ოლეგ კაპანაძე, ნუნუ კაპანაძე

ქართული წინადადების ავტომატური ანალიზი

1. შესავალი

ბუნებრივი ენის ტექნოლოგია არის თანამედროვე კომპიუტერული მეცნიერების სფერო, რომელიც ეფუძნება ტექსტური ინფორმაციის კომპიუტერის მეშვეობით დამუშავებას. მისი მიზანია კომპიუტერული სისტემების შექმნა, რომლებიც შეძლებენ დიდი მოცულობის ტექსტის დამუშავებას (Big Data Processing), მსოფლიო ქსელებში სასურველი ინფორმაციის მოძიებას, ადამიანსა და კომპიუტერს შორის დიალოგის წარმართვას, ბუნებრივი ენის ტექსტისა და მეტყველების ავტომატურ თარგმნას და გამოყენებითი ხასიათის სხვა ამოცანების გადაწყვეტას.

საინფორმაციო და საკომუნიკაციო ტექნოლოგიების სულ უფრო მზარდი გამოყენება, ბუნებრივია, უპირატესობას იმ ენებს ანიჭებს, რომლებიც წარმატებულად მუშავდება კომპიუტერული პროგრამების მეშვეობით. თანამედროვე პროგრამული საშუალებებით უზრუნველყოფილი ენები, რომლებიც ბუნებრივი ენის ტექსტის დასამუშავებლად მძლავრ კომპიუტერულ ტექნოლოგიას იყენებენ, საინფორმაციო მომსახურების თითქმის შეუზღუდავ შესაძლებლობას გვთავაზობენ.

იმის გათვალისწინებით, რომ ენის ტექნოლოგია მსოფლიოში სულ უფრო რეალურად ხელმისაწვდომი ხდება, უპირატესობა განსაკუთრებულად იმ ენებს ენიჭებათ, რომლებზეც ტექსტის შინაარსი ციფრულადაა გადმოცემული და ეს გარემოება არსებით მნიშვნელობას იძენს ნებისმიერი ენის მომავლისთვის. ამდენად, ისეთი ენების „ციფრული სიცოცხლისუნარიანობა“, რომელთაც არ გააჩნიათ კომპიუტერულად მხარდაჭერილი გარემო, საფრთხის წინაშე დგება.

ქვემოთ განხილულია ჩატარებული სამუშაო ქართული ენის ტექსტის მეტყველების ნაწილებად თეგირებისა (tagging) და სალექსიკონო ერთეულებად (lemmatizing) დაყვანისთვის საჭირო ხელსაწყოების ყუთის (toolkit) შექმნის მიზნით, რომელიც საშუალებას იძლევა, შევიმუშაოთ ზმნის სინტაქსურ ვალენტობაზე დამყარებული ენისთვის აუცილებელი ტექნოლოგიური რესურსი – მარტივი წინადადების ავტომატური სინტაქსური ანალიზატორი.

წინამდებარე სტატია წარმოადგენს ქართული ენისთვის კომპიუტერული რესურსების შექმნის ძალისხმევას, რათა შემცირდეს თანამედროვე საინფორმაციო ტექნოლოგიების პირობებში მოწინავე ენებთან არსებული ჩამორჩენა.

2. მორფოპარსერი/POS-თეგერი ქართული ენისთვის

ქართული ენისთვის შემუშავებული აკადემიური გრამატიკებისა და ლექსიკონების სიმრავლის მიუხედავად, ეს რესურსები არ აკმაყოფილებენ კომპიუტერული ხელსაწყოებისადმი არსებულ ტექნოლოგიურ მოთხოვნილებებს. პირველ რიგში, ეს გამოწვეულია იმით, რომ ისინი არ არიან ხელმისაწვდომი იმ ფორმით, რომელიც შესაძლებელს გახდის მათ გამოყენებას ქართული ენის ტექსტის კომპიუტერული პროგრამით დამუშავებისას.

ქართული აგლუტინაციური ენაა, რომელიც სიტყვაფორმის წარმოებისთვის იყენებს როგორც სუფიქსებს, ასევე პრეფიქსებს. ვინაიდან ავტომატური მორფოლოგიური ანალიზი აგლუტინაციური ენებისთვის ერთ-ერთი უმთავრესი ამოცანაა, მორფოპარსერი განიხილება, როგორც აუცილებელი კომპიუტერული რესურსი ქართული ტექსტის ანალიზისთვის.

მიმდინარე ათასწლეულის დაწყებამდე, საბაზისო ტექნოლოგიების განვითარების თვალსაზრისით მხოლოდ „მთავარი ევროპული ენები“ იყვნენ განებივრებული, რომელთა შორის ინგლისური ენა შეუვალი ლიდერი იყო. შესაბამისად, ბუნებრივი ენის ტექნოლოგიის სფეროში მეთოდოლოგიის დიდი ნაწილი შემუშავებულია ინგლისურისთვის, რომელიც მწირი მორფოლოგიით გამოირჩევა. ასეთი ტიპოლოგიური მახასიათებლიდან გამომდინარე, მისი სინტაქსი მეტად შეზღუდულ სტრუქტურას გვიჩვენებს და, ამის გამო, მას სპეციალურ ლიტერატურაში „მკაცრად კონფიგურაციულ“ (strongly configurational) ენად მიიჩნევენ. ბუნებრივი ენების ტექნოლოგიისათვის განკუთვნილი მეთოდოლოგიის უდიდესი ნაწილი, ინგლისურის გარდა, ძირითადად შემუშავებულია სინტაქსურ კონფიგურაციაზე მკაცრი შეზღუდვების მქონე ენებისთვის.

კონფიგურაციული სპექტრის საპირისპირო მხარეს წარმოდგენილია ენები (ფინური, უნგრული, ბასკური, თურქული, ქართული და ა.შ), რომლებიც, ამ მხრივ, ძირეულად განსხვავებულია ინგლისურისაგან. მათი სტრუქტურა მდიდარი მორფოლოგიით გამოირჩევა და, აქედან გამომდინარე, ეს ენები სინტაქსურ შემადგენელთა თავისუფალი რიგით ხასიათდებიან. შესაბამისად, ამ ენებში სინტაქსურ სტრუქტურებს გაცილებით მცირე შეზღუდვები ადევთ წინადადების დონეზე. ამიტომ ამ ტიპის ენებს, მათი ფართო მორფოლოგიური შესაძლებლობებისა და სინტაქსური თვალსაზრისით ნაკლებად შეზღუდულობის გამო, მდიდარი მორფოლოგიისა და ნაკლებად კონფიგურაციულ ენებს (Morphologically Rich and Less-Configurational Languages-MR&LC) უწოდებენ (Fraser, et al., 2013: 58).

ჩვენ მიერ გამოყენებული ქართული ენის მორფოლოგიური გადამსახავი ეყრდნობა პროგრამულ ბიბლიოთეკებს, რომელსაც პროფესიულ წრეებში „ქსეროქსის სასრული მდგომარეობის მთვლელს“ (XEROX Finite-State Calculus – FST) უწოდებენ და სასრული მდგომარეობის ავტომატთა მეთოდოლოგიას ეფუძნება (Beesly et al., 2003). ხსენებული მიდგომის მთავარი ამოსავალია ის, რომ ბუნებრივი ენის მორფოლოგიური ანალიზატორი შეიძლება აიგოს ლინგვისტურ მონაცემთა სტრუქტურების (Linguistic Data Structures) გამოყენებით, რომელსაც სასრული მდგომარეობის ქსელი ეწოდება. ქსეროქსის მთვლელი ენის ლექსიკოგრაფიული ნაწილის აღწერისას იყენებს Lexc პროგრამულ ხელსაწყოს. Lexc ემყარება კონტექსტისაგან დამოუკიდებელი მორფოსინტაქსური წრფივი გრამატიკის (simple right-linear morphosyntactic context-free grammar) ფორმალიზმს, რომელშიც პროდუქციის წესების მარჯვენა მხარე შეიძლება მხოლოდ ერთ არატერმინალურ ერთეულს შეიცავდეს. ამის გათვალისწინებით, ნებისმიერი ბუნებრივი ენის მორფოლოგიური აღწერისას გვაქვს რანგებად დაყოფილი მორფემების ურთიერთგამომრიცხავი სიმრავლეები (ლექსიკონების ერთობლიობა), რომლებიც იწყება ძირების ლექსიკონით (root lexicon). ყოველ ლექსიკონში თითოეულ მორფემას აქვს მისი განმვრცობი ლექსიკონი, რომლებიც, თავის მხრივ, განსაზღვრავენ მორფემების კონკრეტულ სიმრავლეს და მორფემების ჯაჭვში შესაძლოა მოცემული მორფემის გაგრძელებად იქცნენ (Szántó, et al., 2014: 136).

ქართული აკადემიური გრამატიკიდან გამომდინარე, არსებითი სახელების, ნაცვალსახელებისა და ზედსართავი სახელების ღემების ყოველი სიმრავლე განიხილება, როგორც ძირების ერთიანი ლექსიკონი (N_St). ძირებზე დართული მრავალრიცხოვანი სუფიქსების მიმდევრობები ქმნიან სიტყვის გრამატიკულად სწორ ფორმას. მაგალითად, თითოეულ ძირს მარჯვნივ შეიძლება დაერთოს მრავლობითი რიცხვის მარკერი (PL_MK), რომელსაც, თავის მხრივ, შეიძლება მოჰყვებოდეს ბრუნვის შვიდი ნიშნიდან ერთ-ერთი (C_MK), ხოლო მას – თანდებული (PSTF). ეს უკანასკნელი შეიძლება გავრცობილი იყოს ერთმანეთის თანმიმდევრობით ემპათიკური ხმოვნითა (Eph_V) და მავრცობით (Eph_ PT):

N_St + *PL_MK* **+ C_MK +** *Eph_V* **+** *PSTF* + *Eph_V* + *Eph_ PT*

მოცემულ ჯაჭვში მხოლოდ N_St და C_MK წარმოადგენენ აუცილებელ ელემენტებს.

აღწერილი სქემისგან განსხვავებით, ზმნის სასრული ფორმის საწარმოებლად საჭიროა როგორც სუფიქსები, ისე პრეფიქსები.

$$A + B + C + \mathbf{ROOT} + E + F + D + H + G$$

A არის ზმნისწინების ლექსიკონი (სიმრავლე), რომელსაც შეიძლება მოსდევდეს ზმნის სუბიექტ-ობიექტის მარკერების ლექსიკონი, მას კი, თავის მხრივ, გვარისა და ქცევის კატეგორიის მორფემების C ლექსიკონი. B და C ლექსიკონი ერთმანეთთან დამატებითი დისტრიბუციის მიმართებაში იმყოფებიან. ზმნის მორფემების ჯაჭვში მეოთხე ელემენტი ზმნის ძირების ROOT ლექსიკონს აღნიშნავს. ამ ლექსიკონიდან ნებისმიერი ძირი თეორიულად შეიძლება გავრცობილ იქნეს სუფიქსებით მომდევნო ხუთი *(E + F + D + H + G)* ლექსიკონიდან. საერთო ჯამში, ქართული ზმნის სასრული ფორმა შეიძლება აიგოს მარცხნიდან მარჯვნივ, ძირის ჩათვლით, მორფემების ცხრა ლექსიკონზე დაყრდნობით.

XEROX-ის სასრული მდგომარეობის ქართული ენის მორფოპარსერი ეფუძნება სქემას, სადაც ზმნის ძირების ლექსიკონში შეტანილ ყოველ ერთეულს შეიძლება წინ უძღოდეს სამი პრეფიქსალური მწკრივი (A, B, C) და მოსდევდეს სუფიქსების შესაძლო ხუთი სხვადასხვა სიმრავლე. თუმცა, გამონაკლის შემთხვევებში, უღლებადი ზმნის ფორმა შეიძლება ასევე შეგვხვდეს ზმნური ძირების სიმრავლის უშუალო ლექსიკური ერთეულის სახით, როდესაც აფიქსები წარმოდგენილია სამი პრეფიქსული რანგიდან და მომდევნო სუფიქსების ხუთი რიგიდან ნულოვანი ალომორფებით. მაგალითად, „წერ“ (ერთვალენტიანი ზმნა მეორე პირის მხოლობითში).

სადღეისოდ შემუშავებული გვაქვს ქართული FST მორფოპარსერის სამი ვერსია სხვადასხვა გამოსავალი ფორმატით: პირველი, TIGER-XML ფორმატით (Brants, et al., 2000: 1644ff), რომელიც საჭიროა სინტაქსურად ანოტირებული ქართული ხეების ბანკის ასაგებად; მეორე, შესავალი ტექსტიდან მორფოლოგიურად გაანალიზებული სიტყვებისათვის თანდართული ლემებით, და მესამე, გამარტივებული ვარიანტი ლემატიზებული შემავალი სიტყვების გარეშე. თითოეულ ჩამოთვლილ ვერსიას შეუძლია შესავალი ქართული ტექსტის ტოკენიზაცია, POS-თეგირება და მორფოლოგიური ანოტირება. მცირედი მანუალური რედაქტირების შემდეგ, მორფოპარსერის შედეგებმა შეიძლება თითქმის 100%-მდე სიზუსტეს მიაღწიოს.

მორფოლოგიური ანალიზის შედეგი მარტივი წინადადებისათვის **„კაცები აქებენ ჭკვიან ქალებს“** მოყვანილია ქვემოთ:

კაცები
<lemma='კაც 'morph="Pl.Nom", pos="NN"/>
აქებენ
TV[VAL=ND, SR=1, voice=ACT, mood=IND, prs=Pl3]
ჭკვიან
<lemma='ჭკვიან 'morph="Sg*.Stm", pos="Adj"/>
ქალებს
<lemma='ქალ' morph="Pl.Dat", pos="NN"/>
<lemma='ქალ' morph="Pl.Dat", pos="NN"/>
.
<lemma="--" pos="$."/>

ლემით ანოტირებული წინადადების ერთეულები (ტოკენები) დაჭდევებულია მეტყველების ნაწილების თეგებით (POS-tags). მაგალითში ეს არის ჩვეულებრივი არსებითი სახელი NN (Normal Noun), გარდამავალი ზმნა TV (Transitive Verb), ზედსართავი სახელი Adj და წინადადების ბოლო $. (end of clause) ლემის გარეშე.

არსებითი სახელები გაჯერებულია მორფოლოგიური მახასიათებლებით გრამატიკული რიცხვისათვის – Pl(ural) და Sg*(singular tantum) და ბრუნვის მარკერებით – Nom (Nominative) და Dat (Dative). პუნქტუაციის ნიშნები (pos="$,"/ pos="$.") და მორფოლოგიური ფლექსიის არმქონე წინადადების წევრები ლემის გარეშე (lemma ="–".) არის მოცემული. გარდამავალი ზმნა (TV) ანოტირებულია სინტაქსური ვალენტობის მაჩვენებლით (VAL), რომელიც ასახავს წინადადების სინტაქსური შემადგენლების მეთაური სიტყვების (headword) ბრუნვებს: D(ative)N(ominative). ასევე, მითითებულია სერიის (SR=1), გვარის voice=ACT(ive), კილოს mood=IND(icative), პირისა და რიცხვის prs=Pl3 (3rd person plural) გრამატიკული მახასიათებლები.

მოცემულ მაგალითში lemma='**ქალ**' ორჯერადი გენერირება განპირობებულია ქართული ენის ლექსიკონში ისეთი ორი არსებითი სახელით, როგორიც არის – 'ქალი' და 'ქალა'. მორფოპარსერი ორივე შემთხვევაში ანალიზის შედეგად გამოსცემს ერთსა და იმავე lemma='**ქალ**' ერთეულს ბრუნვის მარკერის გარეშე.

ამ სახით აღწერილი მორფოპარსერის გამოსავალი მონაცემები, ავტომატურ სინტაქსურ ანალიზატორთან ერთად, ენის ტექსტის დამუშავებისათვის გამიზნული ნაკადსადენის (pipe-line) ინტეგრალურ ნაწილს წარმოადგენს.

3. ქართული წინადადების ავტომატური სინტაქსური სეგმენტაცია

წინადადების სინტაქსური სეგმენტაცია (Syntactic Chunking), რომელიც ასევე ცნობილია ზედაპირული სინტაქსური ანალიზის (Shallow Parsing) სახელით, მიზნად ისახავს იმ ტიპის სინტაქსური შემადგენლების იდენტიფიცირებას წინადადებაში, როგორიცაა არსებითი სახელის ან ზმნის ფრაზა. თეორიულად, ის შეიძლება შედგებოდეს, VP (ზმნის ფრაზის) და ზმნის ვალენტობიდან გამომდინარე, აუცილებელი NP (არსებითი სახელის) ფრაზებისაგან. წინადადებაში NP-ის რაოდენობა განისაზღვრება ქართული ენის სასრული ზმნის ფორმის სინტაქსური ვალენტობის (VAL) პარამეტრით. VAL გამოითვლება შესავალი ტექსტის მორფოლოგიური ანალიზის პროცესში, ვინაიდან ცნობილია, რომ სინტაქსური ვალენტობის ჩარჩო დამოკიდებულია ზმნის გარდამავლობაზე და შეიძლება შეიცვალოს სერიების მიხედვით.

ქართულ აკადემიურ გრამატიკაში მიღებული სამი სერიის მიხედვით, სინტაქსური ვალენტობის, NP-ების და მათი შესაბამისი მეთაური სიტყვების (headword) ბრუნვის მარკერების განაწილება წარმოდგენილია ცხრილში, რომელიც განსაზღვრავს სინტაქსურ ჩარჩოებს გარდამავალი (TV) და გარდაუვალი (IV) ზმნების სიმრავლის რვა კლასტერისთვის. ცხრილი ეყრდნობა დ. მელიქიშვილის (2001: 60ff) მიერ შემოთავაზებულ ზმნის უღლების სისტემას:

ცხრილი 1. სინტაქსური ჩარჩოების დისტრიბუცია ქართული ზმნების სიმრავლეებისთვის

ლასტერი	I სერია	II სერია	III სერია
IV01	VAL=N	VAL=E	VAL=D
TV02	VAL=NDD	VAL=END	VAL=DN (D+postf)
TV03	VAL=ND	VAL=EN	VAL=DN
IV04	VAL=N	VAL=N	VAL=N
IV05	VAL=ND	VAL=ND	VAL=ND
IV06	VAL=DN	VAL=DN	VAL=DN
IV07	VAL=D	VAL=D	VAL=D
IV08	VAL=zero	VAL= zero	VAL= zero

ცხრილში VAL მიუთითებს სასრული ზმნის ფორმის საშუალებით გამოხატული სინტაქსური ვალენტობის პარამეტრს. ცხრილის მეორე სტრიქონში VAL-ი TV02 კლასტერისთვის I სერიაში არის სამვალენტიანი, NP-ს (სუბიექტი) მეთაური სიტყვით სახელობითში (N) და პირდაპირი და ირიბი ობიექტის NP-თვის მიცემითში (D).

VAL პარამეტრი II სერიაში NP ფრაზული შემადგენლების მეთაური სიტყვებისათვის, შესაბამისად, გვიჩვენებს მოთხრობით (E), სახელობით (N) N და მიცემით (D) D ბრუნვებს. III სერიაში ის დადის ორვალენტიანი ზმნის ჩარჩოზე, რადგან მეორე ფრაზული შემადგენელი მიცემითში თანდებულით (D+postf) ჩნდება და არ შეიძლება იყოს იდენტიფიცირებული, როგორც სინტაქსურ ჩარჩოში ვალენტური ადგილის მფლობელი.

რაც შეეხება მერვე კლასტერს, სადაც VAL=zero, იგი ქართულ ენაში ბუნების მოვლენების აღსაწერ ზმნების მცირერიცხოვან სიმრავლეს შეესაბამება. ცნობილია, რომ მათ სინტაქსური ფუნქციის მქონე ფრაზული შემადგენლები არ ახლავთ წინადადებაში.

ქართული ენის ზმნური მასივის სინტაქსური ვალენტობის ჩარჩოების კლასტერიზება წარმოადგენს აუცილებელ წინაპირობას ქართული მარტივი წინადადების ავტომატურ ფრაზულ შემადგენლებად სეგმენტირებისათვის. როგორც ზემოთ იყო განმარტებული, VAL პარამეტრში ასომთავრული (N, E, D) აღნიშნავს შესატყვის NP ფრაზებში მეთაური სიტყვების ბრუნვის მარკერებს, რომელთა რაოდენობა წინადადებაში შეიძლება მერყეობდეს [0, 1, 2, 3] ფარგლებში. ანალიზის პროცედურა ტექსტის NP-ი ფრაზული სეგმენტების მისაღებად გულისხმობს თითოეული NP-თვის მეთაური სიტყვის (არსებითი სახელი ან მისი ეკვივალენტი) მაქსიმალური პროექციის გამოთვლას შესავალ ტექსტში.

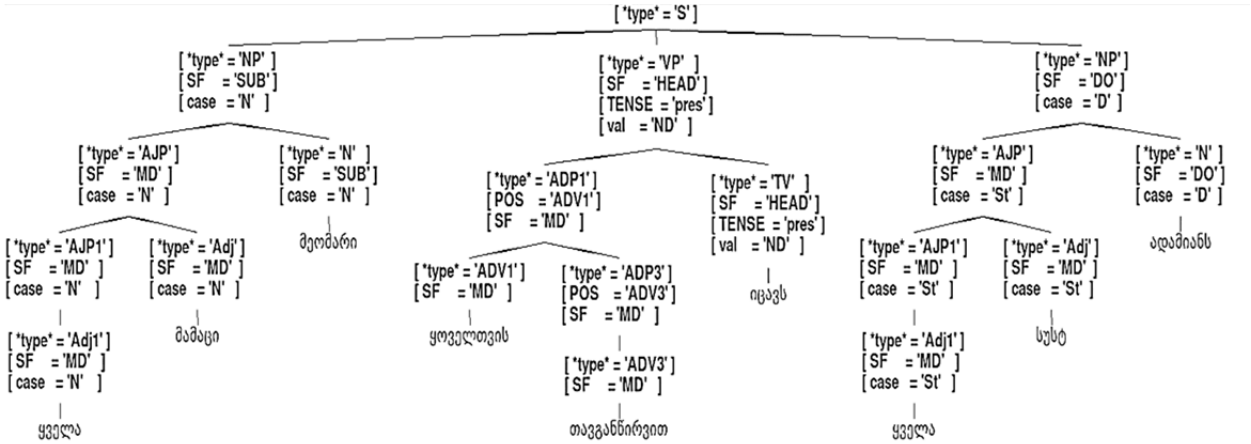
NP-ების რაოდენობა წინადადებაში (resp. გრაფი/სინტაქსური ხე) განისაზღვრება წინადადების მეთაური სიტყვის (ზმნის სასრული ფორმის) სინტაქსური ვალენტობით, რომელიც იდენტიფიცირებულია მორფოპარსერის მიერ და დაფიქსირებულია, როგორც ერთ-ერთი მახასიათებელი VP-ში.

მას შემდეგ, რაც ყველა სავალდებულო და შესაძლო შემადგენელი, როგორც VP-სა და NP-ების მაქსიმალური პროექცია, იდენტიფიცირებული იქნება, სინტაქსურ სეგმენტატორს შეუძლია, ლექსიკური ერთეულების წარმოქმნის წესებზე (Lexical Production Rules – LPR) დაყრდნობით, დააგენერიროს წინადადების სინტაქსური შემადგენლების სტრუქტურა სინტაქსური ხის სახით.

აღწერილი სქემის პრაქტიკულად განხორციელების შესაძლებლობა შემოწმდა Natural Language Toolkit (NLTK) პროგრამული ბიბლიოთეკის მეშვეობით, რომელიც შემუშავებულია Python პროგრამირების ენაზე (Bird, et al., 2009). აღწერილ პრინციპებზე და ქართული ენის LPR-ით შედგენილ გრამატიკაზე დაყრდნობით, NLTK გამოიყენება მორფოლოგიურად და სინტაქსურად ანოტირებული სინტაქსური ხეების ავტომატურად ასაგებად. LPR წესები შეიძლება აიგოს სრულად ხელით ან ნახევრად ავტომატურად. ბოლო ვარიანტი ითვალისწინებს, რომ პარსერი/ანალიზატორი ტექსტს ავტომატურად ანიჭებს სინტაქსურ სტრუქტურას, რომელიც საბოლოოდ შემოწმდება ლინგვისტების მიერ და, საჭიროების შემთხვევაში, შესწორდება.

სურათზე 1 გამოსახულია NLTK მახასიათებლებზე ორიენტირებული კონტექსტისგან დამოუკიდებელი (Feature-Based Context-Free Grammar – FBCFG) პარსერის მუშაობის შედეგი ორვალენტიანი გარდამავალი ზმნის შემთხვევაში ქართული ენის წინადადებისთვის:

„ყველა მამაცი მეომარი ყოველთვის თავგანწირვით იცავს ყველა სუსტ ადამიანს“



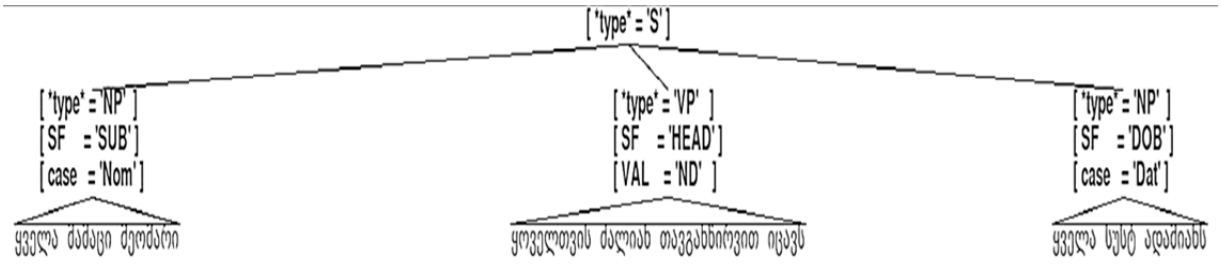
სურათი 1

NLTK-ს გამოსავალი შედგება სინტაქსურ ხეებზე ორიენტირებული გრაფიკული გამოსახულებისაგან. მასში წარმოდგენილ სხვადასხვა დონის სინტაქსური ხის სტრუქტურებისთვის ხის თითოეულ ტერმინალურ ელემენტს („ფოთოლს“ = ტოკენს) და თითოეულ კვანძს (= ფრაზულ შემადგენელს) მინიჭებული აქვს ანალიზისას მისთვის გამოთვლილი სინტაქსური კლასის იდენტიფიკატორი (‘type’). შესავალი წინადადება ვიზუალურად წარმოადგენს გრაფს, რომელიც ემყარება ჰიბრიდულ მიდგომას და აერთიანებს ანოტირებას ფრაზული შემადგენლებისა და მათი შესაბამისი სინტაქსური ფუნქციების მიხედვით. გრაფის კვანძები ასახავს წესებს, რომლებიც რეალიზებულია მახასიათებლებზე ორიენტირებულ კონტექსტისგან დამოუკიდებელ გრამატიკაში (Feature-Based Context-Free Grammar – FBCFG). იგი შეიცავს წინადადების ფრაზულ შემადგენელთა გრამატიკული და ლექსიკური აღწერისათვის საჭირო წარმომშობ (Grammatical and Lexical Production) წესებს. ეს უკანასკნელი აყალიბებს კვანძებს, რომლებიც გაჯერებულია სხვადასხვა მახასიათებლით და მორფოსინტაქსური კლასის ჭდეებით (resp. POS-tags). ‘type’ მახასიათებელი წინადადების აღწერის ფრაზულ დონეზე აღნიშნავს ფრაზულ ჭდეებს (S, VP, NP). იგივე კვანძები ასევე მონიშნულია სინტაქსური ფუნქციის ინდიკატორებით (SF – Syntactic Function), როგორიცაა SUB (სუბიექტი), DDO (პირდაპირი ობიექტი) და HEAD (წინადადების ან ფრაზული შემადგენლის მეთაური სიტყვა – Head of clause/phrase).

შუალედური დონის MD (Modifier) კვანძები სინტაქსური ფუნქციის (SF) თვალსაზრისით მსაზღვრელს წარმოადგენს. VP კვანძი შეიცავს გრამატიკულ მახასიათებელს ser=1 და VAL="ND", რომლებიც რეალიზებულია "type"="TV02" ტერმინალური კვანძიდან (მეორე კლასტერი გარდამავალი ზმნით). VAL-მახასიათებლის (სინტაქსური ვალენტობა) მნიშვნელობა "ND" მიუთითებს მეთაური სიტყვის ბრუნვის მარკერზე (case='N[ominative]') მარცხენა NP-ში სინტაქსური ფუნქციით (SF='SUB') და მეთაური სიტყვის ბრუნვის მარკერზე (case='D[ative]') მარჯვენა NP-ში, როგორც SF='DO' (პირდაპირი დამატება – Direct_Object).

გრაფის/სინტაქსური ხის ზოგადი სქემა ფრაზული შემადგენლების კატეგორიებისთვის და წინადადებაში მათ მიერ გადმოცემული სინტაქსური მიმართებები აგებულია ფრაზის სტრუქტურის X-Bar თეორიის პრინციპებზე (Kornai, et al., 1990: 25ff). წინადადების პრედიკატ-არგუმენტის სტრუქტურა (სინტაქსური ფუნქციები) ეფუძნება სინტაქსური ვალენტობის გრამატიკულ კონცეფციას, „თავ-მახასიათებლის პრინციპის“ (head feature principle) ანალოგს, რომელიც მიღებულია თავზე ორიენტირებული ფრაზის სტრუქტურის გრამატიკაში (Head-Driven Phrase Structure Grammar) (Pollard, et al., 1994: 15ff).

ამავე წინადადების ზედაპირული სინტაქსური სეგმენტაციის შედეგი ნაჩვენებია მე-2 სურთზე, რომელშიც თითოეული ფრაზული შემადგენლის სრული პროექცია სამკუთხედის ქვეშ არის მოქცეული:

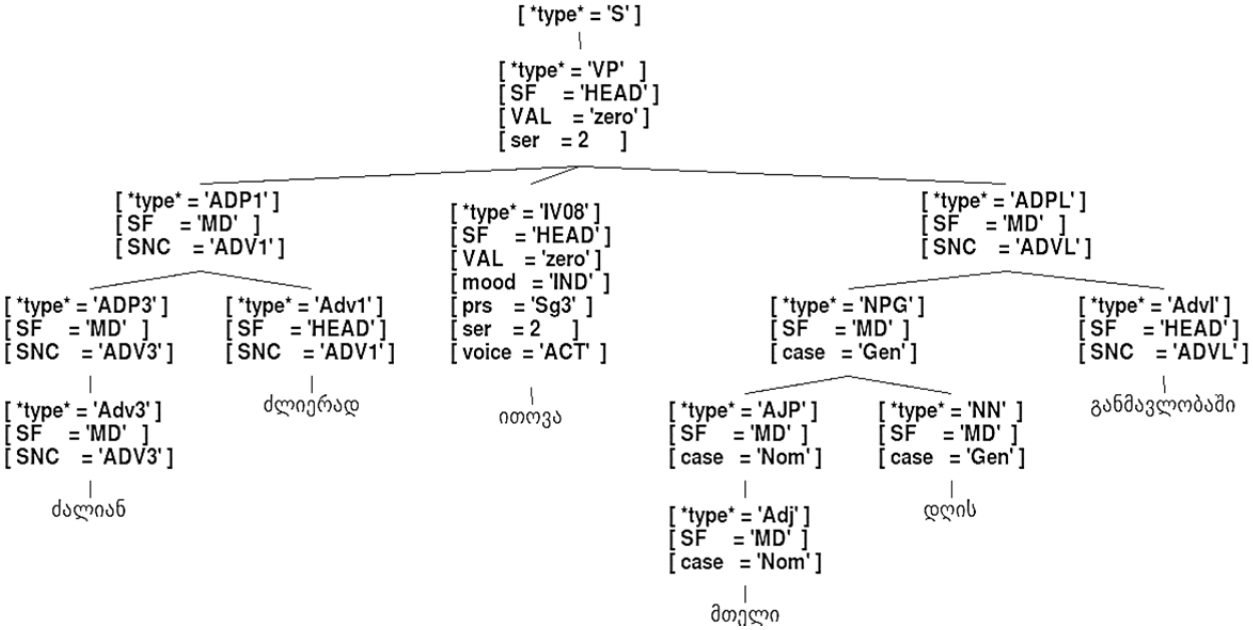


სურათი 2

ავტომატური პარსერის მიერ განსხვავებული სინტაქსური ხეების გენერირების უნარის საილუსტრაციოდ მე-3 სურათზე მოყვანილია ხე წინადადებისათვის:

„ძალიან ძლიერად ითოვა მთელი დღის განმავლობაში“

იგი პრედიკატის სახით მერვე კლასტერის ზმნას შეიცავს, რომელიც ნულოვან ვალენტობას (VAL="zero") გვიჩვენებს და P ფრაზულ შემადგენელს გამორიცხავს:



სურათი 3

პარსერის მიერ მისთვის გამოთვლილი სინტაქსური ხის სტრუქტურა წარმოდგენილია წინადადების საწყის კვანძთან (“type” = ‘S’) უშუალოდ დაკავშირებული შუალედური დონის VP კვანძის მეშვეობით. VP-ის ორივე მხარეს ხის სტრუქტურის იმავე დონეზე სეგმენტირებულია სხვადასხვა ტიპის ზმნიზედის ფრაზები (ADP1, ADPL), რომლებშიც მათი მეთაური სიტყვის სინტაგმატური კლასის მახასიათებლებია (SNC – Syntagmatic Classes) მითითებული.

4. სამომავლო გეგმები

სტატიაში წარმოდგენილია ქართული ენის სინტაქსური სეგმენტატორის/ზედაპირული ავტომატური ანალიზატორის შემუშავების კონცეფცია. შემოთავაზებული მიდგომის ინოვაციური ასპექტია მახასიათებლებზე ორიენტირებული კონტექსტისაგან დამოუკიდებელი გრამატიკის (FBCFG) გამოყენება ნაკლებად კონფიგურაციული ქართული ენის სინტაქსისთვის, თუმცა NLTK ტექნოლოგია არ იყო გამიზნული წინადადებაში სიტყვების თავისუფალი რიგის მქონე ენებისთვის.

ნაკლებად კონფიგურაციული ენებისგან განსხვავებით, ფიქსირებული სინტაქსური სტრუქტურის ენების წინადადების საწყისი (S) კვანძისათვის, როგორც წესი, მხოლოდ რამდენიმე პროდუქციული წესია საკმარისი ფრაზის სინტაქსურ შემადგენელთა გენერირებისათვის.

ქართული სინტაქსისათვის ჩვეული, წინადადებაში სიტყვების თავისუფალი რიგის გათვალისწინებით, სამვალენტიანი ზმნის ჩარჩოსათვის (NP1 NP2 NP3 VP) ფრაზული შემადგენლების ურთიერთის მიმართ გადაადგილების რაოდენობა თეორიულად შეიძლება 4! = 1 x 2 x 3 x 4 ტოლი იყოს. თუმცა წინადადებაში თითოეული მათგანის გამოჩენის ალბათობა განსხვავებული იქნება, რეალურად დასაშვები მიმდევრობების რაოდენობა საკმაოდ დიდია. შესაბამისად, ასეთი პროდუქციული წესების ხელით აგება მეტად შრომატევადია და მათი პრაქტიკული გამოთვლების პროცესში გამოყენებისას უზუსტობის გამორიცხვა არ შეიძლება.

არსებული პრობლემის გადასაჭრელად ვმუშაობთ პროგრამულ მოდულზე, რომელიც შეძლებს ავტომატურად განსაზღვროს ფრაზულ შემადგენელთა სწორი თანმიმდევრობა უშუალოდ შესავალი ტექსტიდან და იგი პროდუქციის წესების გრამატიკულ ნაწილს დაურთოს. მოდული ჩაშენდება, NLTK მოდულისგან დამოუკიდებლად, მონაცემთა ნაკადსადენში (pipe-line). ეს მნიშვნელოვნად გაზრდის ნაკლებად კონფიგურაციული ქართული ენის სინტაქსური ანალიზის ტექნოლოგიური პაკეტის მოქნილობას, რაც გააუმჯობესებს პარსერის ეფექტიანობას.

ბიბლიოგრაფია:

მელიქიშვილი, დამანა. *ქართული ზმნის უღლების სისტემა.* თბილისი, ლოგოსპრესი. 2001. P

Beesley, Keneth R. and Lauri Karttunen. *Finite State Morphology.* CSLI Publications. Leland Stanford Junior University. 2003.

Bird, Stiven, Loper, Ewan and Edward Klein. *Natural Language Processing with Python.* O'Reilly Media Inc. 2009.

Brants, Sabine and Silvia Hansen. >>>Developments in the TIGER Annotation Scheme and their Realization in the Corpus". *Proceedings of the Third Conference on Language Resources and Evaluation (LREC 2002)*, 1643–1649. Las Palmas de Gran Canaria, Spain. 2002.

Fraser, Alexander, Schmid, Helmut, Farkas, Richárd, Wang, Renjing and Hinrich Schütze. "Knowledge Sources for Constituent Parsing of German, a Morphologically Rich and Less-Configurational Language". *Computational Linguistics*, Volume 39, Issue 1, 57-85. MIT Press Cambridge, Ma, USA. 2013. http:// doi: 10.1162/COLI_a_00135

Kornai, Andras and Geoffrey K. Pullum. "The X-bar Theory of Phrase Structure". *Language, 66*(1), 24–50. 1990.

Pollard, Carl and Ivan A. Sag. Head-Driven Phrase-Structure Grammar. Chicago, IL: The University of Chicago Press. 1994.

Szántó, Zsolt and Richárd Farkas, R. (2014). Special Techniques for Constituent Parsing of Morphologically Rich Languages. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 135-144. Gothenburg, Sweden. 20114. http:// doi: 10.3115/v1/E14-1015

OLEG KAPANADZE, NUNU KAPANADZE

THE AUTOMATIC ANALYSIS OF THE GEORGIAN SENTENCE

Summary

Until recently, most basic research in Natural Language Technology (NLT) has been performed on "major" languages such as (predominantly) English but also German, Japanese, Chinese, French, and Spanish. At the same time, Low-Density Languages (LDL) compete to take advantage of modern digital technologies implemented in high-quality computing systems. As a result, the long-term viability of languages not specifically supported by NLT is at risk, which can lead to their digital extinction.

This paper presents an undertaking for developing computational applications involving Georgian to fill a gap with technologically well-equipped languages and to lower the current scarcity of language resources for Georgian text processing.

It is well known that Georgian is a language with rich inflectional morphology and with very few fixed structures on the sentence level. The languages of similar design are called Morphologically Rich and Less-Configurational (MR&LC). This paper concerns issues related to developing crucial NLT tools for the MR&LC Georgian language: We discuss the development of a Feature-Based Context-Free Grammar (FCFG) and a Featured Grammar parser for the Less-Resourced Georgian language.

Generative lexicalised parsing models, which are the mainstay for probabilistic parsing, do not perform as well when applied to languages with free word order or rich morphology. Based on the syntactic valency property of the verb and language-specific features such as productive morphology, we designed a prototype FCFG parser for automatic syntactic chunking/shallow parsing of the Georgian clause, which we present here.

As the initial step to the syntactic analysis, we reimplemented the rule-based Finite-State Morphological Transducer for Georgian text morphological analysis, lemmatization and POS tagging. To build an interface between the TIGER XML scheme and an input format for conceived syntactic chunker, we had to disambiguate manually and reformate the output of the Georgian morphoparser.

As a necessary step in the syntactic valency-driven Feature-Based Grammar parser implementation, we have studied the Georgian verb stock and clustered it according to syntactic valency features. Eight verb clusters with different valency distributions and syntactic frames are identified to date. For each cluster, we developed and started training a prototype Feature-Based Grammar version for Georgian.

As a syntactic parsing testbed, we have utilized a broadly recognized open-source NLTK library developed using the Python programming language.

In the meantime, we are developing a converter module capable of porting automatically the output of the morphoparser at hand into the acceptable format for the NLTK input engine. This would provide an option for piping the morphological transducer with the Feature-Based syntactic parser for linking them in an unsupervised shallow syntactic chunker/parser of the Georgian language text.