Translating the DEMGOL Etymological Dictionary of Greek Mythology with the BEYTrans Wiki

Y. BEY¹  K. Kageura²  C. Boitet³  F. Marzari⁴

¹,³Laboratoire d’Informatique de Grenoble (LIG), UJF, France
²Graduate School of Education, TODAI, Japan
⁴Centro Antropologia e Mondo Antico, Facolta’ di Lettere e Filosofia, Italy

Wikisym2008, Porto, Portugal
More than 10 years ago

⇒ Toward online collaborative organization (1994)

“I’m a lazy guy. I don’t know how to organize. Let’s just let the community organize”

(Ward Cunningham, Microsoft)
During the same period...

- Inspired by «EuroLang Optimizer» TWB (EuroLang, 1992-95)
- Allow no professional translators to use online computer-aided translation tools (CAT)
- Contribution on the improvement of linguistic resources:
  - Equivalence → Bilingual dictionaries
  - Bilingual segment matching → Translation memories (TM)
During the same period...

- Inspired by «EuroLang Optimizer» TWB (EuroLang, 1992-95)
- Allow no professional translators to use online computer-aided translation tools (CAT)
- Contribution on the improvement of linguistic resources:
  - Equivalence $\rightarrow$ Bilingual dictionaries
  - Bilingual segment matching $\rightarrow$ Translation memories (TM)
During the same period...

- Inspired by «EuroLang Optimizer» TWB (EuroLang, 1992-95)
- Allow no professional translators to use online computer-aided translation tools (CAT)
- Contribution on the improvement of linguistic resources:
  - Equivalence → Bilingual dictionaries
  - Bilingual segment matching → Translation memories (TM)
During the same period...

- Inspired by «EuroLang Optimizer» TWB (EuroLang, 1992-95)
- Allow no professional translators to use online computer-aided translation tools (CAT)
- Contribution on the improvement of linguistic resources:
  - Equivalence $\rightarrow$ Bilingual dictionaries
  - Bilingual segment matching $\rightarrow$ Translation memories (TM)
Surprise: more than 10 years later

- **Wikipedia** encyclopedia
  - \( \approx 200M \) words (>Britannica + Encarta)
  - 99,9% of Wikipedia documents are freely editable online

- Size & languages: more than 2M documents in 200 languages
  - English \( > 500 \, 000 \)
  - German, Japanese, French, ... \( > 100 \, 000 \)
  - Polish, Spanish, ... \( > 50 \, 000 \)

- “Wikipedia is not a technical innovation, it's a social innovation” (Jimmy Wales, chairman of Wikipedia)
Surprise: more than 10 years later

- «Wikipedia» encyclopedia
  - \( \approx 200M \) words (>Britannica + Encarta)
  - 99,9% of Wikipedia documents are freely editable online

- Size & languages: more than 2M documents in 200 languages
  - English > 500 000
  - German, Japanese, French, ... > 100 000
  - Polish, Spanish, ... > 50 000

- "Wikipedia is not a technical innovation, its a social innovation" (Jimmy Wales, chairman of Wikipedia)
Surprise: more than 10 years later

- «Wikipedia» encyclopedia
  - $\approx 200M$ words (>Britannica + Encarta)
  - 99,9% of Wikipedia documents are freely editable online

- Size & languages: more than 2M documents in 200 languages
  - English $> 500,000$
  - German, Japanese, French, ... $> 100,000$
  - Polish, Spanish, ... $> 50,000$

- “Wikipedia is not a technical innovation, its a social innovation” (Jimmy Wales, chairman of Wikipedia)
Volunteer translator communities

- Free software
  - Technical translations (e.g. Traduc): 10 volunteer translators, 2 languages
  - Localization/translation projects (e.g. Mozilla): number (?), 70 langues
  - Web specifications translations (e.g. W3C): 301 volunteer translators, 41 langues

- Cultural and humanitarian translations
  - Reports (e.g. Paxhumana and TeaNotWar): 10 volunteer translators, 4 languages
  - Humain rights: 10 volunteer translators, 2 languages

- Wiki-based translation (e.g. Translationwiki.net)
  - News translation
  - Collaborative translations
Volunteer translator communities

- **Free software**
  - Technical translations (e.g. Traduc): 10 volunteer translators, **2 languages**
  - Localization/translation projects (e.g. Mozilla): **number (?), 70 langues**
  - Web specifications translations (e.g. W3C): 301 volunteer translators, **41 langues**

- **Cultural and humanitarian translations**
  - Reports (e.g. Paxhumana and TeaNotWar) : 10 volunteer translators, **4 languages**
  - Humain rights : 10 volunteer translators, **2 languages**

- **Wiki-based translation** (e.g. Translationwiki.net)
  - News translation
  - Collaborative translations
Volunteer translator communities

- Free software
  - Technical translations (e.g. Traduc): 10 volunteer translators, \textbf{2 languages}
  - Localization/translation projects (e.g. Mozilla): \textit{number (?)}, \textbf{70 languages}
  - Web specifications translations (e.g. W3C): 301 volunteer translators, \textbf{41 languages}

- Cultural and humanitarian translations
  - Reports (e.g. Paxhumana and TeaNotWar): 10 volunteer translators, \textbf{4 languages}
  - Human rights: 10 volunteer translators, \textbf{2 languages}

- Wiki-based translation (e.g. Translationwiki.net)
  - News translation
  - Collaborative translations
Abstract

The Resource Description Framework RDF allows users to describe both Web documents and concepts from the real world—people, organisations, topics, things—in a computer-processable way. Publishing such descriptions on the Web creates the Semantic Web. URIs (Uniform Resource Identifiers) are very important, providing both the core of the framework itself and the link between RDF and the Web. This document presents guidelines for their effective use. It discusses two strategies, called 303 URIs and hash URIs. It gives pointers to several Web sites that use these solutions, and briefly discusses why several other proposals have problems.

Status of this document

This section describes the status of this document at the time of its publication. Other documents may supersede this document. A list of current W3C publications and the latest revision of this technical report can be found.

Résumé

Le cadre de description de ressource (Resource Description Framework), ou RDF, permet aux utilisateurs de décrire des documents web ainsi que des concepts du monde réel — personnes, organisations, thèmes, choses — de façon automatique. La publication de telles descriptions sur le Web crée le Web sémantique (Semantic Web). Les adresses URI (Uniform Resource Identifiers) sont très importantes, en fournissant à la fois le cœur de l’environnement (framework) lui-même et le lien entre RDF et le Web. Ce document présente des directives pour les employer efficacement. Il explique deux stratégies dites d’adresses URI avec 303 (303 URIs) et d’adresses URI avec dièse (hash URIs). Il fournit des pointeurs vers plusieurs sites web qui utilisent ces solutions et explique brièvement pourquoi d’autres propositions recèlent des problèmes.

Statut de ce document

Y. BEY, K. Kageura, C. Boitet, F. Marzari
Outline

1. Background
2. Aims
3. Scenario
4. Design: Wiki-based collaborative CAT environment
5. Experimentation
6. Conclusion
Alternative information dissemination is spreading in multilingual form. e.g. Human rights/Sports/Area studies

- Domain translators know each other
- They loosely communicate to avoid duplication
- Potential translators
Alternative information dissemination is spreading in multilingual form. e.g. Human rights/Sports/Area studies

Domain translators know each other

They loosely communicate to avoid duplication

Potential translators
Alternative information dissemination is spreading in multilingual form. e.g. Human rights/Sports/Area studies

Domain translators know each other

They loosely communicate to avoid duplication

Potential translators
Alternative information dissemination is spreading in multilingual form. e.g. Human rights/Sports/Area studies

Domain translators know each other

They loosely communicate to avoid duplication

Potential translators
Levels of translation differ.

- Use mostly standalone electric dictionary and frequently use Google
- No MT, no translation aid system as such
- Not willing to make much efforts
  - OK to register top-page URL for server
  - OK to use `<a>` tag with keyword ”original text” in each text
Levels of translation differ.

Use mostly standalone electric dictionary and frequently use Google

No MT, no translation aid system as such

Not willing to make much efforts

- OK to register top-page URL for server
- OK to use `<a>` tag with keyword "original text" in each text
Levels of translation differ.
Use mostly standalone electric dictionary and frequently use Google
No MT, no translation aid system as such
Not willing to make much efforts
- OK to register top-page URL for server
- OK to use <a> tag with keyword ”original text” in each text
Levels of translation differ.
Use mostly standalone electric dictionary and frequently use Google
No MT, no translation aid system as such
Not willing to make much efforts
- OK to register top-page URL for server
- OK to use `<a>` tag with keyword ”original text” in each text
Aims

Enhancement of translation environments and linguistic functionalities

- Collaborative translation environment
  - Recycling relevant translated texts
  - Integrated editor / linguistic functionalities

- Proactive linguistic suggestions
  - Community translation memories
  - Linguistic resources (dictionaries, glossaries, ...)
  - Machine translation suggestions
Enhancement of translation environments and linguistic functionalities

- Collaborative translation environment
  - Recycling relevant translated texts
  - Integrated editor / linguistic functionalities

- Proactive linguistic suggestions
  - Community translation memories
  - Linguistic resources (dictionaries, glossaries,...)
  - Machine translation suggestions
Enhancement of translation environments and linguistic functionalities

- Collaborative translation environment
  - Recycling relevant translated texts
  - Integrated editor / linguistic functionalities

- Proactive linguistic suggestions
  - Community translation memories
  - Linguistic resources (dictionaries, glossaries,...)
  - Machine translation suggestions
Volunteer translation scenario

- Background
- Aims
- Scenario
- Design: Wiki-based collaborative CAT environment
- Experimentation
- Conclusion

Linguistic tools

Individual translation environments

Who does what?

Document format:
- Help files
- Interface files
- Control file

CVS

Discussion list
XWiki: Web architecture for collaboration

- **Background**
- **Aims**
- **Scenario**
- **Design: Wiki-based collaborative CAT environment**
- **Experimentation**
- **Conclusion**

Y. BEY, K. Kageura, C. Boitet, F. Marzari
BEYTrans: design

**Importation**

- **Source**
  - PDF
  - Doc
  - HTML

**Collaborative translation**

- Multilingual Segmentation
- Historic
- Multilingual research
- Doc.
- TM
- Dico

**Dissemination**

- **Target**
  - HTML
  - HTML
  - HTML

BEYTrans: design

Y. BEY, K. Kageura, C. Boitet, F. Marzari
Translation units management

- One document is a set of translation units
- Translation units are stored in TMX XML structure (LISA) = translation companion (TC)
- Target documents are generated from the translation companion
- Translation companion is the basic unit of the editor “BT-Edit”

TCs sets of translation communities = translation memories
Translation units management

- One document is a set of translation units
- Translation units are stored in TMX XML structure (LISA) = translation companion (TC)
- Target documents are generated from the translation companion
- Translation companion is the basic unit of the editor “BT-Edit”

TCs sets of translation communities = translation memories
Translation units management

- One document is a set of translation units
- Translation units are stored in TMX XML structure (LISA) = translation companion (TC)
- Target documents are generated from the translation companion
- Translation companion is the basic unit of the editor “BT-Edit”

TCs sets of translation communities = translation memories
Translation units management

- One document is a set of translation units
- Translation units are stored in TMX XML structure (LISA) = translation companion (TC)
- Target documents are generated from the translation companion
- Translation companion is the basic unit of the editor “BT-Edit”

TCs sets of translation communities = translation memories
Translation units management

- One document is a set of translation units
- Translation units are stored in TMX XML structure (LISA) = translation companion (TC)
- Target documents are generated from the translation companion
- Translation companion is the basic unit of the editor “BT-Edit”

TCs sets of translation communities = translation memories
BT-Edit: translation companion concept

Translation companion (TC)

btu_11
btu_12
btu_1n
btu_21
btu_22
btu_2n

New language

step (1)

step (2)

BT-Edit
BT-Edit: proactive TM suggestions

- Fuzzy matching method
- Naive algorithm (Wagner & Fisher, 1974)
- Asynchronous suggestions
BT-Edit: synchronization and other functionalities

Dictionary suggestions

Source TU (it): Tokenization, etc.

Target TU: translation (fr)

Dictionary updating
DEMGOLO: introduction

- OnLine Etymological Dictionary of the Greek Mythology
- Sub-project translations: 1200 documents (Italian into French and Spanish)
- 600/1200 have been translated before using BEYTrans
- Translation into French: collaboration with Trieste University (Trieste, Italy) and Stendhal University (Grenoble, France)
DEMGOL: experimentation protocol

Two sets of documents (or notices):

- Set $E_1$: «L» documents (60 documents) - Human translation
- Set $E_2$: «M» documents (30 documents) - BEYTrans

- Translation duration
- Addition of lexical translation
- Online experimentation: one volunteer translator (no professional) at Trieste University (Ms Marzari) and TODAI (me)
Animale dell'India, feroce, antropofago, di colore fulvo o rossiccio. La grafia martiçwvra prevale nelle fonti greche; in latino troviamo per lo più manticora, maschile, che diventerà femminile nelle fonti più tardive e medievali. È descritto dettagliatamente da Ctesia di Cnido (V-IV sec. a.e.v.) negli Indikà (F 45 14-15: riassunto nella Biblioteca di Foazio): vive in India, ha il viso, gli occhi e le orecchie simili all'uomo, le zampe di leone e una coda di scorpione in grado di scagliare come frecce gli aculei che vi crescono.Eliano, Nat. an. 4, 21, paragona curiosamente il suo modo di combattere a quello dei Saci, popolazione degli Sciti, noti come abilissimi arcieri a cavallo. Assai notevole è la triplice fila di denti (trivsticioi de; ojdoynhte") per ogni mascella, che a Plinio (Nat. hist. 8, 30,75) ricordano un pettine (pectinatim), e che ben si accordano alla fama di fiera "divoratrice di uomini": questo sarebbe il significato del suo nome in lingua indiana. Aristotele (Hist. an. 2, 1, 501, 24-b1), molto cauto nei confronti della fonte (eij dei§ pisteu§sai Ktesiva/), la cita comunque sottolineando la salvatichezza e pericolosità dell'animałe (a[grion kai; ajnqrwpofavgon). Pausania (9, 21, 4-5) cerca di spiegare (razionalizzando) che dovrebbe trattarsi della tigre, e attribuisce i tratti più fantasiosi della creatura al grande timore che ispira. È veloce come il cervo, e la sua voce risuona come flauto e tromba al contempo....
DEMGOL: translation without BEYTrans

Evaluating translation duration without BEYTrans "set E1"
DEMGOL: translation using BEYTrans

Evaluating translation duration with BEYTrans "set E2"
(Systran Babel Fich MT + TM)
DEM Gol: translation evaluation

- 174 additions (lexical translation)
- Experimentation: 2 weeks
- Evaluation of 30 notices ($E_1$ and $E_2$)

<table>
<thead>
<tr>
<th>30 notices</th>
<th>BEYTrans</th>
<th>Lexical translation</th>
<th>Total duration</th>
<th>Real duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_1$ (Humain translation)</td>
<td>-</td>
<td>0</td>
<td>3h33</td>
<td>3h33</td>
</tr>
<tr>
<td>$E_2$ (BEYTrans)</td>
<td>+</td>
<td>2h54</td>
<td>5h08</td>
<td>2h13</td>
</tr>
</tbody>
</table>
Conclusion

- Collaborative translation: facilitating the dissemination of free multilingual content on the Web by switching between reading and editing
- Multilingual online editor: integrated proactive and enhanced linguistic functionalities
- Asynchronous translation

- Community translation evaluation
- Localization of free and commercial software (Winsoft - Adobe software)
www.javaling3.imag.fr/beytrans/