

ირინა ლობჯანიძე

ილიას სახელმწიფო უნივერსიტეტი

**მთარგმნელობითი ტექნოლოგიები, სამანქანო
თარგმანი და ლინგვისტური ტაგირების
პრინციპები**

IRINA LOBZHANIDZE

Ilia State University

**TRANSLATION TECHNOLOGIES, MACHINE TRANSLATION AND
PRINCIPLES OF LINGUISTIC TAGGING**

The European Masters in Translation network first published its framework for translator and translation competence, including the well-known EMT “Wheel of Competence”, in January 2009. Since that time, the Wheel of Competence has changed several times. In accordance with the changes implemented in 2017, the central part of the EMT framework is application of computational technologies to the translation studies. The needs of translation industry as a whole oblige young translator-interpreters to provide high-quality translation by means of computer technologies. The mentioned competence encompasses six components including Machine Translation (MT) competence. The translator-interpreter should use MT systems as a part of his/her working process.

This paper presents an overview of Machine Translation (MT) with regards to three main directions. The first one could be considered as a Rule-Based Machine Translation (RBMT) used in older types of technologies, the second one – as a Statistical Machine Translation (SMT) used in comparatively new types of technologies till 2016 and the third – as a Neural Machine Translation (NMT) used in modern type translation technologies since 2016. All these systems are somehow interconnected between each other and modern machine translation systems reveal a shift from statistical to deep learning practices. As a result of the above-mentioned shift, Google, Pangeanic, KantanMT, Omniscien Technologies and SDL Trados translation technologies use neural machine translation. But all these systems reveal difficulties in processing languages like Georgian.

The contents of the paper are as follows: 1. Introduction; 2. Theoretical and technological issues; 3. Components of machine translation with regards to Georgian language and 4. Conclusions.

შესავალი

ევროპის თარგმანთმცოდნეობის სამაგისტრო პროგრამათა ქსელმა პირველად თავისი კომპეტენციების ჩარჩო, მათ შორის EMT-ს „კომპეტენციათა ბორბალი“, 2009 წლის იანვარში გამოსცა. მას შემდეგ კომპეტენციათა ჩარჩომ შემადგენელი ნაწილების შინაარსის თვალსაზრისით ბევრჯერ შეიცვალა. 2017 წელს განხორციელებული ბოლო ცვლილებების შესაბამისად,

EMT-ს ჩარჩოს ცენტრალურ ნაწილს კომპიუტერული ტექნოლოგიების ცოდნა წარმოადგენს. ბაზრის დღევანდელი მოთხოვნებიდან გამომდინარე, იგი განიხილება, როგორც ერთ-ერთი ყველაზე მნიშვნელოვანი. მთარგმნელ-თარჯიმნის პროფესიონალად ჩამოყალიბების კუთხით. აღნიშნული კომპეტენცია ექვს კომპონენტს მოიცავს, რომელთაგან განსაკუთრებით საყურადღებო სამანქანო თარგმანია (MT). მთარგმნელ-თარჯიმანს მუშაობის პროცესში უნდა შეეძლოს როგორც სამანქანო თარგმანის მთარგმნელობით პროცესზე ზეგავლენის საწყისების ცოდნა, ისე მისი გამოყენება თარგმანის შესრულების დროს.

მთარგმნელობითი ტექნოლოგიების მზარდი გამოყენება მათი განვითარებისა და სრულყოფის წინაპირობაა, რაც ვლინდება როგორც ახალი ტექნოლოგიების შექმნაში, ისე ძველი ტექნოლოგიების დახვეწაში. დღესდღეობით მთარგმნელობითი ტექნოლოგიის განუყრელ ნაწილს სამანქანო თარგმანი წარმოადგენს, რომელიც ორ მიდგომად იყოფა. პირველი მიდგომა გამოიყენება შედარებით ძველი ტიპის ტექნოლოგიებში და გულისხმობს კომპიუტერული საშუალებებისა და ადამიანის ენობრივი ცოდნის ინტეგრირებას, ხოლო მეორე - კომპიუტერული საშუალებების გამოყენებით მონაცემების სტატისტიკური დამუშავების საფუძველზე კონცენტრირებულია მანქანური სწავლის შესაძლებლობებზე. მეორე მიდგომის ნაწილს - ნეირონული სამანქანო თარგმანი წარმოადგენს.

სამანქანო თარგმანის ნეირონული მიდგომის განვითარება 2016 წელს დაიწყო, როდესაც იუ. ვუმ, მ. შუსტერმა და სხვ. წარმოადგინეს ტექნიკური ანგარიში გუგლის ნეირონული თარგმანის შესახებ (Wu et al., 2016:1-23). აღნიშნული ანგარიშის მიხედვით, ნეირონული თარგმანი (Neural Machine Translation, NMT) წარმოადგენდა ავტომატური თარგმანის ახალ მიდგომას, რომელიც უზრუნველყოფდა იმ ცდომილებების აღმოფხვრას, რომლებიც სდევდა შესიტყვებაზე ორიენტირებულ მთარგმნელო-

ბით სისტემებს და გულისხმობდა მანქანური დრმა სწავლების მოდელს, რომელშიც გათვალისწინებული იყო 8-დონიანი შიფრატორი და 8-დონიანი დეშიფრატორი. პარალელიზმის უზრუნველყოფისა და სწავლის დროის შემცირების მიზნით, აღნიშნულ მოდელში დეშიფრატორის ქვედა დონე უკავშირდებოდა შიფრატორის ზედა დონეს. თარგმანის სიჩქარის გასაუმჯობესებლად კი სისტემა მიმართავდა დაბალი-სისწორის არითმეტიკულს კომპიუტერული ინტერფერენციის დროს. აღნიშნული მიდგომით დაიწყო ახალი ერა სამანქანო თარგმანის განვითარების თვალსაზრისით, რაც შემდგომში რეალიზდა Google-ის, Pangeanic-ის, KantanMT-ის, Omnicien Technologies-ისა და SDL-ის ტექნოლოგიებში. ბუნებრივია, იბადება კითხვა, მუშაობს თუ არა აღნიშნული მიდგომა ქართული ენის შემთხვევაში და გავგაჩნია თუ არა რესურსი მსგავსი მოდელის გამოსაყენებლად.

სტატიის სტრუქტურა შემდეგია: 1. შესავალი; 2. თეორიული და ტექნოლოგიური საკითხების მიმოხილვა; 3. სამანქანო თარგმანის მახასიათებლები; 4. დასკვნები.

თეორიული და ტექნოლოგიური საკითხები

დღესდღეობით მთარგმნელობითი ტექნოლოგიის მთავარ ამოცანას წარმოადგენს ერთიანი მიდგომის შემუშავება, რომელიც ეფუძნება მედიატორი ენის არსებობას და არ ითხოვს ორ ენას შორის რეალური ენობრივი კომპეტენციის არსებობას. ამ გზით შემუშავებული სისტემა: ა) უნდა თარგმნიდეს მაშინაც კი, როცა ორ ენას შორის წინასწარი ენობრივი კავშირი არ არსებობს და, ბ) უნდა გახდეს გლობალური მთარგმნელობითი სისტემის ნაწილი. თარგმანი კი უნდა ხორციელდებოდეს როგორც წესებზე ორიენტირებული, ისე სტატისტიკურიც და ნეირული მეთოდების გათვალისწინებით.

თარგმნის პროცესი გულისხმობს ორი მხარის არსებობას: ორიგინალი და სამიზნე ენებისა და თუ რამდენად ეფექ-

ტურია ესა თუ ის თარგმანი, დამოკიდებულია აღნიშნული წყვილის წინასწარი დამუშავების შესაძლებლობაზე. მსოფლიოში არსებული სხვადასხვა მიდგომა სამანქანო თარგმანისადმი აჩვენებს როგორც დადებით, ისე უარყოფით შედეგებს. როგორც აღინიშნა, სამანქანო თარგმანის თვალსაზრისით კვლევათა ძირითადი მიმართულება სამ ნაწილად იყოფა და მოიცავს შემდეგს:

1. წესებზე ორიენტირებულ სამანქანო თარგმანს, რომელიც ეფუძნება ლინგვისტურ ინფორმაციას ორიგინალი და სამიზნე ენების შესახებ. ლინგვისტური ინფორმაცია კი, თავის მხრივ, მოიცავს მონოლინგვურ და ბილინგვურ ლექსიკონებსა და კონკრეტული ენის მორფოლოგიური, სინტაქსური და სემანტიკური სტრუქტურების აღწერას (Nirenburg, 1989:5-24). შესაბამისად, სისტემაც თარგმნის წინადადებებს სამიზნე ენაზე, გრამატიკულ მახასიათებლებზე დაყრდნობით. მსგავსი ტიპის სისტემებს განეკუთვნება Systran (1968-2016), GramTrans (2014) და სხვ.

2. შესიტყვებაზე ორიენტირებულ სტატისტიკურ სამანქანო თარგმანს, რომელიც გულისხმობს თარგმანის გენერირებას სტატისტიკურ მოდელზე დაყრდნობით. აღნიშნულ მიდგომას სჭირდება ბილინგვური კორპუსების ანალიზის შესაძლებლობა (Marcu, 2002:133-139). თარგმანის ნეირული მეთოდის შემოღებამდე აღნიშნული მიდგომა საკმაოდ წარმატებულ მეთოდებას უწევდა წესებზე ორიენტირებულ სამანქანო თარგმანს. მსგავსი ტიპის სისტემებს განეკუთვნებოდა Google Translate (2006-2016), Microsoft Translator (2007-2016) და სხვ.

3. ნეირულ სამანქანო თარგმანს, რომელიც იყენებს ხელოვნურ ნეირულ ქსელს სიტყვების თანმიმდევრობის აღბათობის გამოთვლის მიზნით. აღნიშნული მიდგომა სტატისტიკური სამანქანო თარგმანისაგან განსხვავდება იმით, რომ არ განიხილავს ენობრივი ან მთარგმნელობითი მოდელის არსებობას, არამედ ერთიან მოდელს, რომელიც კონცენტრირდება

ერთ სიტყვაზე დროის ერთ მონაკვეთში. აღნიშნული მიდგომა ღრმა სწავლებასაც იყენებს (Zhaopeng და სხვ., 2016). აღსანიშნავია, რომ წესებსა და შესიტყვებაზე ორიენტირებული სტატიკური სამანქანო თარგმანი ზემოხსენებულ პროგრამულ უზრუნველყოფაში 2016 წლიდან ნეირულმა მანქანურმა თარგმანმა ჩაანაცვლა.

ყველა აღნიშნულ მიდგომას აქვს თავისი დადებითი და უარყოფითი მხარე და იმისათვის, რომ გავცეთ პასუხი - შესაძლებელია თუ არა მსგავსი მიდგომების გამოყენება ქართული ენის შემთხვევაში, მნიშვნელოვანია აღნიშნული მახასიათებლების გადახედვა ქართული ენის საჭიროებიდან გამომდინარე.

სამანქანო თარგმანის მახასიათებლები

წესებზე ორიენტირებული სამანქანო თარგმანის მთავარ მახასიათებლებს შემდეგი კომპონენტები უნდა წარმოადგენდეს: ა) ენის გრამატიკაზე ორიენტირებული ბლოკები, როგორებიცაა მორფოლოგიური ანალიზატორი და გენერატორი, სინტაქსური პარსერი და, ბ) ორიგინალი და სამიზნე ენის მონოლინგვური ლექსიკონები და ორიგინალი და სამიზნე ენის ბილინგვური ლექსიკონი. ამოცანიდან გამომდინარე, ორივე ბლოკი უნდა მოიცავდეს როგორც ორიგინალი, ისე სამიზნე ენებისთვის შესაფერის ერთეულებს, წყვილის საჭიროების უზრუნველყოფის მიზნით. აღნიშნული კი გულისხმობს, რომ არაა საკმარისი მხოლოდ ქართული ენის მორფოსინტაქსური კომპიუტერული აღწერა, არამედ საჭიროა სამიზნე ენის შესაფერისი კომპიუტერული აღწერის მოფიქრება შესაბამისი მაკავშირებელი წესების უზრუნველყოფით. მაკავშირებელი წესი კი გულისხმობს, რომ უნდა არსებობდეს მკაფიოდ განსაზღვრული გრამატიკული ფუნქციებისა და სინტაქსური თავების დაკავშირების შესაძლებლობა ორიგინალ და სამიზნე ენაში. აღნიშნული მიდგომის გამოყენებისას მთავარ სირთულეს ქართული ენის შემთხვევაში წარმოადგენს შემდეგი:

1. ლექსიკონების უკმარისობა კონკრეტული ენობრივი წყვილებისათვის და არქონა ენობრივი წყვილების უმეტესობისათვის. მაგალითად, შედარებით კარგად მანქანის მიერ წაკითხვად ფორმატში წარმოდგენილია ინგლისურ-ქართული (Translate, 2010), (მარგალიტაძე, 1995-2020) და ქართული-ინგლისური (Rayfield, 2006) ენობრივი წყვილები, თუმცა სხვა წყვილების შემთხვევაში შემდეგი ვარიანტები შესაძლებელია არსებობდეს ნაბეჭდი, მაგრამ არ არსებობდეს ელექტრონული ან არ არსებობდეს არც ერთი და არც მეორე;

2. გრამატიკული ომონიმების შემთხვევები, რომელებიც ვლინდება სიტყვების მორფოსინტაქსური ტაგირების დროს, მაგ:

და:და+Conj+Coord

და:და+N+Com+Anim+Sg+Nom

სისტემას უჭირს სათარგმნელად სწორი ერთეულების შერჩევა.

3. იდიომური გამოთქმების გამოვლენის შემთხვევები, რომლის თარგმნა ნებისმიერი ენობრივი წყვილის შემთხვევაში რთულდება, ზღრ. ყბად იღებს - ჭორაობის ვინმეზე.

შესიტყვებაზე ორიენტირებული სტატისტიკური სამანქანო თარგმანის მთავარ მახასიათებლებს უნდა წარმოადგენდეს მონოლინგვური მონაცემებისა და პარალელური კორპუსების ეფექტური გამოყენება, რომელიც ვლინდება არა ცალკეული სიტყვისა, არამედ შესიტყვებებისა და წინადადებაში სიტყვათა თანმიმდევრობის სტატისტიკური დამუშავების საფუძველზე სამიზნე ენაზე ორიგინალი ენიდან წინადადების თარგმნის დროს. აღნიშნულ მიდგომას რიგი სისუსტეც ახასიათებს და ყველაზე მნიშვნელოვანი ის, რომ ასეთი სისტემები ყველაზე ცუდად მუშაობს იმ შემთხვევაში, როცა ორიგინალ და სამიზნე ენების სინტაქსური სტრუქტურები არ აძლევს ორიგინალი ენის სიტყვათა თანმიმდევრობის დაცვის შესაძლებლობას. აღნიშნული მიდგომის გამოყენებისას მთავარ სირთულეს ქართული ენის შემთხვევაში წარმოადგენდეს შემდეგი:

1. დარგში არსებული საერთაშორისო სტანდარტების გამოყენებით პარალელური კორპუსების შექმნა თითქმის არ მიმდინარეობს. ერთ-ერთ მცდელობას ქართულ-ინგლისური ტერმინოლოგიის პარალელური კორპუსი წარმოადგენს, რომელშიც ამ ეტაპზე 89,286 წყვილია (კერეჭაშვილი, 2014), და შეიქმნა SDL Trados-ის მძლავრი მთარგმნელობითი ბაზა, რომელიც საფუძვლად დაედო საქართველოს საკანონმდებლო მაცნის ტერმინოლოგიურ ლექსიკონს (საქართველოს საკანონმდებლო მაცნე: ტერმინოლოგიური ლექსიკონი, 2011-2020);

2. აღნიშნული ტიპის სისტემები ძალიან ცუდად მუშაობს იმ შემთხვევაში, თუ სამიზნე ენაში სიტყვათა თანმიმდევრობა მკვეთრად განსხვავდება ორიგინალ ენაში არსებული სიტყვათა თანმიმდევრობისგან. ქართული ენა განეკუთვნება ერგატიული სინტაქსის მქონე მორფოლოგიურად მდიდარ ენებს, რაც წარმოადგენს სიტყვების შედარებით თავისუფალი თანმიმდევრობის წინაპირობას და განსაზღვრულ დონეზე მაღალი დონის სტატისტიკური თარგმანის ხელისშემშლელი ფაქტორია.

ზემოხსენებული პრობლემები აბრკოლებს შესიტყვებაზე ორიენტირებული სტატისტიკური სამანქანო თარგმანის წარმატებული რეალიზების შესაძლებლობას არა მხოლოდ ქართულ-ინგლისური და ინგლისურ-ქართული ენობრივი წყვილის, არამედ სხვა, ნაკლებად გავრცელებული, ენობრივი წყვილების შემთხვევაშიც.

ნეირული სამანქანო თარგმანი ამ ეტაპზე ყველაზე გავრცელებული ტექნოლოგიაა მსოფლიო მაშტაბით თარგმანის უზრუნველყოფის მიზნით. მსგავსი ტიპის სისტემები განსხვავდება სტატისტიკური სამანქანო თარგმანისაგან იმით, რომ ნეირული მთარგმნელობითი მოდელის ყველა ნაწილი მზადდება ერთიანად და ეფუძნება სიტყვებისა და შიდა პოზიციებისთვის ვექტორების გამოყენებას. აღნიშნული მოდელის საფუძველზე კი მუშავდება ისეთი მოდელი, რომელიც ახერხებს თანმიმდევრულად თითო-თითო სიტყვის ამოცნობას. ამ მოდელს იყენებს Google Translate (2006-2016), Microsoft Translator (2007-2016)

და სხვ.. მის ხარისხს ქართული ენის შემთხვევაში ამ ეტაპზე სჭირდება გაუმჯობესება, რაც გამოწვეულია შემდეგი სირთულეებით:

1. ქართული ენა განეკუთვნება ე.წ. მცირე-რესურსების მქონე ენებს, რასაც მისი გამოყენების არეალი და მასზე მოლაპარაკე ხალხის რაოდენობა უდევს საფუძვლად. კომპიუტერული თვალსაზრისით ეს ყველაფერი ვლინდება იმაში, რომ არ არსებობს მონაცემების საკმარისი რაოდენობა ღრმა სწავლების უზრუნველყოფის მიზნით;

2. იშვიათი სიხშირის მქონე სიტყვებზე ნეირულ თარგმანზე ორიენტირებული სისტემები უფრო ცუდად მუშაობს და განსაკუთრებით ცვალებადი მორფოსინტაქსური კონსტრუქციების შემთხვევაში. ქართული ენის შემთხვევაში ეს სირთულე ყველაზე მეტად ზმნური ფორმების თარგმანში ვლინდება, რადგანაც ქართული ზმნის პარადიგმა განეკუთვნება მაღალი-ცვალებადობის კატეგორიათა რიგს;

3. ნეირული თარგმანის სისტემები ნაკლებ სიზუსტეს აჩვენებენ რთული თანწყობილი და ქვეწყობილი წინადადებების თარგმნის შემთხვევაში. ქართული ენის სინტაქსი განაპირობებს გრძელი წინადადებების შექმნის შესაძლებლობას და წინა სირთულესთან ერთად ქმნის წინაპირობას იმისა, რომ ნეირული თარგმანის მთარგმნელობით სისტემებში ქართული ენიდან და ქართულ ენაზე თარგმნის ხარისხის გაუმჯობესებას შეიძლება ბევრი დრო დასჭირდეს.

დასკვნები

ზემოხსენებული სისტემების განხილვის საფუძველზე ვლინდება, რომ ქართულ ენაზე არსებული რესურსი, რომელიც უზრუნველყოფს და შექმნის წინაპირობას ქართული ენიდან და ქართულ ენაზე სამანქანო თარგმნის შესაძლებლობისა, შეზღუდულია და საჭიროებს ენის სრულყოფილი მოდელის შექმნას არა მხოლოდ მორფოსინტაქსური, არამედ სინტაქსური თვალსაზრისითაც.

ლიტერატურა

- Google Translate [კომპიუტერული უზრუნველყოფა] 2006-2016, <https://translate.google.com/>.
- GramTrans [კომპიუტერული უზრუნველყოფა] 2014, <https://gramtrans.com/>.
- KantanMT [კომპიუტერული უზრუნველყოფა] 2017, KantanMT Users Can Now Customise and Deploy Neural Machine Translation Engines. მოპოვებული 2019 წლის 15 September, Slator: <https://slator.com/press-releases/kantanmt-users-can-now-customise-and-deploy-neural-machine-translation-engines/>.
- Marcu, D., Wong, W. 2002,** A phrase-based, joint probability model for statistical machine translation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Philadelphia: Association for Computational Linguistics, 133-139.
- Microsoft Translator [კომპიუტერული უზრუნველყოფა] 2007-2016, <https://www.microsoft.com/en-us/translator/>.
- Nirenburg, S. 1989,** Knowledge-Based Machine Translation. *Machine Translation* 4. Springer, 5-24.
- Rayfield, D. 2006,** A Comprehensive Georgian-English Dictionary. London: Garnett.
- SDL Trados: MultTrans [კომპიუტერული უზრუნველყოფა] 2019 <https://www.sdl.com/software-and-services/translation-software/sdl-trados-studio/>.
- Systran [კომპიუტერული უზრუნველყოფა] 1968-2016 <https://www.systran-soft.com/>.
- Omniscien Technologies [კომპიუტერული უზრუნველყოფა] 2017, Omniscien Technologies Announces Release of Language Studio™ with Next-Generation NMT Technology. მოპოვებული 2019 წლის 15 September, Omniscien Technologies: <https://omniscien.com/omniscien-technologies-announces-release-language-studio-next-generation-neural-machine-translation-technology/>.
- Translate [კომპიუტერული უზრუნველყოფა] 2010, <https://translate.ge/>.
- Wu, Y., Schuster, M., Chen, Z. et al. 2016,** Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. Technical Report.
- Zhaopeng, T., Zhengdong, L., Yang, L., Xiaohua, L., Hang, L. 2016,** Modeling Coverage for Neural Machine Translation. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Berlin: Association for Computational Linguistics, 76-85.

კერეჭაშვილი, გ., მარგილატაძე, თ. 2014, პარალელური კორპუსი. <https://corp.dict.ge/>.

მარგალიტაძე, თ. 1995-2020, დიდი ინგლისურ-ქართული ონლაინლექსიკონი. <https://dictionary.ge/>.

ტერმინოლოგიური ლექსიკონი [კომპიუტერული უზრუნველყოფა] 2011, საქართველოს საკანონმდებლო მაცნე: ტერმინოლოგიური ლექსიკონი: <http://www.matsnedictionary.gov.ge/>.