Coedition of texts and UNL graphs to share text revision across languages and improve MT a posteriori

International Conference on Universal Knowledge and Language (ICUKL2002), Goa, 25-29 November 2002

Christian BOITET & TSAI Wang-Ju
GETA, CLIPS, IMAG, Grenoble

Christian.Boitet@imag.fr, Wang-Ju.Tsai@imag.fr
Plan

- Motivation: to achieve quality MT in a multi-source & multi-target environment
- Why a pivot architecture and why UNL?
- The prototype under construction:
  - principle: divide the correspondence (graph↔text) in 3 parts.
  - interface: necessity of modes, from «naive» to «expert»
- A problem: computing the 2nd correspondence
  - UNL-tree ↔ morpho-syntactic lattice
- A «static» demo on one simple sentence
Motivation: build multisource & multitarget "quality MT"

> intrinsic limits of the classic procedure

30 standard pages of 250 words, in hours (trans+rev):

- 40*n by THum, \((12/3+12+12/3)\)*n = 20 n by THAM
  \((6=,12\approx,12\neq)\)
- 8*n with THAM + MAT-R, and \((30/3)\)*n = 10*n if MAT-R only

⇒ (30/3)*n = 10 or even (30/2)*n = 15, independently of n?
Basic Idea: to edit the « pivot » indirectly, by editing the text

- examples of successful « co-edition »
  Ambassador (old), MULTIMETEO

- 2 problems:
  - To build the correspondence
    abstract structure ↔ concrete text
    Without building new and costly resources
  - To construct a really usable interface
Why use UNL as the pivot?

- brief reminder
  - UNL is
    - a project
    - an artificial language
    - a format of a multilingual document
  - the UNL language has unique features
    even if it is perfectible!
Language: a simple UNL graph
(a) linear form, read by computer

Ronaldo has headed the ball into the left corner of the goal

agt(score(icl>event,agt>human,fld>sport).@entry.@past.@complete, Ronaldo(icl>proper noun))
obj(score(icl>event,agt>human,fld>sport).@entry.@past.@complete, goal(icl>thing))
ins(score(icl>event,agt>human,fld>sport).@entry.@past.@complete, head(pof>body).@def)
plt(score(icl>event,agt>human,fld>sport).@entry.@past.@complete, corner(icl>thing).@def)
pos(corner(icl>thing).@def), goal(icl>thing))
pos(head(pof>body).@def, Ronaldo(icl>proper noun))
mod(corner(icl>thing).@def), left(aoj<thing))
Language: a simple UNL graph
(b) linear form, with colors

Ronaldo has headed the ball into the left corner of the goal

- `agt(score(icl>event,agt>human,fld>sport).@entry.@past.@complete, Ronaldo(icl>proper noun))`
- `obj(score(icl>event,agt>human,fld>sport).@entry.@past.@complete, goal(icl>thing))`
- `ins(score(icl>event,agt>human,fld>sport).@entry.@past.@complete, head(pof>body).@def)`
- `plt(score(icl>event,agt>human,fld>sport).@entry.@past.@complete, corner(icl>thing).@def)`
- `pos(corner(icl>thing).@def, goal(icl>thing))`
- `pos(head(pof>body).@def, Ronaldo(icl>proper noun))`
- `mod(corner(icl>thing).@def, left(aoj<thing))`

- Is this UNL graph OK? Let us draw it.
Language: a simple UNL graph
(c) graphical form

- Ronaldo has headed the ball into the left corner of it

score(icl>event,agt>human,fld>sport).
@entry.@past.@complete

agt

Ronaldo
(icl> proper noun)

obj

head(pof>body).@def

ins

corner(icl>thing).@def

plt

pos

goal(icl>thing)

pos

left(aoj<thing)

mod

•==> Ronaldo has headed the ball into the left corner of it
Language : a simple UNL graph (c) graphical form, colored

• Ronaldo has headed the ball into the left corner of it

score(icl>event,agt>human,fld>sport).
@entry.@past.@complete

agt

Ronaldo (icl>proper noun)

obj

head(pof>body).@def

ins

goal(icl>thing)

plt

corner(icl>thing).@def

mod

left(aoj<thing)
Language: a simple UNL graph
(d) graphical form, corrected!

- Ronaldo has headed the ball into the left corner of the goal <= OK!
Language : a simple UNL graph (e) linear form, corrected

Ronaldo has headed the ball into the left corner of the goal

\texttt{agt} (\texttt{score} (icl>event,agt>human,fld>sport).@entry.@past.@complete, Ronaldo (icl>proper noun))

\texttt{obj} (\texttt{score} (icl>event,agt>human,fld>sport).@entry.@past.@complete, goal (icl>thing))

\texttt{ins} (\texttt{score} (icl>event,agt>human,fld>sport).@entry.@past.@complete, head (pof>body).@def)

\texttt{plt} (\texttt{score} (icl>event,agt>human,fld>sport).@entry.@past.@complete, corner (icl>abstract thing).@def)

\texttt{pos} (corner (icl>thing).@def, goal (icl>concrete thing))

\texttt{pos} (head (pof>body).@def, Ronaldo (icl>proper noun))

\texttt{mod} (corner (icl>thing).@def, left (aoj<thing))
Isaac sees that an apple falls and he explains it.

\[
\text{agt}([\text{explain} \mid \text{do}] \@ \text{entry}, \text{Isaac} \mid \text{proper noun})
\]
\[
\text{obj}([\text{explain} \mid \text{do}] \@ \text{entry}, :01)
\]
\[
\text{obj}:01 ([\text{fall} \mid \text{occur}] \@ \text{entry}, \text{apple})
\]
\[
\text{and} ([\text{explain} \mid \text{do}] \@ \text{entry}, \text{see} \mid \text{do})
\]
\[
\text{agt}([\text{see} \mid \text{do}], \text{Isaac} \mid \text{proper noun})
\]
\[
\text{obj}([\text{see} \mid \text{do}], :01)
\]
What is the UNL language?

- Small ongoing controversy…
- A way to look at a UNL (hyper)graph:
  it corresponds to an utterance \( U-L \) in language \( L \)
  by representing the abstract structure
  - of an equivalent English utterance \( U-E \)
  - as « viewed from \( L \) »

  ==> the semantic attributes not necessarily expressed in \( L \)
  may be absent: frequent underspecification
  - aspect coming from French,
  - determination or number coming from Japanese,
  - etc.
The reasons of using UNL in MT

(and it can be used in many other ways!)

- Technical success of pivot MT does exist
  (ATLAS, PIVOT, ULTRA, KANT)
- UNL derives from the pivot of ATLAS-II (Fujitsu)
  & was designed by the same author (H. Uchida)
- Possible quality & coverage:
  ATLAS-II has been the best E ↔ J system since > 10 years
  It had 586,000 terms in each dictionary in Sept. 2001
- Gives certainly a lesser asymptotic quality than transfer of/via « multi-level structure»
  BUT UNL can be « coedited » from any source language
<HTML><HEAD><TITLE>Example 1 El/UNL</TITLE></HEAD><BODY>

[S:1]
{org:el}I ran in the park yesterday. {/org}

{cn dtime=20020130-2030, deco=man}
我昨天在公園裡跑步 {/cn}

{de dtime=20020130-2035, deco=man}
Ich lief gestern im Park. {/de}

{es dtime=20020130-2031, deco=UNL-SP}
Yo corri ayer en el parque. {/es}

{fr dtime=20020131-0805, deco=UNL-FR}
J’ai couru dans le parc hier. {/fr}[S]

[S:2]
{org:el}My dog barked at me. {/org}

{cn dtime=20020130-2030, deco=man}
我昨天在公園裡跑步 {/cn}

{de dtime=20020130-2036, deco=man}
Mein Hund bellte zu mir. {/de}

{fr dtime=20020131-0806, deco=UNL-FR}
Mon chien aboya pour moi. [/S] [/P][/D]

</BODY></HTML>
The equivalent UNL-xml format

☞ As simple as UNL-html
☞ Open to all XML-related tools

我昨天在公園裡跑步
</unl:org>
</unl:D>

今日は公園に走っています。</unl:GS>
</unl:D>

Ich lief in den Park gestern. </unl:GS>
</unl:D>

Yo corri ayer en el parque.</unl:GS>
</unl:D>

J’ai couru dans le parc hier.</unl:GS>
</unl:D>
Output of the UNL-viewer and display in a browser

<table>
<thead>
<tr>
<th>Output from the viewer (for French)</th>
<th>Display (in IE or Netscape or…)</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>&lt;HTML&gt;&lt;HEAD&gt;&lt;TITLE&gt;</code></td>
<td><strong>Example 1 El/UNL</strong></td>
</tr>
<tr>
<td><code>Example 1 El/UNL</code></td>
<td>J’ai couru dans le parc hier.</td>
</tr>
<tr>
<td><code>&lt;/TITLE&gt;&lt;/HEAD&gt;&lt;BODY&gt;</code></td>
<td>Mon chien aboya pour moi.</td>
</tr>
<tr>
<td><code>J’ai couru dans le parc hier.</code></td>
<td></td>
</tr>
<tr>
<td><code>&lt;p&gt;</code></td>
<td></td>
</tr>
<tr>
<td><code>Mon chien aboya pour moi.</code></td>
<td></td>
</tr>
<tr>
<td><code>&lt;/BODY&gt;&lt;/HTML&gt;</code></td>
<td></td>
</tr>
</tbody>
</table>
Scenario for coedition

- User reads a multilingual document in lang. Li (IE)
- User wishes to correct some errors in Li
- User switches to the coedition environment
- User’s corrections will be executed
  - later on the text
  - immediately on the graph
- User asks for deconversion into Li
- User iterates corrections if not satisfied, asks for deconversion into L1…Ln when OK
- User returns to reading mode
Example of « FB2004 »
(cited in the article)

- FB2004 = Forum Barcelona 2004
- project of « UNL usage normalisation »
- Spanish, Italian, Russian, French, Hindi
- 6-8 months in year 2001
Example

Une cité retrouvera une zone côtière après un forum

A city will retrieve a coastal zone after a forum

→ La cité retrouvera une zone côtière après le Forum

The city will retrieve a coastal zone after the Forum
Initial UNL graph (textual form)

\[
\begin{align*}
\text{agt} & \text{ (retrieve(icl>do).@entry.@future.@complete, city(icl>entity))} \\
\text{tim} & \text{ (retrieve(icl>do).@entry.@future.@complete, after)} \\
\text{obj} & \text{ (retrieve(icl>do).@entry.@future.@complete, zone(icl>place).@indef)} \\
\text{obj} & \text{ (after, forum)} \\
\text{mod} & \text{ (zone(icl>place).@indef, coastal(aoj<thing)))}
\end{align*}
\]
A city will retrieve a coastal zone after a forum.
• A/The? city will retrieve(?) a coastal zone after ¿a/the? forum(?)
Modifications needed

- add ".@def" on the nodes containing "city", "forum".
- add "icl>proper name" as a restriction of the UW "forum".
- Replace "retrieve" by "recover"
Transformed UNL graph

\[ \text{agt}(\text{recover}(\text{icl}>\text{do}).@\text{entry}.@\text{future}.@\text{complete},\text{city}(\text{icl}>\text{entity}).@\text{def}) \]
\[ \text{tim}(\text{recover}(\text{icl}>\text{do}).@\text{entry}.@\text{future}.@\text{complete}, \text{after}) \]
\[ \text{obj}(\text{recover}(\text{icl}>\text{do}).@\text{entry}.@\text{future}.@\text{complete},\text{zone}(\text{icl}>\text{place}).@\text{indef}) \]
\[ \text{obj}(\text{after}, \text{forum}(\text{icl}>\text{proper noun}).@\text{def}) \]
\[ \text{mod}(\text{zone}(\text{icl}>\text{place}).@\text{indef}, \text{coastal}(\text{aoj}<\text{thing})) \]
Transformed UNL graph

- The city will recover a coastal zone after the Forum.
How to do it? First, the interface

- Accessible directly from a browser
- Normal mode
  - without seeing the intermediate structure(s)
- Expert mode (for hackers, fans, young people…)
  — with direct manipulation of structures
After a Forum, a city will recover a coastal zone.

Ciudad recobrará una zona de costal después Foro.

La ciudad habrá recobrado una zona costal después el Foro.

Città ricupererà una zona costiera dopo Forum.

La città ha ricupererà una zona costiera dopo il Forum.

フォーラムの後で、都市は沿岸水域を取り出すことを持っている。

在博覽會之後，城市 將獲得一片海岸域

在博覽會之後，城市将獲得一片海岸域
Une cité retrouvera une zone côtière après un forum.
After a Forum, a city will retrieve a coastal zone.

Ciudad recobrará una zona de costal después Foro.

Città ricupererà una zonacostiera dopo Forum.

フォーラムの後で, 都市は沿岸水域を取り出すことを持っている。

在博覽會之後,城市將獲得一片海岸域
Principles of coedition

괘 In general, it is impossible to deduce a modification on the graph from a modification on the text

For example, to replace "un" ("a") by "le" ("the") doesn’t imply that the following noun is determined (\@def)

"il aime la montagne" = "he likes mountains"

괘 The revision is not made directly by modifying the text, but by using a system of menus

 OnPropertyChanged

괘 The menu items have a « language side » and a hidden « UNL side »
Construction of the correspondence

- Division in three sub-correspondences
  - UNL graph
    - UNL tree obtained from unfolding the graph
  - SMS lattice or morpho-syntactic automaton
  - Text

- For =2=, computation of the links by using
  1 or 2 dictionaries $L0 \leftrightarrow$English or UNL
  Known relations between $L0$ attributes and UNL attributes
  Maximal projectivity principle (tree rotation)
Demo of mockup on Internet Explorer…

…is not available here, sorry, but we can see the idea with images from the demo.

 Situation
 User is in (co)edit mode and has chosen a sentence

 What will be illustrated
 IE screen shots (from March obsolete mockup!)
 Given UNL-graph
 Computed morphosyntactic lattice
 Steps in the construction of the correspondence
Une cité retrouvera une zone côtière après un forum.

Original text: Une cité retrouvera une zone côtière après un forum.
Possible Modifications: 
Second Deconversion: 
Manual Insertion: La cité retrouvera une zone côtière après le forum.

Languages:
- English: After a Forum, a city will retrieve a coastal zone.
- Spanish: Ciudad recuperará una zona de costas después del Foro.
- Italian: Città recupererà una zona costiera dopo il Forum.
- Japanese: フォーラムの後で、都市は海岸域を取り出す。
- Chinese: 在論壇之後，城市將獲得一片海岸域。
A cidade recuperará uma zona costeira após o forum.

Graph: lemma lattice
Show Graph  Decoersion  Find Lemma  Find Correspondence  Save Graph

Original text: 
Possible Modifications: 
Second Decoersion: 
Manual Insertion: 

English: After the Forum, the city will have retrieved a coastal zone.
Spanish: Ciudad recuperará una zona costera después del Foro.
Italian: Una città recupererà una zona costiera dopo Forum.
Japanese: フォーラムの後に、都市は沿岸水域を取り出す。
Chinese: 在議會之後，城市將獲得一片海水域。
Input UNL-graph (here a tree)
UNL-French tree + French lemmas

QuickTime™ et un décompresseur TIFF (non compressé) sont requis pour visionner cette image.
UNL-French disambiguated tree

"zone-N" remains, but "diviser-V" disappears

QuickTime™ et un décompresseur TIFF (non compressé) sont requis pour visionner cette image.
On-line free French lemmatizer: PILAF
Morphosyntactic lattice

1. Output of French morphological analysis
2. Enrichment by FRE-ENG dictionary

QuickTime™ et un décompresseur BMP sont requis pour visualiser cette image.
3. Enrich the UNL-tree (UNL/ENG-FR) & link
SMS ↔ UNL-tree attributive links

4. Link UNL & French attribute values or places

QuickTime™ et un décompresseur TIFF (non compressé) sont requis pour visionner cette image.
New UNL-tree ↔ Chinese text

Alignment of tree-text is not always projective
New UNL-tree ↔ Spanish text

Here, 2 errors remain (de costal, el Foro)  
===> need for feed-back to the deconversion sites

QuickTime™ et un décompresseur TIFF (non compressé) sont requis pour visionner cette image.
Conclusion

« Shared revision » seems possible
No claim that all corrections could be done/shared this way!
Needs further prototyping (1 more Ph.D. year + students)
Idea of small experiment formed at ICUKL-2002

« Coedition » is a concept unifying
Machine Translation
and
Multilingual Generation

Thanks a lot to our guests & colleagues for
the high scientific interest of the conference
the excellent organization
and the fabulous hospitality