The Translation Model Based on Sentential Primitives

Liana Lortkipanidze[1], Nino Amirezashvili[2], George Chikoidze[2], Nino Javashvili[2]

[1]Ivane Javakhishvili Tbilisi State University, Georgia

[2]Archil Eliashvili Institute of Control Systems of Georgian Technical University, Georgia

Abstract.

The main problem of language modelling is the meaning of the utterance, which is the ultimate goal of one of the directions of language model functioning (analysis) and the starting point for its opposite direction (synthesis). The paper suggests one of the options of the problem solution. In particular, dividing sentences into the set of (quasi)-synonymous “sentential primitives" and on the basis of the "primitives", building such a structure, which reflects the appropriate content of this set quite transparently. Determination of the estimated set of primitives should be based on the search of the constituents according to their characteristics, particularly, to their syntactic properties. Verbs and their dominant role in sentences are considered according to "multilevel" syntax. By this theory, a primitive is a central component (core) of the layered structure, which is "surrounded" by the periphery i.e. traditional adverbs. The central component itself involves the verb and its actants. At the same time, internal relations of primitives, i.e. dependencies of actants on the predicates and peripheries on central structures are defined by the semantic roles. Such approach to language content conveys the meaning of the utterance through the sentential primitives where the primitives are the language units. In the frame of the presentation, we will offer a translation model, which is simplified to the level of primitives. For testing the model, Georgian-English parallel corpus will be used.

Keywords: language; modelling; meaning; semantics; utterance.

1. Introduction

If we take into account the information search process on the one hand and the variety of languages on the other, it will be easy to see that translation is rather useful, but at the same time quite complicated and diverse process that needs to be simplified. Accordingly, automation of translation has become one of the purposes of using computer intellectually. However, the systems of machine translation came across complications because of the diversity of languages. All these complications were caused by the variety of means of language, through which the same meaning is expressed. The difference is quite significant on the surface levels of a language (phonology, morphology) and it decreases on syntax level, however, it keeps peculiarities of each language.

The translation system that is based on the syntax structures of two different languages requires creation of a complex interactive component, which is different for any pair of languages. We can say that complications do not exclude the creation of translation system, which, as a specific syntactic structure of statements, will be include superficial statements. Based on such a structure, on the one hand, and by means of lexical matches, on the other, the system will be able to determine the correspondence of the same statement within the syntax structure of the other language. Finally, it will be able to create an expression of information using this "target language," in other words; will be able to synthesize the language of the source content.

It should be noted that the approach on which our work is focused does not exclude analytical or synthetic components. It means simplifying and generalizing the middle link only, that includes the essence of translation, or specifies correspondence between the result of analysis and the initial point of synthesis. More specifically, both sides of the correspondence must be represented using structures that are deeper and therefore more connected with each other. Of course, we can consider these types of structures as semantic structures more related to the meaning of utterance rather than syntactic structures.

Presenting an utterance on semantic level requires an additional extension for both opposing processes. Namely, for analysis it requires adding the last stage and for synthesis - the initial one. As a compensation for complications, we can expect semantic structures of different languages becoming close. Consequently, we expect simplification and specification in establishing correspondences between them, and so, we expect high quality of translation.

Whether the task solution will be successful or not, first of all, depends on the choice of semantic structure and its representation. In the paper, we present formation of the semantic structure and one of its representation, which is very important in our opinion.

Generally, the structure of any system is considered as the unity of its parts, that are related to each other and these relations unite them as one whole. This approach distinguishes the main task - it is definition of units and relations. It is reasonable to begin with the units as they have internal structure that also conditions the "external" relations. All this makes them unite semantic structure. In other words, the basic basis of the approach is the assumption that any language utterance is compensation of separate facts, situations that can be expressed by the most simplified language means, so called “sentential primitives”.

2. Methodology

We consider the content of the utterance as a role structure. Its constituent units are sentential primitives i.e. maximally simplified sentences. Each of the primitives must clearly reflect any particular situation – being it a process or a state. The role structure obtained by composition of sentential primitives should be equal to the content of the whole utterance.

The process of presentation of the utterance content will be based on the following scheme:

Formation of a primitive group → Selection of a dominant primitive → Establishment of the role relationships → Generation of (quasi-)synonymous sentences.

For the implementation of the approach, the selection of the primitive primary form is a very important aspect. The means of the selection from the variety of lexical units should be included in the presumable initial semantic- syntactic structure of the sentential primitives. The latter can be engaged from the "synonymous series" (Apresjan, 2003), especially if we consider that the lexical quasi-synonymy must have one of the crucial roles in the further process of forming. “When we talk about different ways of expressing the same content or about the semantic equivalence of outwardly different expressions we assume the existence of some semantic language or conceptual language which is inaccessible to direct observation. The production of sensible sentences can be regarded as a process of translating from the semantic language into a natural language, and understanding sentences can be regarded as a process of translating from natural language into the semantic language.” (Apresjan, 1973), (Apresjan, 1992)

According to the "Lexical semantics" approach, the members united in the synonymous series should have a common semantic crossing and we should consider the member as a dominant, whose semantics is the closest with this crossing.

The separation of the "Dominant primitive" emphasizes it as a “core” (Van Valin & LaPolla 1997), the central "fact” of the group. The rest of the group members are located around it. They expand and define the general content and, thus, are the periphery of this "dominant" "core".

Lexical information (synonymous series or lexical functions) should be reflected in the system dictionaries, as on the one hand, they provide potential of generation, on the other hand, the possibility to select a version of the expression for the concrete utterance, which is closest to the original.

Such a generator of synonymous statements based on primitives and role structures should ensure the functioning of both opposite directions of the language model (analysis, synthesis). While forming the final result the system controls both opposite processes: the analysis (standardization of primitives) and the synthesis (generation of quasi-synonyms).

For building the primitive based content structure and using it as a translate system it is necessary to create special conditions for testing the system and try to improve it.

One of the possible ways of testing must be the language parallel corpora. i.e. couples of corpora, one of which corresponds to the translation of the other. For example, one of them may be a combination of Georgian texts and the second – their English equivalence.

If we imagine that at this stage some systems, which split an utterance into primitives and construct corresponding semantic structures, have already been formed, we will get content structures of both sides. By comparison and alignment, it will be possible to determine what is needed to make sure that its result is close to the results in the parallel corpora.

“There exists a finite set of indecomposable meanings — semantic primitives. Semantic primitives have an elementary syntax whereby they combine to form 'simple propositions” (Goddard & Wierzbicka 1994) - clauses. “Semantic primitives and their elementary syntax exist as a minimal subset of ordinary natural language.” (Goddard & Wierzbicka 1994)

Both directions of the process of utterance can be considered as a combination of its mental and linguistic compositions. The generation of the utterance must begin with the mental analysis of the initial idea, which is based on the "starting point” that produces certain language expressions. On the contrary, the perception of a statement must begin with an analysis of the language, which consists in determining independent language content for the mental synthesis.

For some simplification of the task, we will consider the text as the beginning of the analysis where word forms are obviously separated from each other by places and punctuation marks. Let us submit the sentence transformed to the sequence of characteristics of each form of a word.

The top layer representing a primitive of each component of the offer is the central structure as which core serves the verbal form (Chikoidze 2010). Besides the top layer possesses direct participants of this process. Such as "causer" (what caused the action to occur) (CS), agent (AG), Object (OB), Addressee (recipient) (AD.) (Chikoidze 2015).

The sentence consists of certain groups of word forms - phrases (PH). In addition to the core, each group consists of some additional features. For example, when the core of a phrase is represented by a noun, we get the Name Phrase (NP). Some additional characteristics are attributed to him, both elementary: *grʒeli/mok’le/č’k’viani/suleli/lamazi/...* (long / short / smart / stupid / beautiful ...) and complex, which are realized through a dependent sentence. For example, *gogo, romelic k’argad mγeris; masc’avlebeli, romelmac amixsna es teorema; čem mier ašenebuli saxli; modurad čacmuli č’abuk’i* (a girl who sings well; the teacher who explained this theorem; the house built by me; modestly dressed young man).

An important initial stage in the formation of the content structure is the selection of the “dominant” situation, which directs this structure and hierarchically controls the rest of the composition. We choose a situation, or a sentential primitive, which dominates the syntactic structure of the expression. In particular, the core of V (verb) is accompanied by a “dominant” or “independent” verb defined by the syntax of the expression. From the dominant peak, the primary “rays” connect directly with the members of the dominant situation and, thus, unite the upper layer into the central structure of the nucleus, from which the lines are directed to the second layer (periphery). At the same time, any member of both levels can have its own attribute, which in some cases is represented as the following situation with proper sentential primitives.

The position and function of each component must be expressed by means of markers emerging from them rays. For example:

gušin rom momšivda mivedi Cems saxltam axlos mdebare maγaziaši p'uris saqidlad. 'Yesterday when I got hungry, I went to a shop near my house to buy bread.' Here is the dominant:

D *(me) (AG) mivedi maγaziaši (AD)*
'I (AG) went to the store (AD)'

The considered sequence actually expresses only the top layer of the dominant, which is accompanied by the periphery, primarily a time indicator:

T *gušin* 'yesterday'

The lower layer should also be attributed to the cause of the dominant situation:

C *me (AD) momšivda*
I (AD) got hungry

The dominant act (to come to the store) is followed by the purpose of the Act (R):

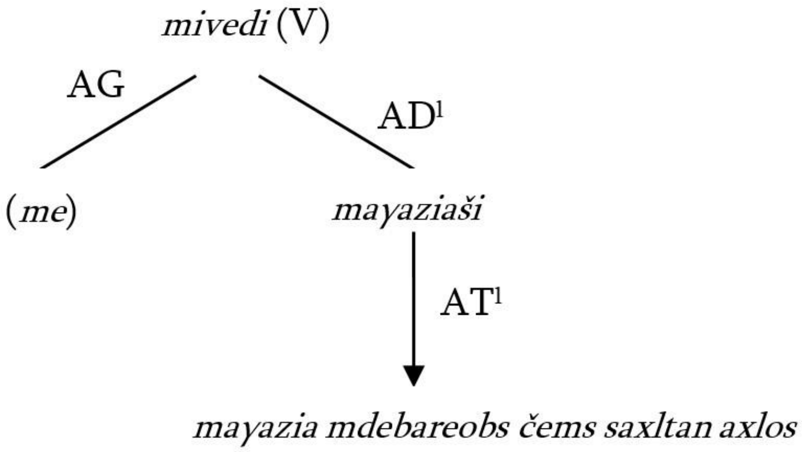
R *p'urs (OB) me (AG)viqididi*
Bread (OB) I (AG) will buy

Finally, this sample includes an example of the attribute (AT) situation:

AT *maγazia (AG) mdebareobs čems saxltan axlos (L)*
The Shop (AG) is located near my house (L)

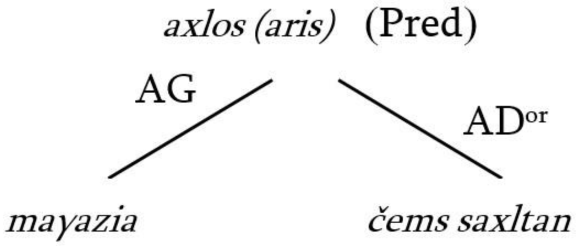
The above elements can be combined in a joint scheme of content, the head of which is the dominant verb (see Fig. 1):

Figure 1: Joint scheme with dominant verb and attribute



The attribute component can be replaced by synonyms: (store) *axlos (aris) čems saxltan* (close to my home). As a result, we get the scheme of this component (see Fig. 2):

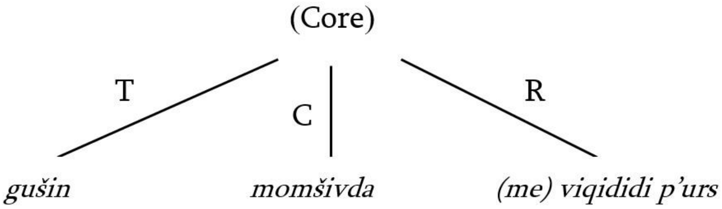
Figure 2: The scheme of the attribute component replaced by synonyms



Where the superscript "or" indicates that the addressee now performs the function of "reference point".

Let us finish the scheme with the designation of the ratio of the upper layer and periphery (see Fig. 3):

Figure 3: The finish scheme



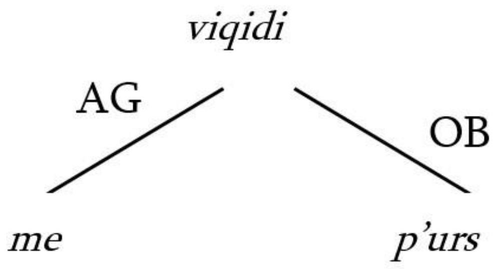
"Hungry" expresses the situation:

me(AD) momšivda 'I (AD) got hungry'

Where the implied pronoun 'I' means the role of the addressee of the feeling of 'hunger' (AD) (Chikoidze 2015).

In the case of the target (R) situation, if we choose the shape with future tense of verb, we will have (see Fig. 4):

Figure 4: The scheme of the shape with future tense of verb



The combination of the above schemes (as per our assumptions) must express the integrity of content of the initial expression in which all the private situations within the content of the expression and semantic forms are characterized by the interconnections of these components.

In perspective, this approach implies the formation of a universal generator of synonyms that will produce such statements, which will have the identical meanings, based on the original structure of the content.

3. Conclusion

Development of the approach is justified for its simplicity. We cannot go beyond the borders of natural language even while language cognition (Fillmore 1968), (Heidegger 1959) that will help us find “the key” of the concept of language content that is the most complicated problem inside the language.

According to the represented paper, the transformation model must be preceded by determining the main syntactic structure of the input text. Of course, this is complicated and comprehensive solution of this enormous task, and is practically impossible in the course of solving the core (also comprehensive) task. Considering the size and complexity of the tasks, it is natural to consider the current work as a definite step towards the ultimate solution of the problem.

The importance and purpose of the initial phase should be to generate the basic principles of the future comprehensive system and test the suitability and effectiveness of the assumptions on a limited material.

In particular, it should be noted in this regard that the transformation system will take into account only simple Georgian proposals, and those that are constructed under the dominance of the verb narrative form. The behaviour of these types of expressions within the quasisynonyms transformation was the main object of the research conducted.

References

[1] Apresjan Ju., D., (2003) *The New Explanatory Dictionary of Russian Synonyms.* (in Russian) Moscow.

[2] Apresjan Ju., D., (1973). *Principles and Methods of Contemporary Structural Linguistics*, Transl. by Crockett D. B.. The Hague: Mou-ton.

[3] Apresjan Ju., D., (1992). *Lexical Semantics. A Guide to Russian Vocabulary,* Ann-Arbor, Karoma Publishers.

[4] Van Valin, R. D. Jr. and LaPolla R. J. (1997). *Syntax. Structure, Meaning and Function*, Cambridge University Press.

[5] Goddard, Cl. and Wierzbicka A. (1994). *Semantic and Lexical Universals,* John Benjamins. Amsterdam/Philadelphia.

[6] Chikoidze, G. (2010). *Systematization of Some Meanings of Language Unit Classes*, Archil Eliashvili Institute of Control Systems, Tbilisi, Monograph (in Russian).

[7] Chikoidze, G. (2015). *Semantics of Units Defining a Sentence Structure*, Monograph, "Universali", Tbilisi (in Georgian).

[8] Fillmore, Ch. (1968). *The Case for Case. In "Universals of Linguistic Theory"*, Publ. Holt Rimebort and Winston Inc.

[9] Heidegger, M. (1959), *On the Way to Language*. Berlin.