Advances in Computer Science and Information Technology (ACSIT)
p-ISSN: 2393-9907; e-ISSN: 2393-9915; Volume 6, Issue 2; April-June, 2019, pp. 111-115
© Krishi Sanskriti Publications
http://www.krishisanskriti.org/Publication.html

Machine Translation with UNL: A Study of Kashmiri

Naziya Rasool, Sumaira Nabi and Younis Rashid Dar

Ph.D Scholars, Department of Linguistics, University of Kashmir
E-mail: naziyarasoo14@gmail.com, sumaira.nabi@rediffmail.com, younisrashid_7@yahoo.com

Abstract—*Universal Networking Language (UNL) is a computer language created to represent and process information across language barriers (Uchida et al., 2001).UNL is basically a knowledge representation language i.e. it is used to represent information conveyed by natural languages (Cardenosa et al, 2009).UNL has been used by various researchers as a tool for machine translation (MT). IAN and EUGENE are the two main tools in the UNL system. The conversion of source language to UNL is known as UNLisation. While as, the process of generating natural language text from the UNLised text is known as NLisation. IAN and EUGENE, the two online tools created by the UNDL Foundation perform the process of UNLisation and NLisation efficiently. The aim of the present work is to incorporate Kashmiri in the UNL framework. This work involves the analysis of selected Kashmiri corpus to UNL using IAN, and then the generation of Kashmiri from the UNL expressions using EUGENE. For the present work, UC- A1 corpus (available online) was taken and the accuracy calculated for the selected corpus is estimated to be around 90%. This work will act as an important milestone in developing a robust multilingual MT system for Kashmiri.*

Keywords: *Machine Translation (MT), UNL, IAN, EUGENE.*

1. INTRODUCTION

Machine Translation (MT), also known as automatic translation is the translation of text from one natural language to another by a machine. The area of Machine Translation (MT) has witnessed integration of research work from other fields like Statistics, Mathematics, Artificial Intelligence, etc during its expansion. Different approaches like Direct MT, Rule-based MT, Corpus-based MT, and Knowledge-based MT have been put forth from time to time for carrying out the process of MT. Universal Networking Language (UNL) based MT is also an attempt in this direction.

1.1 Universal Networking Language

"Universal Networking Language is a declarative formal language designed to represent the semantic data extracted from natural language texts.." In the UNL approach, information extracted from the NL text is represented in the form of a semantic graph.UNL expressions are semantically complete and unambiguous which is the biggest advantage of this system. UNL works independently for any language.

UNL system involves two processes:

- UNLisation: It is the enconversion of Source Language text to UNL format and is done by the Encoverter tool using the respective language's NL-UNL dictionary and T rules.
- NLisation: It is the deconversion of UNL expression to the target language and is done by the Deconverter tool using the target language's UNL-NL dictionary and generation rules.

1.2 The Components of UNL system

UNL system involves the following components**:**

1.2.1 Language Servers

- Language Servers involve the following tools:
- EnConverter - This tool enconverts natural language text into the UNL expression.
- DeConverter - This tool deconverts the UNL expression into native languages.

1.2.2 Software Tools

Software Tools involve the following components:

- UNL Editors – UNL Editors are used to make UNL documents.
- UNL Explorers – UNL Explorers are used to view/manage UNL document by accessing UNL language servers, UNLKB & UNL Documents.
- UNL Verifiers – It verifies UNL expression for correctness.
- UNL Proxy servers – It Provides communication with language servers.
- Concept Definitions – It defines concepts in connection with other concepts.
- UNL Documents – These are the documents in which UNL expression is described for each sentence of natural language.

2. RELATED WORKS

Some of the research works done in the field of UNL are:

2.1 Decodification of UNL-Portuguese (Martins et al., 1997)

Martins, et al. (1997) described the set of grammar rules for the decodification of UNL-Portuguese using DeCoL 1.0 (Deconverter System for Latin Languages, version 1.0), the decodifier provided by the UNL Center. The set of rules has been based upon two experimental corpora, which comprise of 20 sentences of the UNL language booklet and Chapters XIV and XV of the UNU Charter. The rules express, then, a meaningful, but incomplete, part of the grammatical aspects of the Portuguese language. They are divided into three groups, which focus on distinct aspects of linguistic processing i.e. the insertion of lexical items in a sentence and the intrarelationship between such items, and the morphological treatment, which processes mostly word suffixing. Rules to graphically process the inputs are also defined.

2.2 Arabic Generation in the Framework of the Universal Networking Language (Daoud, 2005)

Daoud (2005) proposed an Arabic DeConversion system. The system proposes the architecture along with the strategy of this Deconversion system . In addition to this, problems related to the UNL representation that affect the quality of generation have also been discussed in this study.

2.3 UNL Nepali Deconverter (Keshari and Bista, 2005)

Keshari and Bista (2005) proposed the architecture and design of UNL Nepali Deconverter (Generator) that has been implemented using a tool called DeCo, a language neutral generator. There are basically two modules for UNL Nepali Deconversion. One is the syntax planning module and the other is the morphology generation module.

3. METHODOLOGY

The present work is analytical in nature. For this work, some of the UCA1 corpus (available online) and some randomly selected corpus was taken and UNLised and NLised using IAN (Interactive Analyser) and EUGENE (dEep to sUrface-Generator), the two online tools provided by the UNDL Foundation. First step involved the classification of words into Permanent and Temporary words. Second step involved the framing of NL-UNL and UNL-NL Dictionaries and final step was to frame analysis and generation rules in the respective tools.

4. ANALYSIS/GENERATION OF THE SELECTED CORPUS

This section presents the analysis and generation of the selected corpus. The selected corpus was UNLised using IAN and NLised using EUGENE. UNLisation and NLisation are two independent processes and this makes UNL a best tool for carrying out the process of MT. The work done can be presented as:

4.1 Temporary words

The analysis of Temporary words is given in the table below:

Table 1: Temporary Words in IAN

S. No.	Input	T-rule	UNL expression
1.	aedesf	(TEMP,%a)(BLK,%b)(TEMP,%c):=(%a&%b&%c);	"aedesf"
2.	aedesf aedesf	(TEMP,%a)(BLK,%b)(TEMP,%c):=(%a&%b&%c);	"aedesf aedesf"
3.	aedesf aedesf aedesf	(TEMP,%a)(BLK,%b)(TEMP,%c):=(%a&%b&%c);	"aedesf aedesf aedesf"
6.	aedesf aedesf aedesf aedesf aedesf aedesf	(TEMP,%a)(BLK,%b)(TEMP,%c):=(%a&%b&%c);	"aedesf aedesf aedesf aedesf aedesf aedesf"
7.	H_2o	(TEMP,%a)(BLK,%b)(TEMP,%c):=(%a&%b&%c);	"H_2o"
14.	C	(TEMP,%a)(BLK,%b)(TEMP,%c):=(%a&%b&%c);	"C"
15.	www.facebook.com	(TEMP,%a)(BLK,%b)(TEMP,%c):=(%a&%b&%c);	"www.facebook.com"
16.	Universal Networking Digital Language Foundation	(TEMP,%a)(BLK,%b)(TEMP,%c):=(%a&%b&%c);	"Universal Networking Digital Language Foundation"
17.	78.000	(TEMP,%a)(BLK,%b)(TEMP,%c):=(%a&%b&%c);	"78.000"
18.	H	(TEMP,%a)(BLK,%b)(TEMP,%c):=(%a&%b&%c);	"H"
19.	Googler	(TEMP,%a)(BLK,%b)(TEMP,%c):=(%a&%b&%c);	"googler"
20.	1001:43:9001	(TEMP,%a)(BLK,%b)(TEMP,%c):=(%a&%b&%c);	"1001:43:9001"
21.	1/2	(TEMP,%a)(BLK,%b)(TEMP,%c):=(%a&%b&%c);	"1/2"

4.2. Analysis of Noun Phrases in IAN

Analysis of Nouns in IAN can be explained with the help of some examples given below:

S.No	Result	Rules	Dictionary (NL- UNL)	English	Kashmiri Corpus
1.	pen.@both	Rule No. 1: (" "):=; Rule No. 2: (%a,D,POS=QUA,@both)(%b,N,POS=NOU,NUM=INV):=(%b,+att=%a)	[دۄشوے]{}"" (LEX=D,POS=QUA,att=@both)<kas,0,0>; both books [قلم]{}"" (LEX=N,POS=NUM,NUM=INV)<kas,0,0>; both books	*both pens*	دۄشوے قلم /dɔʃvaj kalam/
2.	book.@no	Rule No. 1: (" "):=; Rule No. 2: (%a,D,POS=QUA,@no)(%b,N,POS=NOU,NUM=SNG):=(%b,+att=%a)	[کٖہنی]{}"" (LEX=D,POS=QUA,att=@no)<kas,255,0>; no book [کِتاب]{}"book" (LEX=N,POS=NUM,NUM=SNG)<kas,0,0>; both books	*no book*	کٖہنی کِتاب /kɨhi:nʲ kita:b/
3.	book.@any	Rule No. 1: (" "):=; Rule No. 2: (%a,D,POS=QUA,@any)(%b,N,POS=NOU,NUM=SNG):=(%b,+att=%a)	[کانٛہہ]{}"" (LEX=D,POS=QUA,att=@any)<eng,255,0>; any book [کوٗر]{}"girl" (LEX=N,POS=NUM,NUM=SNG)<kas,0,0>;	*any girl*	کانٛہہ کوٗر / ka˜h ku:r/
4.	sparrow.@other	Rule No. 1: (" "):=; Rule No. 2: (%a,D,POS=QUA,@other)(%b,N,POS=NOU,NUM=SNG):=(%b,+att=%a)	[بیٚاکھ]{}"" (LEX=D,POS=DEM,att=@other)<eng,255,0>; another book [ژٔر]{}"sparrow" (LEX=N,POS=NUM,NUM=SNG)<kas,0,0>;	*other sparrow*	بیٚاکھ ژٔر $_h$ / bʲa:k tsər/
5.	girl.@other	Rule No. 1: (" "):=; Rule No. 2: (%a,D,POS=QUA,@other)(%b,N,POS=NOU,NUM=SNG):=(%b,+att=%a)	[بیٚاکھ]{}"" (LEX=D,POS=DEM,att=@other)<eng,255,0>; another book [کوٗر]{}"girl" (LEX=N,POS=NUM,NUM=SNG)<kas,0,0	*other girl*	بیٚاکھ کوٗر $_h$ /bʲa:k ku:r/

Table 2: Analysis of Noun Phrases

4.2.1 Generation of Noun phrases in EUGENE

Table 3: Generation of Noun Phrases

S. No.	Result	English	Rules	Dictionary (UNL-NL)	UNL input
1.	دۄشوے قلم /dɔʃvaj kalam/	*both pens*	(%a,N,@both):=("دۄشوے"(" ")(%a,-@both);	[قلم]{}”pen”(LEX=N,POS=NOU,NUM=INV)	pen.@both
2.	کٖہنی کِتاب /kɨhi:nʲ kita:b/	*no book*	(%a,N,@no):=("کٖہنی "("")(%a,-@no);	[کِتاب]{}”book”(LEX=N,POS=NOU,NUM=INV)	book.@no
3.	کانٛہہ کوٗر /ka˜h ku:r/	*any girl*	(%a,N,@any):=("کانٛہہ"("")(%a,-@any);	[کوٗر]{}”girl”(LEX=N,POS=NOU,NUM=INV)	book.@any
4.	بیٚاکھ ژٔر $_h$ /bʲa:k tsər/	*other sparrow*	(%a,N,@other):=(" "بیٚاکھ"("")(%a,-@other);	[ژٔر]{}”sparrow” (LEX=N,POS=NOU,NUM=INV)	sparrow.@other

5.	بیاکھ کوٗر /bʲa:kʰ ku:r/	*other girl*	(%a,N,@other):=("بیاکھ"("")(%a,-@other);	[کوٗر]{}"girl"(LEX=N,POS=NOU,NUM=INV)	girl.@other

4.3 Analysis of Adjectival Phrases in the UNL system

In UNL, Adjectival phrases are represented as **mod (noun, adj).**

Table 4: Representation of Adjectives in IAN

S. No	Result	T-rule	Dictionary (NL-UNL)	English	Corpus
1.	mod(sky, white)	Rule No. 1: (" "):=; Rule No. 2: (%a,J)(%b,N):=(mod(%b;%a),+N);	[سفید]{}"white"(LEX=J,POS=ADJ)<kas,0,0>; white sky [آسمان]{}"sky"(LEX=N,POS=NOU,NUM=INV)<kas,0,0>; white sky	*white sky*	سفید آسمان /safe:d a:sma:n/
2.	mod(mug,glass)	Rule No. 1: (" "):=; Rule No. 2: (%a,J)(%b,N):=(mod(%b;%a),+N);	[شیشُو]{}"glass"(LEX=J,POS=ADJ)<kas,0,0>; glass mug [مگ]{}"mug"(LEX=N,POS=NOU,NUM=SNG)<kas,0,0>; white sky	*glass mug*	شیشُو مگ /ʃi:ʃu: mag/

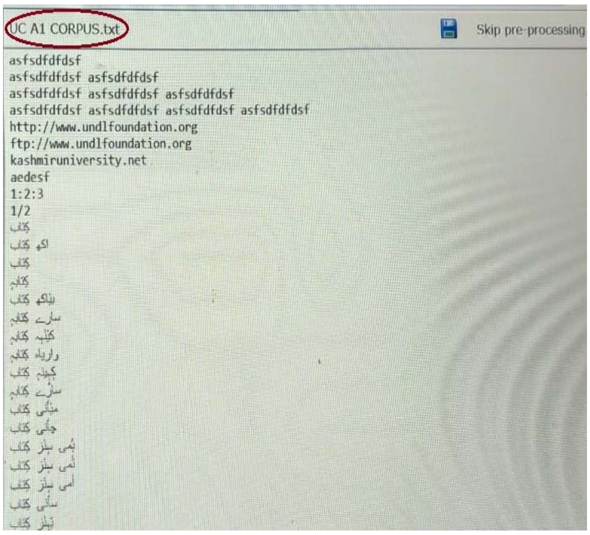
4.3.1 Generation of Adjectival Phrases in EUGENE

The generation of Attributive Adjectives is given in the table below:

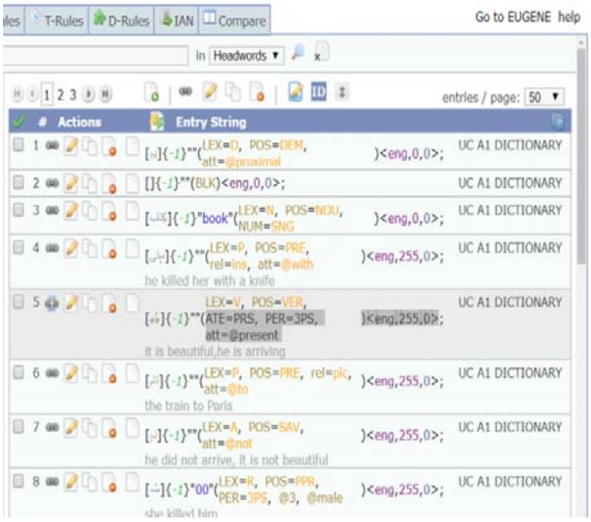
Table 5: Representation of Adjectives in EUGENE

S. No.	Result	English	Rules	Dictionary (UNL-NL)	UNL input
1.	سفید آسمان /safe:d a:sma:n /	*white sky*	mod(N,%a;J,%b):=(%b)(" ")(%a);	[سفید]{}"white"(LEX=J,POS=ADJ,NUM=SNGT)<kas,0,0>; white sky [آسمان]{}"sky"(LEX=N,POS=NOU,NUM=INV)<kas,0,0>; white sky	mod(sky,white)
2.	شیشُو مگ /ʃi:ʃu: mag/	*glass mug*	mod(N,%a;J,%b):=(%b)(" ")(%a);	[شیشُو]{}"glass"(LEX=J,POS=ADJ,NUM=SNGT)<kas,0,0>; glass mug [مگ]{}"mug"(LEX=N,POS=NOU,NUM=SNG)<kas,0,0>; white sky	mod(mug,glass)

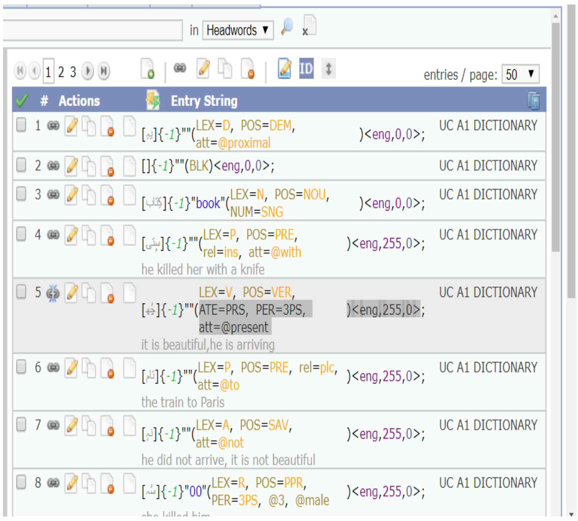
Likewise, the same procedure was adopted to include other parts of speech and the same is represented below:



```
asfsdfdfdsf
asfsdfdfdsf asfsdfdfdsf
asfsdfdfdsf asfsdfdfdsf asfsdfdfdsf
asfsdfdfdsf asfsdfdfdsf asfsdfdfdsf asfsdfdfdsf
http://www.undlfoundation.org
ftp://www.undlfoundation.org
kashmiruniversity.net
aedesf
1:2:3
1/2
کتاب
اکھ کتاب
کتاب
کتابہ
بیاکھ کتاب
سارے کتابہ
کٲیہہ کتابہ
واریاہ کتابہ
کیٛنژ کتاب
سارے کتابہ
منٛلی کتاب
جٲنی کتاب
بُمی بنز کتاب
لٔمی بنز کتاب
اٚمی بنز کتاب
سأنی کتاب
تیٛلر کتاب
```



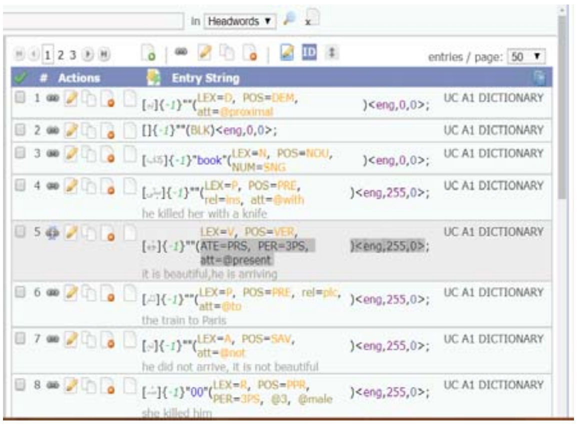
#	Entry String	
1	[یہ]{-1}""(LEX=D, POS=DEM, att=@proximal)<eng,0,0>;	UC A1 DICTIONARY
2	[]{-1}""(BLK)<eng,0,0>;	UC A1 DICTIONARY
3	[کتاب]{-1}"book"(LEX=N, POS=NOU, NUM=SNG)<eng,0,0>;	UC A1 DICTIONARY
4	[[illegible]]{-1}""(LEX=P, POS=PRE, rel=ins, att=@with)<eng,255,0>; he killed her with a knife	UC A1 DICTIONARY
5	[[illegible]]{-1}""(LEX=V, POS=VER, ATE=PRS, PER=3PS, att=@present)<eng,255,0>; it is beautiful,he is arriving	UC A1 DICTIONARY
6	[[illegible]]{-1}""(LEX=P, POS=PRE, rel=plc, att=@to)<eng,255,0>; the train to Paris	UC A1 DICTIONARY
7	[[illegible]]{-1}""(LEX=A, POS=SAV, att=@not)<eng,255,0>; he did not arrive, it is not beautiful	UC A1 DICTIONARY
8	[[illegible]]{-1}"00"(LEX=R, POS=PPR, PER=3PS, @3, @male)<eng,255,0>; she killed him	UC A1 DICTIONARY



in Headwords

1 2 3 entries / page: 50

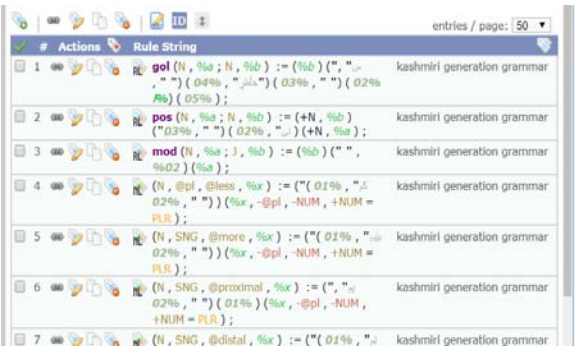
#	Actions	Entry String	
1		[يہ]{-1}""(LEX=D, POS=DEM, att=@proximal)<eng,0,0>;	UC A1 DICTIONARY
2		[]{-1}""(BLK)<eng,0,0>;	UC A1 DICTIONARY
3		[كِتاب]{-1}"book"(LEX=N, POS=NOU, NUM=SNG)<eng,0,0>;	UC A1 DICTIONARY
4		[سِتّی]{-1}""(LEX=P, POS=PRE, rel=ins, att=@with)<eng,255,0>; he killed her with a knife	UC A1 DICTIONARY
5		[چھُ]{-1}""(LEX=V, POS=VER, ATE=PRS, PER=3PS, att=@present)<eng,255,0>; it is beautiful,he is arriving	UC A1 DICTIONARY
6		[تام]{-1}""(LEX=P, POS=PRE, rel=plc, att=@to)<eng,255,0>; the train to Paris	UC A1 DICTIONARY
7		[نہٕ]{-1}""(LEX=A, POS=SAV, att=@not)<eng,255,0>; he did not arrive, it is not beautiful	UC A1 DICTIONARY
8		[تَس]{-1}"00"(LEX=R, POS=PPR, PER=3PS, @3, @male)<eng,255,0>; she killed him	UC A1 DICTIONARY



in Headwords

1 2 3 entries / page: 50

#	Actions	Entry String	
1		[يہ]{-1}""(LEX=D, POS=DEM, att=@proximal)<eng,0,0>;	UC A1 DICTIONARY
2		[]{-1}""(BLK)<eng,0,0>;	UC A1 DICTIONARY
3		[كِتاب]{-1}"book"(LEX=N, POS=NOU, NUM=SNG)<eng,0,0>;	UC A1 DICTIONARY
4		[سِتّی]{-1}""(LEX=P, POS=PRE, rel=ins, att=@with)<eng,255,0>; he killed her with a knife	UC A1 DICTIONARY
5		[چھُ]{-1}""(LEX=V, POS=VER, ATE=PRS, PER=3PS, att=@present)<eng,255,0>; it is beautiful,he is arriving	UC A1 DICTIONARY
6		[تام]{-1}""(LEX=P, POS=PRE, rel=plc, att=@to)<eng,255,0>; the train to Paris	UC A1 DICTIONARY
7		[نہٕ]{-1}""(LEX=A, POS=SAV, att=@not)<eng,255,0>; he did not arrive, it is not beautiful	UC A1 DICTIONARY
8		[تَس]{-1}"00"(LEX=R, POS=PPR, PER=3PS, @3, @male)<eng,255,0>; she killed him	UC A1 DICTIONARY



entries / page: 50

#	Actions	Rule String	
1		gol (N , %a ; N , %b) := (%b) (", "ہ , " ") (04% , "دلہ") (03% , " ") (02% /%) (05%) ;	kashmiri generation grammar
2		pos (N , %a ; N , %b) := (+N , %b) ("03% , " ") (02% , "ہ) (+N , %a) ;	kashmiri generation grammar
3		mod (N , %a ; J , %b) := (%b) (" " , %02) (%a) ;	kashmiri generation grammar
4		(N , @pl , @less , %x) := ("(01% , "کـ 02% , " ")) (%x , -@pl , -NUM , +NUM = PLR) ;	kashmiri generation grammar
5		(N , SNG , @more , %x) := ("(01% , "ـہ 02% , " ")) (%x , -@pl , -NUM , +NUM = PLR) ;	kashmiri generation grammar
6		(N , SNG , @proximal , %x) := (", "یہ 02% , " ") (01%) (%x , -@pl , -NUM , +NUM = PLR) ;	kashmiri generation grammar
7		(N , SNG , @distal , %x) := ("(01% , "ہۄ	kashmiri generation grammar

5. CONCLUSION:

The aim of this paper is to involve Kashmiri in the UNL framework. Till date, very little has been done regarding the development of machine translation system for Kashmiri, so this work is an important step in this direction. This work involved the analysis and generation of selected Kashmiri corpus using IAN and EUGENE. The rules framed during this work are open ended in nature and can be reapplied to other language structures with little or no modification. The accuracy of the system was measured using a technique known as F Score which came out to be .89 for this work meaning an accuracy of 89%.

6. SCOPE FOR FUTURE WORK:

This work paves a way for developing a robust multilingual machine translation system involving Kashmiri as one of the languages. This work leaves scope for future in entertaining more and more structures of Kashmiri language in the UNL system in order to make it complete and efficient in every way.

REFERENCES

[1] Adly, Noha, and Sameh Al Ansary. "Evaluation of Arabic Machine Translation System based on the Universal Networking Language." Natural Language Processing and Information Systems, edited by Helmut Horacek et al., Springer, 2009, pp. 243-257.

[2] Ali, Md Nawab Yousuf, et al. "Morphological Analysis of Bangla Words for Universal Networking Language." *Digital Information Management,* 2008. *IEEE Xplore,* doi.org/10.1109/ICDIM.2008.4746734.

[3] Alansary, Sameh, et al. "A Semantic Based Approach for Multilingual Translation of Massive Documents." *The 7th International Symposium on Natural Language Processing (SNLP),* Pattaya, Thailand, 2007.

[4] Bokil Hrushiklesh, M. "Towards Marathi Sentence Generation from Universal Networking Language." 2002. IIT Bombay, M Tech dissertation.

[5] Boudhh, Sangharsh, and Pushpak Bhattacharyya. "Unification of Universal Word Dictionaries Using WordNet Ontology and Similarity Measures." *proceedings of the 7th International Conference on Computer Science and Information Technologies*, Armenia, 2009, pp. 271-283.

[6] Cardeñosa, Jesús, et al., editors. *Universal Networking Language: Advances in Theory and Applications.* Centre for Computing Research of IPN, 2005.

[7] Cardeñosa, Jesús, et al. "Standardization of the Generation Process in a Multilingual Environment." *Proceedings of the International Conference Convergences*, 2003.

[8] Jurafsky, Daniel, and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Prentice Hall, 2000