also a marker of subjuctive mood (***ts'aiġ-o*** : kargi iqneba tsigni shen rom ***ts'aiġ-o*** (it will be good if you take the book)); "**s**" is a marker of Dative Case and the third person marker of subjective verbs: A) ***deda-s*** Dative case marker; B) ***sts'er-s*** (He writs) III Subj. marker.

When "**s**" is already depicted in the program as a dative case marker and a specialist does not agree, he/she can add the exact classification.

As differential analysis of the above-given morphemes is fixed in the program, specialists have opportunity to select exact qualification in every concrete case.

Several parts of speech can be simultaneously main parts of speech and functional words: "***unda***" is a **verb** (wants/wants to go) and **modal component** "*unda"* (must go/ is obliged) to go.

One and the same segment, for example, "**da**" can be major part of speech (noun), the minor part of speech (linker), morpheme (preverb: ***da-ts'era***) and adposition dialect form (***chemda*** (for me) etc.

Morphemes and words can stay without categories (due to morpho-

syntax specifics of languages), however the program gives an opportunity to make a comprehensive analysis - it is possible to add non-existent categories to standard version (i.e. to adapt morpho-syntax of the Georgian language). This makes the program to be even more perfect.

The program enables us to show the diverse dialectal forms as well as literary ones. In the variants of the data, one can find accumulated literary, dialectal, old and new language forms and also, lexical and grammatical units of relative languages. For example: "***darga***" *(*sow) and *"**dargo**"*, or "***dapkveuli***" (milled) and "***dapkvevli***", "***avadmq'opi***" (ill) and "***avantq'opi***" are divided as literary and dialectal language forms.

"*khvimiri*" (place for grain/corn in the mill) and "*godori"*(long twiggen/woven basket) are represented as archaism and modern language forms. The lexicon also shows the language contacts: ***topali*** (Turkish, lamed), ***zastavo*** (Russian, border post), ***lodosi*** (Greek, warm cloudy weather that speeds the melting of the snow) and ***jumadia*** (Lazian, uncle).

The texts incorporated in the project are considered to be a corpus, which gives dialectal and literaty forms (phonetic and morphologic) and literaty forms can have several versions in one and multiple subsystems.

By providing complete material of a dialect, we can later create comparative lexicons. From the field of data, we can clearly see the peculiarities of the English Language but not Georgian. For example, Grammatical gender, article, supinum and gerund etc. For the successful functioning of the program, adding the new fields (it is possible in the program using "Add new Sense" icon), which would depict the different markers of the Georgian Language (mostly all the essential markers of the verb, especially, polipersonalism, screeves,

thematic markers, contact, additional functions of preverbs, semantic details of passive voice and a lot more) became necessary (Shanidze, 1973).

In order to successfully operate the above given packages, it's necessary to perfectly show the Grammatical structure of the Georgian language. After the comparative study of the certain issue, it is necessary to accurately and qualititetively segment the data and choose the appropriate terms (Surmava, Beridze, 2008) and also, it's important to solve the problem of the terminology differences.

The advantage of FLEx is automatization of analysis. As A. Blake and G. Simmons state, the more elaborate system becomes, the better automatization of the text analysis it gets: "When a grammatical morpheme is glossed in interlinear text analysis, the MGA presents a view of the complete feature catalog as a choice list for possible glosses. As glosses are selected, they are added to a language-specific feature system which is being automatically constructed behind the scenes" (Black, Simons, 2006).

Automatic description of the grammar is a very faraway perspective. And the most interesting thing is that FLEx is capable of checking the hypothesis of the linguists using the factual data it includes (ibid).

What should be done now? The main difficulty now is the shortage of data: "Thousands of word forms could be needed to establish basic patterns of allomorphy, for example, or the structure of an inflection-class system" (Kim, 2020).

## Conclusions

Therefore, by accurate processing of multiple texts of different contents, by recording them in authentic situations, it will become possible to depict the peculiarities of agglutinative languages. Making the exact and qualitative analysis, reviewing certain issues and elaborating the

terms will promote the stable development of the Georgian language. The availability of the resources will become the basis for some fundamental interdisciplinary researches.

Documenting the languages and cultures and managing the data is firm basis of the future interdisciplinary researches. This gains even more importance for endangered and unexplored languages. FLEx and Elan, which are modern packages of managing the language data give us an unique opportunity to show the diversity of language materials and to synchronically and diachronically study a wide range of issues; to check the scientific hypothesis; to create new lexicons; to describe precisely the paradigm of declension and conjugation; to create the grammar framework; to give automatic analysis of the data, and finally, to create a united standard of preserving and analyzing the data.

## References:

Black, Simons, 2006 – Black A. & Simons G. The SIL FieldWorks language explorer approach to morphological parsing. In *Computational Linguistics for Less-studied Languages: Proceedings of Texas Linguistics Society*, 10, 3-5 November, Austin. 2006. [accessed 19.05.2020]. Available online at: https://scholars.sil.org/sites/scholars/files/gary_f_simons/preprint/flexparser_preprint.pdf

Gippert, Himmelmann, Mosel, 2006 – Gippert, J., Himmelmann, N. P. & Mosel U. (Eds.). *Essentials of language documentation*. Berlin, Germany: Mouton de Gruyter. 2006. [accessed 19.05.2020]. Available online at: https://scholarspace.manoa.hawaii.edu/bitstream/10125/4353/evans.pdf

Gippert, Tandashvili, 2012 – Gippert, J., Tandashvili, M. "Structuring a Diachronic Corpus. The Georgian National Corpus project" Proceedings of the international conference „Historical Corpora 2012", Frankfurt. 2012. [accessed 19.05.2020]. Available online at: http://armazi.uni-frankfurt.de/gnc/gnc_2012.pdf

Himmelmann, 1998 – Himmelmann, N. Documentary and descriptive linguistics. Linguistics 36(1) (pp.161–195). 1998.

Kim, 2020 – Kim, Y. Morphology and Language Documentation. 2020. [accessed 19.05.2020]. Available online at: https://oxfordre.com/linguistics/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-518

Lomia, Gersamia, 2012 – Lomia, M., Gersamia, R. Interlineal Morphemic Glossing (Morfological Analysis of Megrelian Texsts). Ilia State University, Part 1, Tbilisi. 2012.

Shanidze, 1973 – Shanidze, A. Morphology of Georgian Language, Tbilisi. 1973.

Surmava, Beridze, 2008 – Surmava, N., Beridze, M. The basic concepts of Corpus Linguistics and the attempt to select Georgian matches for them, Natural processing of languages. Conference theses. 2008. [accessed 19.05.2020]. Available online at: http://www.ice.ge/conferenciebi/Bunebriv%20enata%20damushaveba.html

Tandashvili, 2016 – Tandashvili, M. Digital documentation of the language (Introduction to Doculinguistics), Batumi Shota Rustaveli State University, Batumi, 2016.

Tandashvili, Khalvashi, Beridze, Khakhutaishvili, Tsetskhladze, 2017 – Tandashvili, M., Khalvashi, R., Beridze, Kh., Khakhutaishvili, M., Tsetskhladze, N. Batumi Linguacultural Digital Archive, International Journal of multilingual Education 2017-10 (pp52-68) 2017. ISSN 1512-3146. [accessed 19.05.2020]. Available online at: http://www.multilingualeducation.org/en/article/37

Tandashvili, Phurtskhvanidze, 2013 – Tandashvili, M., Phurtskhvanidze, Z. Glossary of Corpuslinguistics, Frankfurt, Tbilisi. 2013.