ირინა ლობჟანიძე, ერეკლე მაღრაძე, სვეტლანა ბერიკაშვილი, ანზორ გოზალიშვილი, თამარ ჯალაღონია
ილიას სახელმწიფო უნივერსიტეტი, საქართველო
irina_lobzhanidze@iliauni.edu.ge

ქართული ენის უნივერსალური სინტაქსური ანოტირების პრინციპები*

ნაწილი 1. შესავალი

ბუნებრივი ენის დამუშავების მთავარ ამოცანას წარმოადგენს უნივერსალური ენობრივი რესურსების შექმნა მსოფლიოს სხვადასხვა ენისთვის, რომლებსაც განეკუთვნება სხვადასხვა ენობრივი პროექტის ფარგლებში შექმნილი როგორც პროგრამული მომსახურება, ასევე მონაცემთა ბაზები (მაგ. უნივერსალური სინტაქსური დამოკიდებულებები (UD),[1] უნივერსალური მორფოლოგია (UniMorph),[2] ფრაზეოლოგიური შესიტყვებების პარსინგი (PARSEME)[3] და სხვ.). უნივერსალური დამოკიდებულებების პერსპექტივიდან გამომდინარე, ერთ-ერთ ყველაზე მნიშვნელოვან რესურსს უნივერსალურ დამოკიდებულებათა ინიციატივა წარმოადგენს, რომელიც უზრუნველყოფს სხვადასხვა ენის კროსლინგვისტურ ანოტირებას სინტაქსურ ხეთა ბანკების გაზიარების მიზნით. ამ მხრივ, აღსანიშნავია, რომ ეს რესურსი არც ერთ ქართველურ ენას არ მოიცავს, არც ქართულს მოიცავდა 2023 წლამდე. აღნიშნული კი გამოწვეული იყო შემდეგი გარემოებებით: პირველი, ქართული განიცდის მონაცემების უკმარისობას; მეორე, რესურსები, რომლებიც იქმნება სხვა ენებისათვის, ნაკლებად გამოსადეგია ქართულისათვის ანოტირების მორფოსინტაქსური სქემატური სხვაობების გამო. ამავდროულად, რესურსები, რომლებიც შეიქმნა ქართულისათვის, როგორიცაა მორფოლოგიური ანალიზატორი და

* კვლევა განხორციელდა შოთა რუსთაველის საქართველოს ეროვნული სამეცნიერო ფონდის მხარდაჭერით [FR-22-20496]. წინამდებარე პუბლიკაციაში გამოთქმული ეკუთვნის ავტორებს და შესაძლებელია არ ასახავდეს ფონდის შეხედულებებს.

1 იხ. https://universaldependencies.org/, ბოლო წვდომა: 2023 წლის 2 დეკემბერი.

2 იხ. https://unimorph.github.io/, ბოლო წვდომა: 2023 წლის 2 დეკემბერი.

3 იხ. https://typo.uni-konstanz.de/parseme/, ბოლო წვდომა: 2023 წლის 2 დეკემბერი.

გენერატორი (Lobzhanidze 2022) ახდენენ ლემების, მეტყველების ნაწილებისა და მორფოსინტაქსური ტეგების მიწერას, თუმცა არ ახდენენ ომონიმურ განმსგავსებას და სინტაქსურ ანალიზს; ასევე, მონაცემები ქართული ენის კორპუსიდან (GLC)[4] და KartNLP[5] არ არის საკმარისი ზემოხსენებული მიზნებისათვის.

ქართული ენის სინტაქსური ანოტირების სქემების შემუშავება საფუძველს წარმოადგენს რთული მორფოლოგიის მქონე ენების სინტაქსური პარსინგისათვის და, ასევე, მცირე რესურსების მქონე ენების (როგორიცაა ქართველური ენები) დამუშავების გაუმჯობესებისათვის. წინამდებარე ანგარიში აღწერს ქართული ენის სინტაქსური ანოტირების სქემების შემუშავებასთან დაკავშირებულ ზოგიერთ საკითხს და წარმოადგენს ქართული ენის მორფოსინტაქსური კომპიუტერული ანალიზისა და უნივერსალურ სინტაქსურ დამოკიდებულებათა პროექტის (# FR-22-20496) ანგარიშის ნაწილს.

ნაშრომი შედგება ოთხი ნაწილისაგან. პირველი ნაწილი მოკლედ აღწერს სხვადასხვა ენობრივ რესურსს, რომელიც ხელმისაწვდომია ქართული ენის შემთხვევაში (XPOS დონე) ანოტირების და ტეგირების სხვადასხვა დონეზე, არსებულ კორპუსებს და მათი მონიშვნის დონეებს და აღნიშნავს სინტაქსური ანოტირების საჭიროებას. მეორე ნაწილი ყურადღებას ამახვილებს პროგრამულ უზრუნველყოფაზე, რომელიც გამოიყენება ქართული ენის არსებული ტეგსეტების უნივერსალურ დამოკიდებულებათა ფორმატთან დასაკავშირებლად ენაზე დამოკიდებული ტეგსეტის (UPOS) და თვისებათა ტეგსეტის (FEATS) დონეზე და იმ ფუნქციებზე, რომლებიც გამოტოვებულია UPOS და FEATS დონეზე, ხოლო მესამე ნაწილში განიხილება ქართული ენის სინტაქსური ანოტირების პრინციპები შესაბამისი ფორმატის ფაილების მაგალითზე (Georgian-ud-test.conllu, README.md და ა.შ.) და ენობრივი დოკუმენტირების ფაილები (introduction.md და index.md), რომლებიც უკვე ატვირთულია GitHub[6]-ზე. მეოთხე ნაწილი აჯამებს შესრულებულ სამუშაოს და აღწერს სინტაქსური ხეთა ბანკის განვითარების სამომავლო პერსპექტივას.

ნაწილი 2. ქართული ტეგსეტის უნივერსალურ დამოკიდებულებათა ტეგსეტთან დაკავშირების პრინციპებისათვის

სინტაქსური უნივერსალიების კვლევის თეორიული საფუძველი, როგორც წესი, ორ თეორიულ ჩარჩოს მოიცავს, კერძოდ, ფუნქციონა-

4 იხ. http://corpora.iliauni.edu.ge/, ბოლო წვდომა: 2023 წლის 2 დეკემბერი.

5 იხ. https://qartnlp.iliauni.edu.ge/, ბოლო წვდომა: 2023 წლის 2 დეკემბერი.

6 იხ. https://github.com/UniversalDependencies/UD_Georgian-GLC/tree/dev/ და https://github.com/UniversalDependencies/docs/blob/pages-source/_ka/, ბოლო წვდომა; 2023 წლის 2 დეკემბერი.

ლურ-ტიპოლოგიურს (გრინბერგი 1966 და სხვ.) და ფორმალურ-გენერაციულს (ჩომსკი 1976 და სხვ.). ბუნებრივი ენის დამუშავების ფარგლებში ზემოაღნიშნული თეორიული ჩარჩოების გამოყენება, ძირითადად, ორიენტირებულია ერთიან მიდგომაზე, რომელიც უზრუნველყოფს კროსლინგვისტურ ენობრივ აღწერას მეტყველების ნაწილების (PoS) ტეგირების, კომპონენტებისა და დამოკიდებულებათა პარსინგისა და კორპუსზე დაფუძნებული მიდგომების დონეზე. ქართულისათვის ანოტირების სქემის შემუშავების თეორიულ საფუძველს წარმოადგენს ენობრივი უნივერსალიებისა და შესაბამისი ქართული ენისათვის დამახასიათებელი სინტაქსური სტრუქტურების აღწერა, ხოლო პრაქტიკულ საფუძველს – კომპიუტერული ენობრივი რესურსების გამოყენება სინტაქსური ანალიზისთვის. აღნიშნული წინაპირობები გამოიყენება შემდეგი მიზნების მისაღწევად: ა) ქართული ენის სინტაქსური ფუნქციებისა და დამოკიდებულებების განსაზღვრისათვის; ბ) ქართულის ენის ანოტირების სახელმძღვანელოს შესადგენად; გ) ცალკეული მეტყველების ნაწილებისათვის (PoS) უნივერსალურ დამოკიდებულებათა ანოტირების სქემის შესამუშავებლად წინადადების შემდეგი ტიპებისათვის: მარტივი, რთული თანწყობილი, რთული ქვეწყობილი; დ) სატესტო ხეთა ბანკის შესაქმნელად. ზემოაღნიშნული მიზნების მიღწევა მჭიდროდ უკავშირდება დამოკიდებულებათა გრამატიკის მეთოდებს, რომლებიც გამოიყენება ქართული წინადადებების მონაცემთა სახით წარმოსადგენად, რადგანაც ქართულს, ისევე როგორც სხვა ენებს, აქვს იერარქიული სტრუქტურა (კვაჭაძე 1996), რომელიც მოიცავს სახელურ შესიტყვებებს, ზმნურ შესიტყვებებს და ა.შ.

არსებული მონაცემების შეჯერების მიზნით, ქართული ენის არსებული ტეგსეტები (Lobzhanidze 2021, 2022) შედარდა უნივერსალურ დამოკიდებულებათა ტეგსეტს და შემუშავდა უნივერსალურ დამოკიდებულებათა ქართული ენის შესაბამისი ტეგსეტების ორი სია:

- უნივერსალური მეტყველების ნაწილების (UPOS) ტეგების სია ემსახურება მეტყველების ძირითადი ნაწილების მონიშვნას. UPOS ტეგების გარდა, უნივერსალურ დამოკიდებულებათა (CoNLL-U) ფორმატი უზრუნველყოფს ქართული ენისათვის დამახასიათებელი მეტყველების ნაწილების ტეგირებას პირობითი ენობრივი (XPOS) ტეგების გამოყენებით. შესაბამისად, .conllu ფორმატის სატესტო ფაილიც მოიცავს შესაბამის ველს, რომელიც ემსახურება XPOS ტეგების ჩვენებას. ტეგსეტების შეჯერებისას გამოვლინდა რამდენიმე განმასხვავებელი ნიშანი, რომლითაც უნივერსალური ტეგსეტი განსხვავდება ქართული

ენის ტრადიციული ტეგსეტისგან, კერძოდ, უნივერსალური დამოკიდებულებების მიხედვით, ზოგადი არსებითი სახელები აღინიშნება როგორც NOUN, ხოლო შესაბამისი საკუთარი არსებითი სახელები როგორც PROPN. მაშინ როცა ქართული ენის ამოსავალ ტეგსეტებში (2021, 2022) ზოგადი არსებითი სახელები წარმოდგენილია როგორც +Noun ან Nc, ხოლო საკუთარი სახელები – როგორც +Noun+Prop ან Np.

- ლექსიკური თვისებების (FEATS) ტეგების სია, რომელიც იყოფა ლექსიკურ თვისებებად, მაგ. ნაცვალსახელის ტიპი (PronType), რიცხვითი სახელის ტიპი (NumType) და სხვ. და ფლექსიურ თვისებებად სახელური და ზმნური პარადიგმებისათვის, მაგ. რიცხვი (Number), ბრუნვა (Case) და სხვ. სახელური და ზმნური პარადიგმების ტეგების სიები, ბუნებრივია, განასხვავებენ თვისებებს, რომლებიც გამოიყენება მხოლოდ სახელებთან ან მხოლოდ ზმნებთან. გამოვლინდა, ასევე, ზოგიერთი ქართული ენისთვის დამახასიათებელი თავისებურებაც, მაგ. მიახლოებითი რიცხვითი სახელები, გვარის სხვადასხვა ტიპი და ა.შ.

ქართული ენის სინტაქსური ტეგები წარმოდგენილია უნივერსალურ დამოკიდებულებათა ტეგების სიის სახით და შედარებულია უნივერსალურ დამოკიდებულებათა სიასთან, რომელიც ხელმისაწვდომია https://universaldependencies.org/u/dep/index.html. ქართული ენისთვის დამახასიათებელი ინფორმაცია, როგორიცაა ტოკენიზაცია და სეგმენტაცია, ინფორმაცია ქართული ენის მორფოლოგიისა და სინტაქსის შესახებ, ინფორმაცია ზემოაღნიშნული ტეგებისა და თვისებების შესახებ, აიტვირთა GitHub-ის საცავში ამოსავალი introduction.md და index.md ფაილების სახით.[7]

ნაწილი 3. ქართული ენის სინტაქსური ანოტირების პრინციპები

ქართული ენის კორპუსი (დობორჯგინიძე და ლობჟანიძე 2012-2019) და ქართული ენის ანოტირების ხელსაწყოები, როგორიცაა: ტოკენიზატორი, მორფოლოგიური ანალიზატორი და გენერატორი (Lobzhanidze 2022) გამოიყენება ცალკეული ტოკენების ლემატიზაციისა და მორფოსინტაქსური ანოტირებისათვის. აღნიშნულ რესურსებს დაემატა სინტაქსური ფუნქციების მონიშვნის სკრიპტები. შესაბამისად, პროექტზე მუშაობისას მოხდა თეორიული და კომპიუტერული მიდგომების გამოყენება წინადადებების საზღვრე-

7 იხ. https://github.com/UniversalDependencies/docs/blob/pages-source/_ka/, ბოლო წვდომა: 2023 წლის 2 დეკემბერი.

ბისა და სინტაქსური დამოკიდებულებების მისანიჭებლად, კერძოდ:

- ქართული ენის სინტაქსური ფუნქციებისა და დამოკიდებულებების განსაზღვრა, რომელიც მოიცავს შესიტყვების სტრუქტურის მონაცემთა აღწერასა და ანალიზს, წინადადების შემადგენელ ნაწილებს შორის ფუნქციონალური ურთიერთობებისა და ქართული ენის დამოკიდებულებათა მიმოხილვას;
- ქართულის სინტაქსური ანოტირების ინსტრუქციების შედგენა, რომელშიც წარმოდგენილია ინფორმაცია სახელური შესიტყვებების (მაგ. განსაზღვრებისა და ფუნქციური სიტყვების დამოკიდებულების და ა.შ.) შესახებ, მარტივი (მაგ. გარდაუვალი და გარდამავალი წინადადებების, ვალენტობის ცვლილებისა და ა.შ.) და რთული (მაგ. ქვეწყობილი და თანწყობილი) წინადადებებისა და სხვა კონსტრუქციების შესახებ (საჭიროების შემთხვევაში);
- უნივერსალურ დამოკიდებულებათა ანოტირების სქემების შემუშავება ცალკეული მეტყველების ნაწილებისათვის წინადადებების ტიპების გათვალისწინებით: მარტივი, რთული ქვეწყობილი, რთული თანწყობილი, რომელიც მოიცავს დოკუმენტირებული ტეგების, თვისებებისა და სინტაქსური ურთიერთობების მიმოხილვას სინტაქსური ანოტირების წინასწარ შემუშავებული ინსტრუქციების შესაბამისად, რომლებიც მოიცავენ ინფორმაციას სიტყვებს შორის დამოკიდებულებათა ურთიერთობების, ფუნქციონალური სიტყვების სტატუსისა და შერჩეული დამოკიდებულებების ტაქსონომიის შესახებ.

ქართული ენის წინასწარ შემუშავებული საწყისი ფაილი (Georgian-ud-test.conllu) მოიცავს 151 წინადადებას, რაც 2123 ტოკენს შეადგენს. წინადადებების რაოდენობა განისაზღვრა მოდელის მინიმალური მოთხოვნების დასაკმაყოფილებლად. მოდელში იგულისხმება დასწავლის გზით UDPIPE[8] რესურსის გამოყენებით მოდელის შექმნა. წინადადებები შემთხვევითი პრინციპით შეირჩა ქართული ენის კორპუსიდან (დობორჯგინიძე და სხვ. 2012-2019). შემუშავებული ფაილი 10 სვეტისგან შედგება, მათ შორის:

- ID: დამოკიდებულების ხის მიხედვით სეგმენტირებული წინადადება, რომლის ტოკენიზაცია შესრულებულია ქართული ენის კორპუსის ანოტირების მიხედვით, რომელსაც შეემატა მრავალსიტყვიანი ტოკენების ავტომატური ტოკენიზაცია. წინადადებების ნუმერაცია 00001-დან იწყება;

8 იხ. https://ufal.mff.cuni.cz/udpipe/2/models, ბოლო წვდომა: 2023 წლის 2 დეკემბერი.

- FORM: სიტყვის ფორმა ან პუნქტუაციის სიმბოლო ქართული ენის კორპუსის ანოტირებიდან;
- LEMMA: სიტყვის ამოსავალი ლემა, რომელიც ავტომატურად განისაზღვრება ქართული ენის კორპუსის ანოტირებიდან;
- UPOS: მეტყველების ნაწილების უნივერსალური ტეგები, რომლებიც უკავშირდება ქართული ენის კორპუსის ამოსავალ ანოტირებას;
- XPOS: ენის მეტყველების ნაწილის ტეგსეტი დობორჯგინიძის და ლობჟანიძის (2012-2019) და ლობჟანიძის (2022) მიხედვით;
- FEATS: მორფოლოგიური თვისებების სია, რომელიც შედგენილია ქართული ენის კორპუსის ანოტირების მიხედვით;
- HEAD: მიმდინარე სიტყვის თავი ამოიცნობა და მოინიშნება ავტომატურად ხელით ჩასწორების გზით;
- DEPREL: სიტყვების უნივერსალური დამოკიდებულებები ამოიცნობა და წინადადების თავს (HEAD) დაუკავშირდება ავტომატურად ხელით ჩასწორების გზით;
- DEPS: დამოკიდებულების გრაფიკი მეთაური წყვილების სახით გასნაზღვრება ავტომატურად ხელით ჩასწორების გზით;
- MISC: მოიცავს ინფორმაციას ტოკენების ავტომატური სეგმენტირებისა და ტრანსლიტერაციის შესახებ.

ნაწილი 4. დასკვნები

უნივერსალურ დამოკიდებულებათა ხელსაწყოების გამოყენებით შესაძლებელი ხდება ხეთა ბანკის კროსლინგვისტური თანმიმდევრული ანოტირება, რაც მნიშვნელოვანია მულტილინგვური პარსერებისა და კროსლინგვისტური დასწავლის უზრუნველყოფისათვის ენობრივი ტიპოლოგიის პერსპექტივის გათვალისწინებით. და ვინაიდან დღემდე არ არსებობს ქართული ენის ხეთა ბანკი, უნივერსალური დამოკიდებულებების სინტაქსური მოდელის კვლევა ძალიან მნიშვნელოვანია სხვადასხვა პერსპექტივიდან გამომდინარე. ამრიგად, ანგარიშში წარმოდგენილია ქართულის სინტაქსური ანოტირების სქემის შექმნის საწყისი ეტაპი, რომელიც ემსახურება ქართული ენობრივი მონაცემების თავსებადობის უზრუნველყოფას კროსლინგვისტურ ჭრილში და სამომავლოდ გამოყენებული იქნება ქართული ენის ხეთა ბანკის შესაქმნელად. ამ მომენტისათვის ზოგიერთი თვისება კიდევ უნდა დაემატოს ენობრივ დოკუმენტაციას და საცავი ხელახლა განახლდება ივნისისათვის. სამომავლო პერსპექტივები მჭიდროდ

უკავშირდება ქართული ენის ლინგვისტური მოდელის გაწვრთნას UDPIPE-ის გამოყენებით.

დამოწმებანი

დობორჯგინიძე, ნინო და ირინა ლობჟანიძე. 2012-2019. *ქართული ენის კორპუსი.* თბილისი. http://corpora.iliauni.edu.ge/.

კვაჭაძე, ლეო. 1996. *თანამედროვე ქართული ენის სინტაქსი*. თბილისი: რუბიკონი.

Chomsky, Noam. 1976. *Reflections on Language*. London: Temple Smith.

de Marneffe, Marie-Catherine, Bill MacCartney and Christopher D. Manning. 2021. Generating Typed Dependency Parses from Phrase Structure Parses. Computational Linguistics, vol. 47, Issue 2: 255-308.

Greenberg, Joseph. 1966. *Universals of Language.* Cambridge: MIT Press.

Lobzhanidze, Irina. 2022. *Finite-State Computational Morphology: An Analyzer and Generator for Georgian.* Cham: Springer.

----------. 2021. *MULTEXT-East Morphosyntactic Specifications, Revised Version 6: Georgian Specifications.* August 20. http://nl.ijs.si/ME/V6/msd/html/msd-ka.html.

Nivre, Joakim, Bandmann Megyesi, Beáta. 2007. Bootstrapping a Swedish Treebank Using Cross-Corpus Harmonization and Annotation Projection. In *Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories,* eds. Koenraad De Smedt and Jan Hajič and Sandra Kübler. NEALT Proceedings Series, vol. 1: 97-102. Bergen: Northern European Association for Language Technology (NEALT).

Nivre, Joakim and Chiao-Ting Fang. 2017. Universal Dependency Evaluation. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW),* eds. Marie-Catherine de Marneffe, Joakim Nivre and Sebastian Schuster, 86–95. Gothenburg, Sweden: Association for Computational Linguistics.