Towards a Georgian Controlled Language in Crisis Management

Nino Amiridze[1], Anna Chutkerashvili[1,2], Besik Dundua[1,3], and Irina Temnikova[4]

[1] Ivane Javakhishvili Tbilisi State University, Georgia
[2] Georgian Technical University, Georgia
[3] International Black Sea University, Georgia
[4] Sofia University, Bulgaria

Controlled natural languages (CNLs) are engineered languages that are based on natural language, but have their vocabulary, syntax, and/or semantics restricted [9]. The motivation is to have a language that, on one hand, looks as natural as possible and, on the other hand, is simple and unambiguous.

The application area of CNLs is quite broad. There are general-purpose CNLs (e.g., Attempto Common English [6]), whose vocabulary can be adapted to specific areas, and domain-specific ones, which reflect also the domain syntax. Due to space limitations, we just list some interesting practical applications of the domain-specific CNLs: writing technical documentation [3], mathematical authoring [5], crisis management [12, 13, 2], legal texts [4], interchanging business rules among organizations [11], and editing ontologies in OWL [10].

While English dominates the landscape of CNLs ([9] analyzes 100 English CNLs), controlled (sub)languages have been developed for many other languages (e.g., Bulgarian, French, Spanish, Chinese, Russian, German, Greek, Spanish, and Japanese). In this talk we present our project on Georgian controlled language in crisis management, which aims at establishing clear (and unambiguous) communication between different management specialists and the general population.

Georgian is a language from the Kartvelian family of languages, spoken mainly in Georgia, where it is an official language. Georgian has its own alphabet and long literary tradition, which goes back to 5th century AD. From the computational point of view, despite some recent and ongoing attempts, Georgian is still pretty under-resourced (e.g., [1, 8]).

One of the application areas of controlled natural languages is crisis management. During the last decades several big disasters happened in Georgia which had substantial impact on people's lives. In 2017, 100 hectares of forests were burned in Borjomi Gorge. In 2015, the Vere river was flooded, which caused severe damage on the city's infrastructure and resulted in at least 19 fatal casualties. In 1991, an earthquake occurred in the province of Racha, which killed more than 270 people and left more than 100,000 inhabitants homeless. Such big disasters require clear communication with large number of people in a stressful situation. Unfortunately, until now, the country does not have a central emergency alarm system, but since 2018 there is a discussion in the government to implement it.

Our project includes: collection of a small corpus of emergency instructions in Georgian, linguistic analysis of the ambiguities in the texts, development of CNL grammar rules (some

of which map to first-order logic) and define the allowed syntactic structures, word order, and lexicon; and evaluation of the resulting CNL.

In this talk we present some of the work done. We present some ambiguity cases and propose alternative clearer solutions. We analyze which of the existing ambiguity cases for other languages appear also in Georgian, and which are typical only for it. Being a complex language not closely related to any well-resourced language, developing a CNL for Georgian is an additional challenge. In fact, Georgian has one of the most complex morphologies: e.g., the verb form is very complex – it has agglutinative and inflection morphemes, and many categories (person, number, tense, mood, voice, etc.) which complicates parsing.

Some of the most frequent types of complexities and ambiguities in emergency texts are: writing too long sentences, not following lexical consistency, using syntactically or lexically ambiguous modifiers [7], using more archaic or complex forms or highly specialized terms, etc.

The examples below show some of the ambiguity cases in Georgian emergency instructions, that we identified in real texts, along with our proposed solutions.

(1) Syntactic ambiguity, which may prevent from correctly executing the instruction:

daķeţet   ķar+eb-eb-i,   panjr-eb-i   da napral-eb-i
you.PL.close.them door+PL-PL-NOM window-PL-NOM and opening-PL-NOM
amokolet   svel-i   pirsaxoc-it.
you.PL.block.them wet-NOM towel-INST

   a. 'Close the doors and windows and block the openings with wet cloths.'

   b. 'Close the doors and block the windows and openings with wet cloths.'

The form *ķar+eb-eb-i* in (1) is a non-standard, colloquial form referring to the plural referent 'doors'. The instruction manual from which we cited the example avoided using the standard Georgian plural form *ķar-eb-i* (2a), as the latter is ambiguous – in colloquial Georgian it refers to the singular referent 'door' (2b).

(2) Lexical ambiguity:

   a. Standard Georgian
      ķar-eb-i
      door-PL-NOM
      'doors'

   b. Colloquial Georgian
      ķar+eb-i
      door+PL-NOM
      'a door'

To avoid both the lexical ambiguity (which the use of *karebi* might cause (cf. (2a) vs. (2b))) and the use of the non-standard *ķar+eb-eb-i* (as in (1)), we would prefer to employ the quantifier *qvela* 'all' (3):

(3)   Standard Georgian

    qvela ḳar-i
    all   door-NOM

    'all doors'

Our talk will present more intriguing cases, taken from real documents, along with the CNL pattern rules to avoid such ambiguities. Our work is an useful example for many other under-resourced languages.

References

1. Lasha Abzianidze. An HPSG-based formal grammar of a core fragment of georgian implemented in TRALE. Master's thesis, Charles University in Prague, 2011.
2. AeroSpace and Defence Industries Association of Europe. *ASD Simplified Technical English: Specification ASD-STE100.* ASD AeroSpace and Defence, 2005.
3. Arendse Bernth. Easyenglish: A tool for improving document quality. In *5th Applied Natural Language Processing Conference, ANLP 1997, Marriott Hotel, Washington, USA, March 31 - April 3, 1997*, pages 159–165. ACL, 1997.
4. John J. Camilleri, Gordon J. Pace, and Michael Rosner. Controlled natural language in a game for legal assistance. In Michael Rosner and Norbert E. Fuchs, editors, *Controlled Natural Language - Second International Workshop, CNL 2010, Marettimo Island, Italy, September 13-15, 2010. Revised Papers*, volume 7175 of *Lecture Notes in Computer Science*, pages 137–153. Springer, 2010.
5. Marcos Cramer, Bernhard Fisseni, Peter Koepke, Daniel Kühlwein, Bernhard Schröder, and Jip Veldman. The Naproche Project Controlled Natural Language Proof Checking of Mathematical Texts. In *Controlled Natural Language, Workshop on Controlled Natural Language, CNL 2009, Marettimo Island, Italy, June 8-10, 2009.*, pages 170–186, 2009.
6. Norbert E Fuchs and Rolf Schwitter. Attempto Controlled English (ACE). *arXiv preprint cmp-lg/9603003*, 1996.
7. Trevor A Harley. *The psychology of language: From data to theory.* Psychology press, 2013.
8. Oleg Kapanadze and Alla Mishchenko. A multilingual GRUG treebank for underresourced languages. In Alexander F. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing - 14th International Conference, CICLing 2013, Samos, Greece, March 24-30, 2013, Proceedings, Part I*, volume 7816 of *Lecture Notes in Computer Science*, pages 50–59. Springer, 2013.
9. Tobias Kuhn. A survey and classification of controlled natural languages. *Computational Linguistics*, 40(1):121–170, 2014.
10. Richard Power. OWL Simplified English: a finite-state language for ontology editing. In *International Workshop on Controlled Natural Language*, pages 44–60. Springer, 2012.
11. Ronald G Ross. *Principles of the business rule approach.* Addison-Wesley Professional, 2003.
12. Irina Temnikova and Ruslana Margova. Towards a controlled language in crisis management: The case of bulgarian. In *Proceedings of the International Symposium on Data and Sense Mining, Machine Translation and Controlled Languages (ISMTCL), Besancon, France*, 2009.
13. Irina P Temnikova, Constantin Orasan, and Ruslan Mitkov. Clcm-a linguistic resource for effective simplification of instructions in the crisis management domain and its evaluations. In *LREC*, pages 3007–3014, 2012.