

DEVELOPMENT OF ARMENIAN LANGUAGE MODULE OF UNL*

Vahan AVETISYAN, Vladimir SAHAKYAN, Lussine PETROSYAN¹
Robert URUTYAN, Angela MANUKYAN, Liana HOVSEPYAN²

Abstract — We present some results on the development of Armenian Language Module for Universal Networking Language. The solutions for morphological phase and the results for the first phase of development of Armenian UNL Server is shown. This research is supported by UNDP

INTRODUCTION REMARKS

Historically Armenia was situated on the crossroads of great ancient civilizations and the Armenian language is one of the oldest languages of the world. It belongs to the Indo-European family of languages as a separate branch which has no nearest cognates among the other branches. The written sources of Armenian go back to the 5th century A.D., so it has rather long literary traditions. From the initial period of its formation being located in the surrounding of the cultural languages of the Near East Armenian had the opportunity to develop its structure and enrich its vocabulary so as it could express the most complicated scientific conceptions and possessed all means of expressiveness for high poetry and fiction.

At present the literary Armenian has two variants – East Armenian and West Armenian. The first is the national language of the Republic of Armenia and the second is spoken by the majority of the Armenian Diaspora. Armenian has also territorial dialects spread both in Armenia and abroad.

The grammatical structure of Modern Armenian is rather complicated – with numerous and multiform morphological elements. As a whole it belongs to the agglutinative-synthetic type of languages but contains also a significant number of inflectional and analytical constructions. The standard Modern Armenian has a considerable stylistic and functional stratification of the vocabulary, a rich phraseology, and well-developed means of word-formation. Therefore Armenian can be ranked among the most developed modern languages owing inter alia to its ability to express through translation all the semantic and stylistic nuances of the original language.

Since the very beginning of the classical Armenian literature at the dawn of the 5th c. parallel to the original writings of Armenian authors a great number of works of Greek, Syrian, and Arabian scientists and writers have been

translated into Armenian. The Armenian versions of works of Plato, Aristotle, Euclid, Dionis of Thrace, Philo of Alexandria and others are of great importance because of the high degree of exactness of the translation and their literary quality. Some of these authors' works are maintained only in Armenian versions thus keeping the value of originals.

Nowadays, when scientific-technical progress has opened the borders between many countries, the language differences have become one of the most noticeable obstacles on the way of integration of entire human knowledge. The common natural language creation efforts have failed, and the translation is still the only way to overcome this obstacle. Thus, while taking in consideration the huge resources needed for translation, becomes obvious that only a small part of today's entire information is translatable. Yet, at present the machine translation is still the highest priority problem.

This problem arises while dealing with the World Wide Web, where the language boundaries are reducing the information retrieval.

As in the past, today also the main question that highlights during the development of machine translation applications is whether the algorithm should make translation in terms of the translator skill or the linguist skill. The entire research made in this area till today doesn't give enough ground to solve this problem.

In this paper we try to present the works that have been done in machine translation field in Armenia and highlight the new approaches to participate in the development of a world wide translation system.

The first attempts to use machines for translation automation were made when the first computers emerged in 60s of past century. Armenia also was involved in this process.

As globally adopted in practice it is possible to pass from one language code to another and is practically implemented due to the common principles existing in most languages. Consequently the word, structural and semantic levels are equivalent.

The past experience of Armenian scientists is connected with the first two levels, so the semantic analysis and synthesis have never been taken into consideration. As a result of input sentence analysis a tree of syntactical dependences was generated and replaced with the equivalent

¹ Vahan AVETISYAN, Vladimir SAHAKYAN, Lussine PETROSYAN, Institute for Informatics and Automation Problems of National Academy of Sciences of the Republic of Armenia 1, P. Sevak str., 375014, Yerevan, Armenia va@ipia.sci.am

² Robert URUTYAN, Angela MANUKYAN, Liana HOVSEPYAN, Institute of Linguistics of National Academy of Sciences of the Republic of Armenia 15 Abovyan str. Yerevan, Armenia

tree on the target language afterwards, using equivalent entry replacement principle.

In 1963-67s in the Institute for Informatics and Automation Problems of National Academy of Sciences of the Republic of Armenia a group of scientists created a dedicated machine "Garni", which was performing translation from Russian to Armenian. In 3 point scale "Garni" was estimated 1.5 – 2 points. [2], [3]

The further practice showed that the translation algorithm needs to be accomplished with semantic analysis schemas. The old equivalent replacement principle should be relinquished. A new multi-meaning translation approach emerged, which Armenian scientists were researching in cooperation with Russian colleagues.

ARMENIAN LANGUAGE MODULE

As the word has a two-plan representation (expression plan and meaning plan) during the translation the algorithm goes from the expression plan to meaning and back to expression. And if taking into account that the meaning plan is common for most languages the idea of an artificial language comes up. Usage of a such language would give a chance to create multi-language translation system. Particularly as a basis for our work the Universal Networking Language (UNL) [10] – [13] was taken. The UNL program considers that each member of the society (representing his language) must develop his module of UNL to perform conversion from native language to UNL and vice versa, using approach presented in Figure 1 of the Appendix, and finally join to the global translation system. The team of scientist from the Institute for Informatics and Automation Problems and Institute of Linguistics of National Academy of Sciences of the Republic of Armenia also started the development of Armenian Module of UNL.

As UNL provides a special syntax for dictionaries of native words and morphemes, grammatical rules for conversion from native language to UNL and vice versa, knowledge representation and grammatical attributes. So for the development of Armenian Module we face the creation of all these resources.

The first phase of our work considered two main tasks:

- Formalization of morphology of Armenian language.
- Development of a special environment for the management of grammatical and semantic resources.

As an outcome of first phase we formalized all templates of Armenian morphology using enconversion rules of UNL, presented and tested in 1000 most usable morphemes.

Also a web based environment was created, which serves as a management tool for the resources needed for translation. It contains Dictionary management service, Knowledge Base (Universal Words) management service, Rule management and Attribute management services. The

Environment also has tools for testing and debugging of rules (Figures 2 - 4).

MORPHOLOGY

Armenian language is notable for variety of different morphological categories, a large amount of syntactic and semantic level information can be obtained on morphological level. This generates a need for formalized description of modern Eastern Armenian morphology and attributes set developed for its following presentation in UNL format [4] – [6], [14] – [15].

For the formalized description of morphology the solution of the following tasks is necessary:

- 1) To develop the rules of word-forms articulation on morphs and generation of word-forms; on the basis of these rules the vocabularies and tables of morphs and the calculus of rules for generation and analysis are developed;
- 2) To determine dependences of grammatical categories from morphs, that constitute the word form and to develop the rules of definition of the word-form grammatical characteristics;
- 3) To determine calculus of all word-forms, produced by the given stem;

We shall show the solution of the listed tasks on an example of a verb as the most morphologically rich part of speech of modern Eastern Armenian language.

1* For the description of generation and analysis of the verbal forms we shall use operations concatenation (*) and addition (+). Then the rules of the verb form generation will be as follows (calculus of rules):

The direct forms of participles

Part=O *Sp • **fi**É**, • **fi*áú**

The negative forms of participles

ã*Part1 **ã* fi**É**

The simple personal forms of a verb

O *Fv • **fi**Û**

The negative simple personal forms of a verb

ã*O *Fv **ã* fi**Û**

The indirect forms of participles

Part1 *Fn, • **fi**É*áí**

The forms of a conditional mood

İ*O *Fv **İ* fi**Û**

The compound forms

Part2+Va • **fi*áú+»Û**

The negative compound forms

Van+Part2 **ã»Û • fi*áú**

The forms of a compulsive mood

á Çi Ç+ Part2/Part2+ á Çi Ç

The negative forms of a compulsive mood

ã*á Çi Ç+ Part2,

where

Part1 = {Inf, Res, Sub, Fut}, (Inf -infinitive, Res – perfect, Sub - subjective, Fut – future) – declinable participles,

Part2 = {Imp, Hyp, Plu, Adv, Neg}, (Imp - imperfect, Hyp - hypothetical (future), Plu - pluperfect, Adv – adverbial, Neg –negative) – indeclinable participles,

O - stem,

Sp – suffix of participle,

Fv - verb flexion,

Fn - noun flexion,

Fs – participle suffix,

Va - auxiliary verb,

Van – negative auxiliary verb.

2* The verb is characterized by the following 10 grammatical categories:

Conjugation: Con1 = {Pos, Neg}, where Pos - positive and Neg - negative conjugations,

Types of conjugation: Con2 = {→É, →Ē}.

Mood. M= {Ind, Imp, Opt, Con, Comp}, where Ind – indicative, Imp – imperative, Opt – optative, Con – conditional, Comp – compulsive mood.

Tense. Tns = {Pres, Fut, Past}, where Pres- present, Fut - future, Past - past tense.

Voice. Voc = {Act, Pas, Med, Caus} where act - active, pas- passive, med - medial voice, Caus - causative..

Aspect. Asp = {Perf, Imp, It}, where Perf –perfect, Imp - imperfect, It- iterative.

Number. Num = {Sg, Pl, St, Pt}, where Sg -singular, Pl - plural, St - the forms not having plural number (singularia tantum), Pt - forms which are not having singular number (pluralia tantum).

Person. Pers = {P1, P2, P3}, where P1 - first, P2 - second, P3 - third person.

Case. Case = {Nom, Gen, Dat, Acc, Abl, Instr, Loc}, where Nom - nominative, Gen - genitive, Dat - dative, Acc - accusative, Abl - ablative, Ins - instrumental, Loc - locative case.

Actualization. Def = {def1, def2, Pers1, Pers2}, where def1 - absence of an article, def2 - definite article Ի/Ի, Pers1 – possessive article of first person (–Է), Pers2 – possessive article of second person (–Ի).

Two last categories (Case and Def) are nominal and characterize only forms of declinable participles.

The combinations of values of these categories constitute the characteristics of morphs and wordforms. For example, if the characteristic of Fv is Pres, Sg, P1 (→Ü), of O is Reg1,Act,Con1 and of Ps is Imp, then the characteristic of the word-form is Reg1,Act, Con1, Imp, Pres, Sg, P1.

3* About 700 word-forms can be constituted from one verb stem. The verb stem is described by Patterns. In total about 30 Patterns for verbs are developed. The Pattern1 describes the stem with the characteristic Reg1,Act, Con1.

Pattern1 O(Reg1,Act, Con1)

© Convergences '03

(Inf, Res, Sub, Fut, chInf, chRes, chSub, chFut,

Imp, Hyp, Plu, Adv, Neg,

MInd, MInd(PastPerf), MImp, MOpt, MCond, MComp, chMInd, chMInd(PastPerf), chMImp, chMOpt, chMCond, chMComp,

Inf(v), Res(v), Sub(v), Fut (v), chInf(v), chRes(v), chSub(v), ch Fut (v),

Imp(v), Hyp(v), Plu(v), Adv(), Neg(v),

MInd(v), MImp(v), MOpt(v), MCond(v), MComp(v),

chMInd(v), chMImp(v), chMOpt(v), chMCond(v), chMComp(v)).

The negative forms are designated by means of addition to a basis at the left of negative particle (ch- օ), and form of a passive voice (pas) - by addition to the stem on the right before inflexion or the participle suffix the passive afflex (v – ւ).

The paradigm of a verb includes 4 initial forms of declinable participles Part1, being declined on a sample of nominal declination. In Armenian language for each initial form of a declinable participle 128 forms are realized. The total number of forms for the declinable participles is 512.

The calculus of the simple and compound regular verbal forms of modern Armenian language constituting from one stem is equal to 220 forms, and together with the participles – to 732.

Examples of rules for information preliminary transformation for UNL representation are presented below.

O v, Part1Res, Sp = 3 Ի	→ +@ state	Վե՛ւ → ւ Ի
O v, Part1Res, Sp = 3 Ի, Va=»Ü (1 Sg, pres)	→ +@ state+@ present	Վե՛ւ → ւ Ի »Ü
O v, Part1Res, Sp = 3 Ի, Va=Ճ (1 Sg, past)	→ +@ state+@ past	Վե՛ւ → ւ Ի Ճ
O v, Part2Imp, Va=»Ü (1 Sg pres)	→ +@ present+@ progress	• ն-ձժ »Ü
O v, Part2Imp, Va=Ճ (1 Sg past)	→ +@ past+@ progress	• ն-ձժ Ճ
O v, M=Ind, Fv=»Օ (1 Sg pastperf)	→ +@ past+@ complete	• ն-»Օ Ճ
O v, Part2Pl, Va=»Ü (1 Sg, pres)	→ +@ end +@ past	Ի 3 ն 1 → Օ »É »Ü
O v, Part2Hyp, Va=»Ü (1 Sg, pres)	→ +@ future	• ն-»Էձ »Ü
O v, M=Opt, PreP=Ի, Fv=»Ü (1 Sg, pres)	→ +@ future	Ի → ն-»Ü
O v, Fv=Շ (Sg)	→ +@ in perative	• ն-Շ
O v, Fv= »Պ/»Օ »Պ (Pl)	→ +@ in perative+@ pl	ն-»Պ/ն-»Օ »Պ
M=Opt, PreP=ձ Շ Է	→ +@ intention	ձ Շ Է • ն »Ü
O v(Iter), Part2Imp+Va (»Ü) (Pres 1Sg)	→ +@ present,+@ progress, +@ repeat	Ի Ի ն Ի ձժ »Ü
O v(Iter), Part2Imp+Va (Ճ) (Past 1Sg)	→ +@ past,+@ progress, +@ repeat	Ի Ի ն Ի ձժ Ճ
O v, Part=Part2Pl, Va=»Ü (2Sg pres),		»ն ւ շ Է Ի 3 ն 1 → Օ »É »Է

»ն ւ շ Է Ի 3 ն 1 → Օ »É »Է

December 2 - 6, 2003, Alexandria, EGYPT

International Conference on the Convergence of Knowledge, Culture, Language and Information Technologies

The submitted formalized model of the noun's morphology is realized in the Armenian Module of UNL as well as the morphology model of verb, pronoun etc. They are represented by 1000 most usable morphemes. This set of words covers all morphological forms present in Armenian language, so the future extension of the dictionary won't need new templates to be declared.

This formalized models are kept in the vocabulary of stems with the characteristics (patterns) and tables of auxiliary morphs with grammatical characteristics. Also the grammatical rules of the analysis and synthesis of the forms are presented according to UNL rules format. The software described below allows to carry out the analysis and synthesis of Armenian language words, based on described model.

Parallel to the extension of the dictionary works on the development of rules for syntactic analysis and generation has started. These works consider the formalization of grammatical rules of Armenian Language, their presentation in UNL enconversion and deconversion rules format.

DEVELOPMENT ENVIRONMENT

During the development of dictionaries and rules of the Armenian Language Server developers deal with big amount of data representing dictionaries, attributes and rules. A need for a special data management software becomes obvious. The development of such application has started and is in process. The application called Development Environment for Armenian UNL Server serves as for the management of the linguistic data so as for testing and presentation purposes. The Environment is developed as a web application for the best multi-user performance over the internet. As a platform Java's JSP and JavaBeans technology was taken. For the storage of the linguistic data "MySQL" relational database was chosen as back-end. In the database the dictionaries, attributes and rules are stored in tables related via many-to-many and one-to-many relations. This approach helps to distinguish the informational units of the language such as native word, grammatical attribute or universal word from the knowledge representing the structure of the grammatical and semantic relations of the natural language.

The application operates over the database and generates an object model for the data entry and language processing services. All further operations are being performed over this object model.

The Web Application of Armenian UNL Server is designed to be a full functional dynamic environment for all activities connected with UNL. The application is being built using a flexible structure which allows new services to be included. For now the following services are already developed and included:

- Dictionary management service,
- Enconversion rules management service,
- Attributes management service,

- Knowledge Base (Universal Words) management service,
 - Tools for testing and debugging of rules.
 - Content service as an information source providing latest news about the progress of the development and
 - User management service for setting user access levels for services
- (Figures 2-4).

The following tasks are described for the second phase of development of Development Environment for Armenian UNL Server.

- Deconversion rules management service.
- Tools for exporting the databases into textual format (to give ability to represent our resources in UNL's ENCO/DECO applicable format).
- Integration of ENCO/DECO software onto the Environment.
- Additional tools for testing and debugging.

The creation of such software is a step towards the popularization and simplification of the development of UNL Language Servers. This environment is developed for Armenian language, but is designed as language independent system, which can be easily customized for any other language. The main purpose for the creation of this software is to place the UNL processing under database basis, instead of textual representations of the data, which is used now. This approach brings to more flexible data management, speed optimization and minimizes potential errors and bugs.

The authors express their thankfulness gratitude to prof. Yu. Shoukouryan and prof. I. Zaslavsky for valuable remarks and consultations.

REFERENCES

1. Automated translation. Progress, Moscow, 1971
2. Mathematical Problems of Cybernetics and Computer Science. (Dedicated Machine for Automated Translation). Yerevan, 1972
3. R.A. Bazmajyan, M.I. Beletsky, V.M. Grigoryan "About the algorithm of Armenian-Russian Machine Translation" Problems of Cybernetics, N14, Moscow, 1965
4. Angela Manukyan, Liana Hovsepian "The Automated Recognition of Compound Verbs in Old Armenian Text". Computers in Armenian Philology. Academy Press. Yerevan 1993
5. Angela Manukyan, Robert Urutyan "The Formalized Description of Modern Armenian Word-formation" Math. of the International Conference on Armenian Linguistics. Yerevan 1994 (Arm.)
6. Angela Manukyan "A Formalized Model for Ancient Armenian Verb" Proceedings of International

- Conference on Computer Science and Information Technologies, Abstracts. Yerevan 2001(Rus.)
7. Robert Urutyan “Problems of Transformations of Dependency Trees” Text Analysis Problems. Yerevan. 1975 (Rus.).
 8. Robert Urutyan “Armenian Substantives” American Journal of Computational Linguistics. New-York. 1979.
 9. Robert Urutyan “Analysis of Equivalence in Language by means of D-grammars“ Symposium on Grammars of analysis and synthesis and their representational structures. Tallinn. 1983.
 10. Uchida Hiroshi, Zhu Meiyang, Tarcisio Della Senta, The UNL, “A Gift for a Millenium”, UNU/IAS, Tokyo,1999.
 11. Uchida, Hiroshi, Zhu, Meiyang. “The Universal Networking Language beyond the Machine Translation” International Symposium on Language in Cyberspace. Seoul, 2001.
 12. Andre Bortolon, Hugo Cesar Hoeschl, Joel Ossamu Mitsui, Jaime Leonel de Paula Júnior, Ricardo Miranda Barcia, “A proposal of an UNL Application Development Environment” 2002
 13. UNL Center. Enconverter Specifications. Version 3.3 Tokyo, 2002.
 14. G. Jahukyan. The Universal Linguistic Model. Yerevan. 2000. (Rus.).
 15. A. Manukyan. Formalized Model for Ancient Armenian Verb//Proceedings of the International Conference on Computer Science and Information Technologies, Abstracts. Yerevan. 2001.(Rus.).
 16. V. Avetisyan, A. Manukyan. Development Environment for the Armenian Language Server of UNL. Proceedings of the International Conference on Computer Science and Information Technologies, Abstracts. Yerevan. 2003.
 - 17 T. Dhanabalan, K.Saravanan, T.V. Geetha, Tamil to UNL EnConverter. Goa. India. 2001

