A Platform for Experimenting with UNL

Wang-Ju Tsai

GETA, CLIPS-IMAG
BP53, F-38041 Grenoble cedex 09 France
Wang-Ju.Tsai@imag.fr

Abstract. We introduce an integrated environment, which provides the initiation, information, validation, experimentation, and research on UNL. This platform is based on a web site, which means any user can have access to it from anywhere. Also we propose an XML form of UNL document as the base of future implementation of UNL on the Internet.

1 Introduction

Since proposed 5 years ago, UNL project has attracted 16 international teams to join and is regarded as a very promising semantic Interlingua for knowledge representation on the Internet. The articles and applications of UNL have been found in many domains such as: machine translation, information retrieval, multilingual document generation, etc. Now we can find on the Internet not only the web sites of UNL language centres but also some discussions. The applications to facilitate the usage of UNL have been produced as well. Now we see the need to create a platform to integrate these applications also to introduce UNL to new ordinary users. We create this platform on a web site SWIIVRE (http://www-clips.imag.fr/geta/User/wang-ju.tsai/welcome.html), which has several goals: for the initiation, information, verification, research, and experimentation of UNL. And since this platform is based on a web site, any user from anywhere can have access to it.

2 Introduction of the Site SWIIVRE

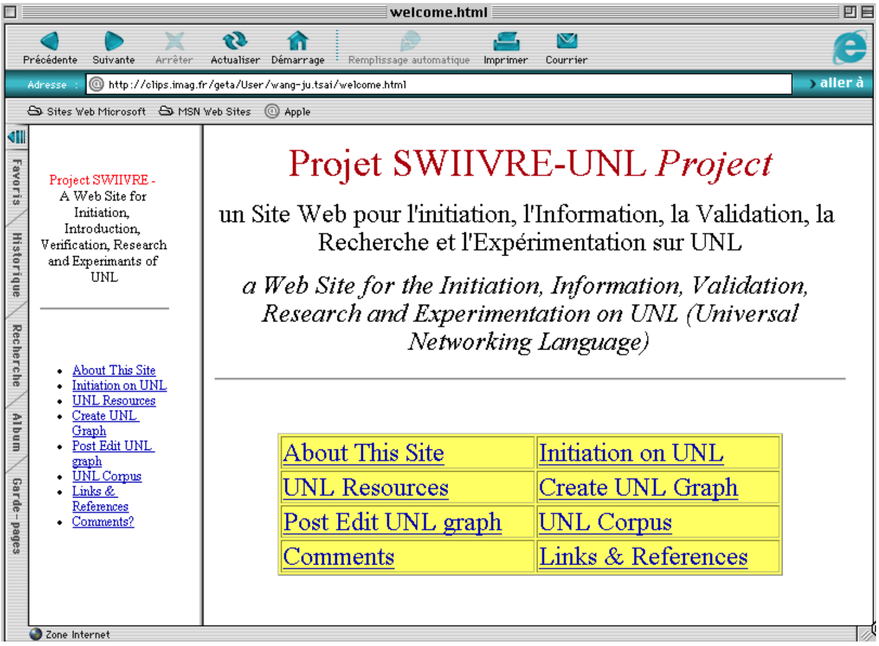
In Appendix I we list all the resources accessible for UNL society members from internet. We can find out that most of the LC's connect vertically to UNL Centre but the horizontal connection among LC's is not enough, which means any user who wants to try the multilingualism of UNL will feel frustrated, since he will need to spend a lot of time try out every LC to know what service he can get.

The main purpose of this site is rather to integrate the current UNL applications and complete the services of Language Centres', when the function is available on a Language Centre, we simply provide the link to it, we also produce some applications to integrate or provide new functions, which all serve to facilitate the usage of UNL. Also we collect the useful information and publications on UNL, the web site is updated regularly. Lastly, by collecting the useful information and recording the related

© J. Cardeñosa, A. Gelbukh, E. Tovar (Eds.)
Universal Network Language: Advances in Theory and Applications.
Research on Computing Science 12, 2005, pp. 261–267.

data, this site finally can serve as an evaluation of the performance of UNL community.

Here we show the welcome page of this site:



Project SWIIVRE - A Web Site for Initiation, Introduction, Verification, Research and Experimants of UNL

- About This Site
- Initiation on UNL
- UNL Resources
- Create UNL Graph
- Post Edit UNL graph
- UNL Corpus
- Links & References
- Comments?

Projet SWIIVRE-UNL *Project*

un Site Web pour l'initiation, l'Information, la Validation, la Recherche et l'Expérimentation sur UNL

a Web Site for the Initiation, Information, Validation, Research and Experimentation on UNL (Universal Networking Language)

About This Site	Initiation on UNL
UNL Resources	Create UNL Graph
Post Edit UNL graph	UNL Corpus
Comments	Links & References

The following is a description of each link on the welcome page:

About This Site This page provides the introduction, why and how this site exists, the site log and current status of this site, also the new projects to come on this site, lastly all the recent activities of UNL community. When clicked, a news flesh will also show the most recent UNL activities and the new updates on this site. In the future, we think we will at least UNL-ise this page to demonstrate the multilingualism of UNL.

Initiation on UNL This page is to help users to take a first step in UNL, understand how UNL works. We first provide a copy of most recent UNL specifications, for the moment only Spanish Centre has prepared a “multilingual interactive page” can serve as the tutorial and give examples to each UNL relations, thus we put a link to this page. When UNL becomes more well known, there will be more and more tutorials for beginners in the future. Or we might finally create an graphical interface for user to manipulate and show the spirit of UNL. We would also like to introduce the XML-UNL document here. We put an example of XML-UNL document here and with the help of XSLT, we can create the same effect like UNL browser, then the users can choose to read the document in the language they wish. We will explain later in the article why we want to XML-ise a UNL document.

UNL Resources This page provides all the UNL<->NL deconverters / enconverters, dictionaries that are accessible on the Internet. Some deconverters accept the deconversion of one single UW (Universal Word), in this case they can serve as the UNL-NL dictionaries. We can simply add some scripts in our site to help users to access these deconverters as if they are accessing dictionaries. In the future, the status report of each server will be added; we hope we can provide "UNL daily bulletin" to report the updates and status of each server. Currently only French server report can be seen. To complete the services, we developed a "multilingual simultaneous deconverter" (Preedarat 2001), which can handle several deconversions at one time. Users can click on the language versions they want as output, the program will contact these servers at once, thus they don't need to do the deconversions one by one, and they can experience the automatic multilingual generation.

Create UNL Graph Since ordinary users are not able to write UNL graph without being trained, to help users create UNL graph will be an important function to develop. In this page we collect the links to accessible UNL editors, including editor for professional writers or for beginners. We have put a link to our "Basic UNL graph editor" (Preedarat 2001), which is implemented by using a similar XML-UNL format and XSL transformation. The users can manipulate the UNL graph represented in tree-like structure, and save the result in XML format. We also put a link to the "interactive multilingual page" of Spanish Language Centre, here users can manipulate the UNL graph by the options provided, actually users can already generate many sentences based on these examples.

Post-Edit UNL Graph This function is still under development. Our idea is to provide the users the possibility to correct the UNL document after it is deconverted. It provides ordinary users with the ability to correct the faults in the UNL graph and improve the quality of graph.

UNL corpus We collect all the UNL corpora here, and also we are currently working on designing a data base to store these corpora thus to facilitate the further exploitation or calculation. We can finally design an interface to allow users to upload the corpora in different forms, or produce the forms they desire. In Appendix II we show the first statistics we made on the corpus FB2004.

Comments To send comments to the maintainers of the site.

Links & References We collect all the links to UNL Centre, Language Centres, articles, papers, discussion of UNL, and users can trigger the search engines here to find more information about UNL when they want.

3 XML-UNL document

The applications compatible to XML have been increasing a lot and XML can replace HTML as the next norm of a web-based document. And from an XML form, we can further produce other form, exchange or integrate the existing data easily. It would thus be reasonable to XML-ise the UNL document. We would like to propose here an XML form of UNL document as in Appendix III. We created this DTD according to the UNL specification Version 3 Edition 1 (20/02/2002). Based on this DTD, we can create the UNL document in XML form, with an XSL Transformation we can produce the same effect as an UNL browser. Further more, we can easily expand this DTD to enable the XML-UNL document to register all the modifications and corrections on a UNL document, this can be very useful in our post-edition project.

4 Conclusion

We have made the first step in the integration of all the UNL components on a website. Next step is to streamline the procedures between current functions and to include more services.

References

Boitet Ch. (2001) Four technical and organizational keys for handling more languages and improving quality (on demand) in MT, " MT-SUMMIT VIII (2001) ", Proceedings of the Workshop (Towards a Road Map for MT), p.14-21. 18/09/2001

Coch & Chevreau (2001) Interactive Multilingual Generation. Proc. CICLing-2001 (Computational Linguistics and Intelligent Text Proceeding), Mexico, Springer, pp. 239-250.

Sérasset G. and Boitet Ch. (2000) "On UNL as the future "html of the linguistic content" & the reuse of existing NLP components in UNL-related applications with the example of a UNL-French deconverter", COLING 2000, Saarbruecken, Germany 31/07-04/08, p.768-774

Sérasset G. & BOITET Ch. (1999),"UNL-French deconversion as transfer & generation from an interlingua with possible quality enhancement through offline human interaction" MT Summit 99, 13-17 september 1999, Singapore, pp 220-228.

Boitet Ch. (1999) A research perspective on how to democratize machine translation and translation aids aiming at high quality final output, Machine Translation Summit VII (1999), Singapore, 13-17/9/99

Munpyo HONG & Olivier STREITER (1998) "Overcoming the Language Barriers in the Web: The UNL-Approach" , in 11.Jahrestagung der Gesellschaft für Linguistische Datenverarbeitung (GLDV'99), 1999, Frankfurt am Main.

Preedarat JITKUE (2001) Participation au projet SWIIVRE-UNL et première version d'un environnement web de déconversion multilingue et d'éditeur UNL de base, report de stage de Maîtrise Informatique Université Joseph Fourier – Grenoble 28/05-31/08

Appendix I: The resources accessible at each LC for UNL society members

	Enco	Deco	Dico	Introduction of UNL system	Linked by UNLC	Remarks
Arabic	√	√	√	Arabic	√	
Chinese		√		English	√	
French		√				
Indonesian				Indonesian		
Italian		√		Italian	√	
Russian		√	√	English		
Spanish		√		English Spanish	√	Tutorials / Interactive Page / Document Repository
Thai				Thai	√	
UNLC			√	English		UNL specs/ development modules

Appendix II: Some Statistics about FB2004 Corpus

Corpus Name : FB2004
Original Language : English
Other available versions : French, Spanish, Italian, Russian, Hindi, UNL
No. of Sentences : 122
No. of Words : 2799
No. of Relations in UNL: 1519

Part I. The relation count

Relation	Outside scope	In scope	TOTAL	Relation	Outside Scope	In scope	TOTAL
AGT	**66**	**10**	**76**	SEQ	0	0	0
AOJ	**64**	**37**	**101**	FMT	5	0	5
OBJ	**225**	**89**	**314**	FRM	6	3	9
AND	**63**	**120**	**183**	PLF	0	0	0
OR	26	3	29	SRC	2	0	2
BAS	2	2	4	GOL	17	7	24
CAG	0	0	0	PLT	1	0	1
CAO	0	0	0	TO	5	1	6
COB	1	1	2	INS	0	0	0

PTN	4	1	5	**MAN**	**49**	**17**	**66**
BEN	7	5	12	MET	10	3	13
PUR	28	1	29	PER	0	0	0
CNT	22	6	28	QUA	12	5	17
MOD	**263**	**186**	**439**	PLC	17	3	20
NAM	21	15	36	SCN	13	5	18
POF	5	2	7	TMF	2	0	2
POS	17	8	25	TMT	0	1	1
CON	2	0	2	VIA	1	0	1
RSN	1	0	1	DUR	5	4	9
COO	4	2	6	TIM	20	5	25

Total no. of relations: 1519

Remarks

(a) The 6 most frequently used relations are marked in bold type. The result is not surprising, since these relations have either an important or a broad usage. MAN and AGT's usage are frequent though straight forward. Besides its own static verb and copula usage, AOJ also shares part of adjective-noun relation, otherwise the frequency of MOD will be even higher.

(b) AND relation appears much more frequently within a scope, which is not surprising, since scope is used to represent the union of the similar things or ideas, and AND relation links these UW's in te scope.

(c) Some other relations' usage is not very braod, so they didn't appear.

Part II. Attribute count

(1) Time Attribute

.@past 40 / **.@present 114** / **.@future 187**

(2) Aspect Attribute

.@complete 20 / .@progress 13 / .@state 16 / else 0

(3) Reference Attribute

.@generic 9 / **.@def 659** / .@indef 79 / .@not 2 / .@ordinal 8

(4) Focus Attribute

.@entry 530 / .@topic 48 / .@title 21 / else 0

(5) Attitude Attribute

.@exclamation 1 / else 0

(6) Viewpoint Attribute

.@ability 7 / .@obligation 7 / .@possibility 8 / .@should 2 /
.@unexpected-consequence 2 / else 0

(7) Convention Attribute

.@pl 558 / elso 0

Remarks

The original text langue is English, so the frequency of .@pl, .@def, .@indef and time attributes are among the highest. If the original language is one of those isolated languages, such as Thai, Vietnamese, Chinese, which does not provide so much information about definitiveness or time, it might be difficult to use or to decide these attributes. It's not because that the graph authors or enconverters are bad, it's simply because they can't find this information from the text when encoding.

Appendix III. An XML form of UNL document

```
<!DOCTYPE D [
<!ELEMENT D (P+) >
<!ELEMENT P (S+)>
<!ELEMENT S (org,unl,GS+)>
<!ELEMENT org (#CDATA)>
<!ELEMENT unl (#CDATA)>
<!ELEMENT GS (#CDATA)>

<!ATTLIST D dn CDATA
    #REQUIRED
    on CDATA #REQUIRED
    did CDATA #IMPLIED
    dt CDATA #IMPLIED
    mid CDTAT #IMPLIED>
<!ATTLIST P number CDATA
    #REQUIRED>
<!ATTLIST S number CDATA
    #REQUIRED>
<!ATTLIST org lang CDATA
    #REQUIRED
            code CDATA #IMPLIED
    >
<!ATTLIST unl sn CDATA
    #IMPLIED
    pn CDATA #IMPLIED
    rel CDATA #IMPLIED
    dt CDATA #IMPLIED
    mid CDTAT #IMPLIED>
<!ATTLIST GS lang CDATA
    #REQUIRED
    code CDATA #IMPLIED
    sn CDATA #IMPLIED
    pn CDATA #IMPLIED
    rel CDATA #IMPLIED
    dt CDATA #IMPLIED
    mid CDTAT #IMPLIED>

]>

<!-- GS = generated sentence -->
<!-- dn = document name -->
<!-- on = owner name -->
<!-- did = document id -->
<!-- dt = date -->
<!-- mid = mail address -->
<!-- lang = lang tag -->
<!-- code = character code name -->
<!-- sn = system name -->
<!-- pn = post editor name -->
<!-- rel = reliability -->
]>
```