Macau, March 11th

XI UNL School

Day #1

Day #1

- Welcome
- Context
  - The UNL
  - The UNL System
  - The UNDL Foundation Road Map
  - The UNL^{versity}
- Corpus
- Dictionaries
- X-bar

Welcome

Participants

- Ana Luísa Varani Leal
- Angela Manukian
- Chittaphone Chanylilath
- Chunlei Yang
- Dang Quang Vinh
- Jeanette Tan Yi Wen
- Kim Sokphyrum
- Lilit Paremuzyan
- Muhammad Zulhelmy Bin Mohd Rosman
- Ronaldo Martins
- Somdev Kar
- Suos Samak
- Teng Wei Min
- Violeta Liu
- Viraj Ramyanath Karunananda
- Viraj Welgama

The UNL

The Universal Networking Language (UNL)

UNL

translation
knowledge representation

☼1996

UNITED NATIONS
UNIVERSITY

Universal
Networking
Language

Commitments

1. The UNL must represent information

   The UNL must represent "what was meant" (and not "what was said").

2. The UNL must be a language for computers

   The UNL must be computable.

3. The UNL must be self-sufficient

   The UNL representation must not depend on any implicit knowledge.

4. The UNL must be general-purpose

   The UNL must not be bound to translation.

5. The UNL must be independent from any particular natural language

   As a language of the UN, the UNL must be neutral.

Properties

- Non-Ambiguity
  - the boys saw the girl with the telescope
  - [[the boys] [[[saw(icl>perceive) [the girl]] [with the telescope]]]]
- Non-Redundancy
  - Peter killed Mary ≅ Mary was killed by Peter ≅ Peter caused Mary to die
- Compositionality
  - John devoured thousands of books = John read many books
- Declarativeness
  - Can you pass me the salt? = (you pass the salt to me).@request.@polite
- Completeness
  - The monkey took the banana and ate it
  - The monkey_i took the banana_j and the monkey_i ate the banana_j

Structure

Information can be represented by semantic networks made of three different types of discrete semantic entities:

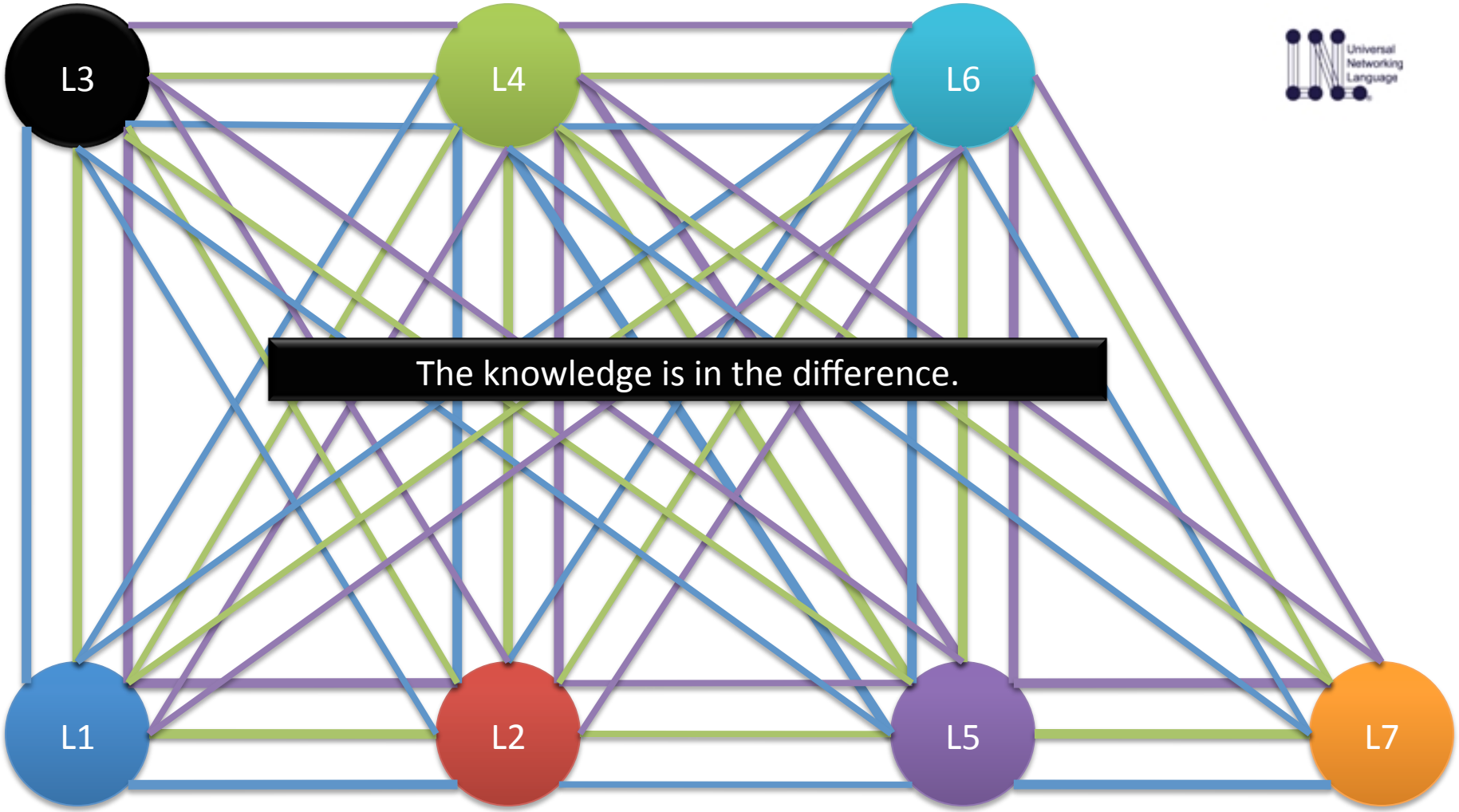
CONCEPTS = Universal Words (UWs)

CONCEPT SPECIFIERS = Universal Attributes

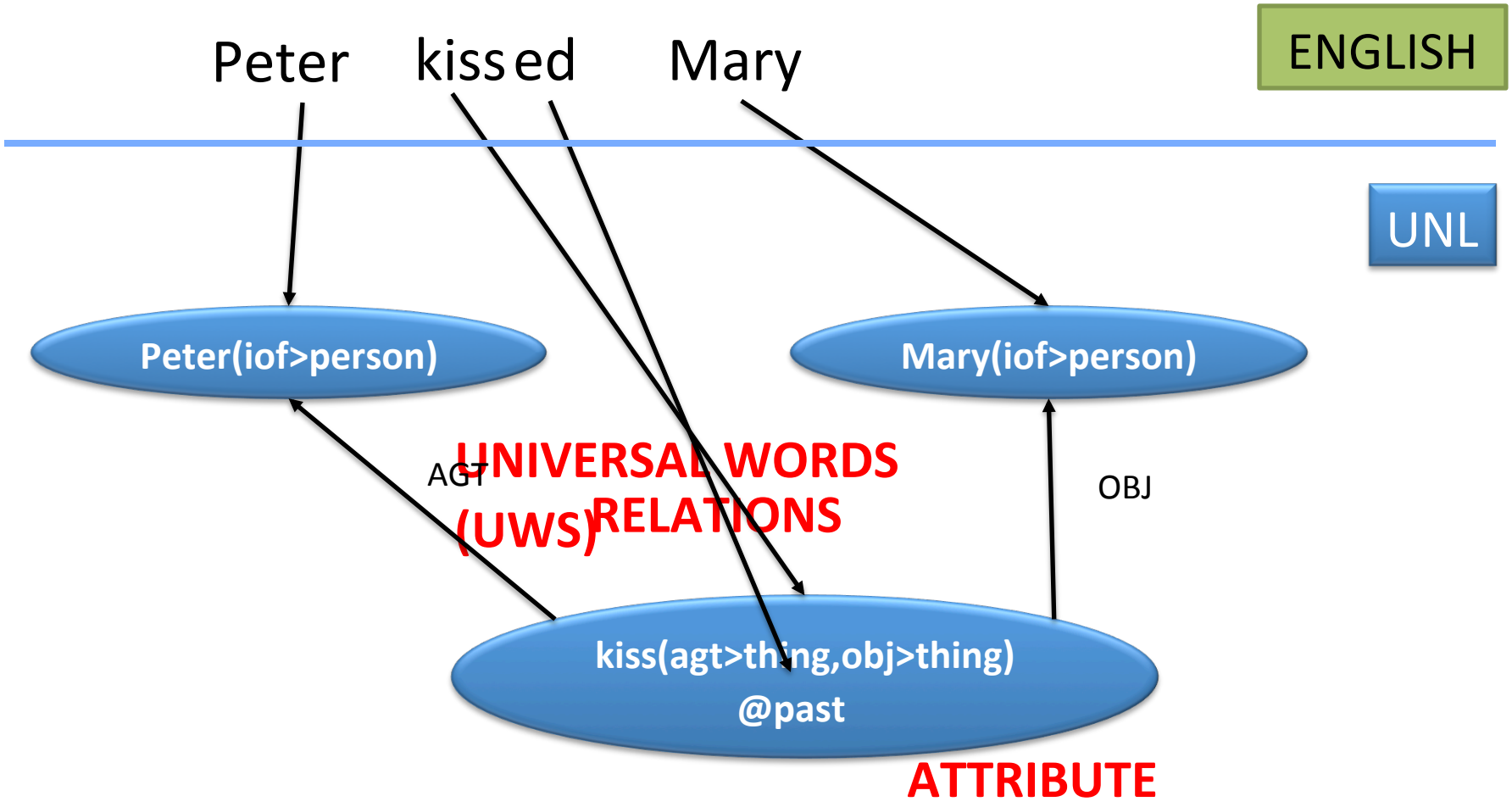
RELATIONS BETWEEN CONCEPTS = Universal Relations

UNDL FOUNDATION

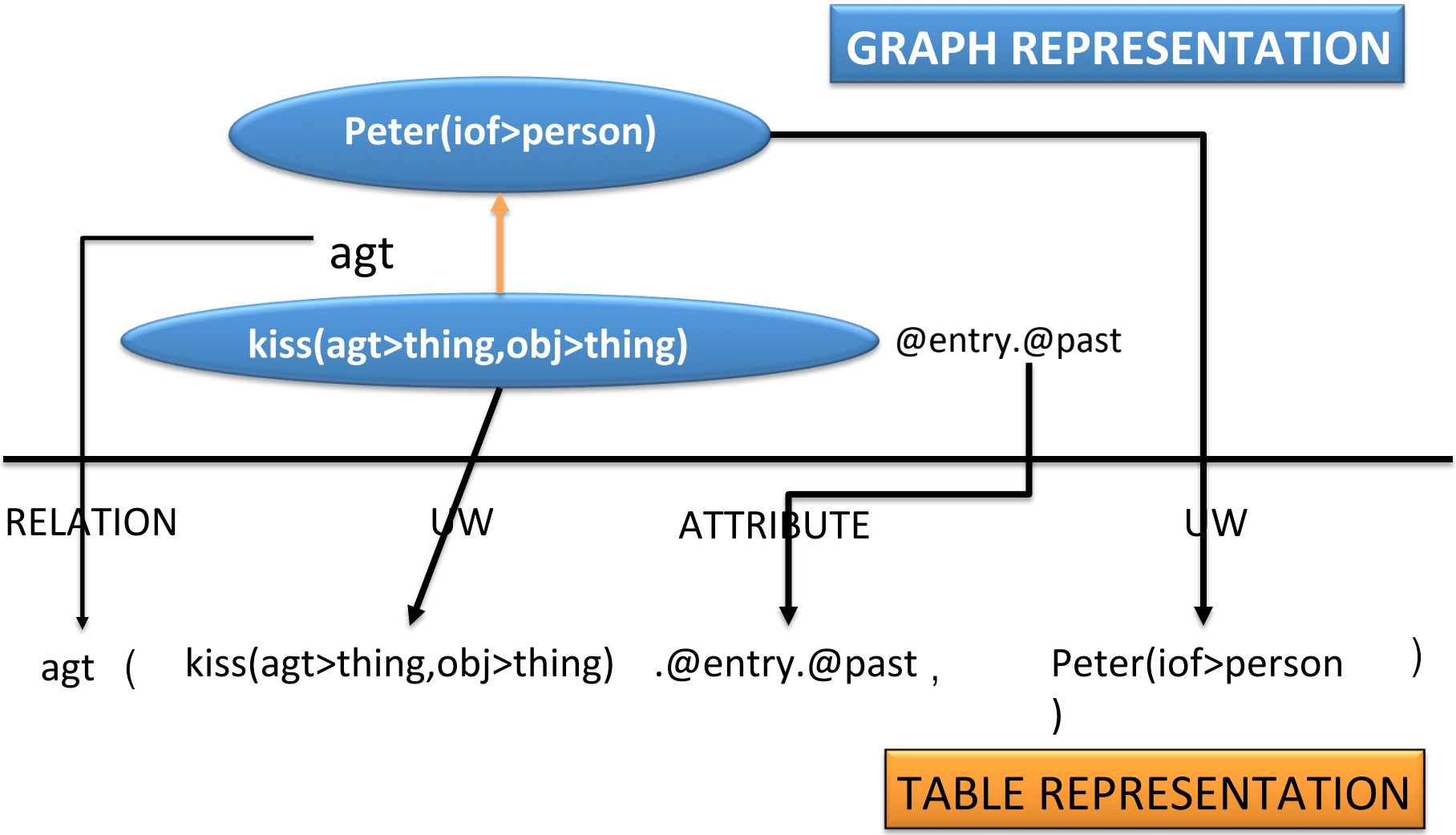
The Universal NETWORKING Language



Natural Language-to-UNL (UNL-ization)



Syntax of UNL



UNL document

[D]

[S:01]

{org:en}

Peter kissed Mary.

{/org}

{unl}

agt(kiss(agt>thing,obj>thing).@entry.@past, Peter(iof>person))

obj(kiss(agt>thing,obj>thing).@entry.@past, Mary(iof>person))

{/unl}

[/S]

[/D]

UNL (http://anydomain/anydocument.unl)

<unl>
agt(kiss(agt>thing,obj>thing).@entry.@past, Peter(iof>person))
obj(kiss(agt>thing,obj>thing).@entry.@past, Mary(iof>person))
</unl>

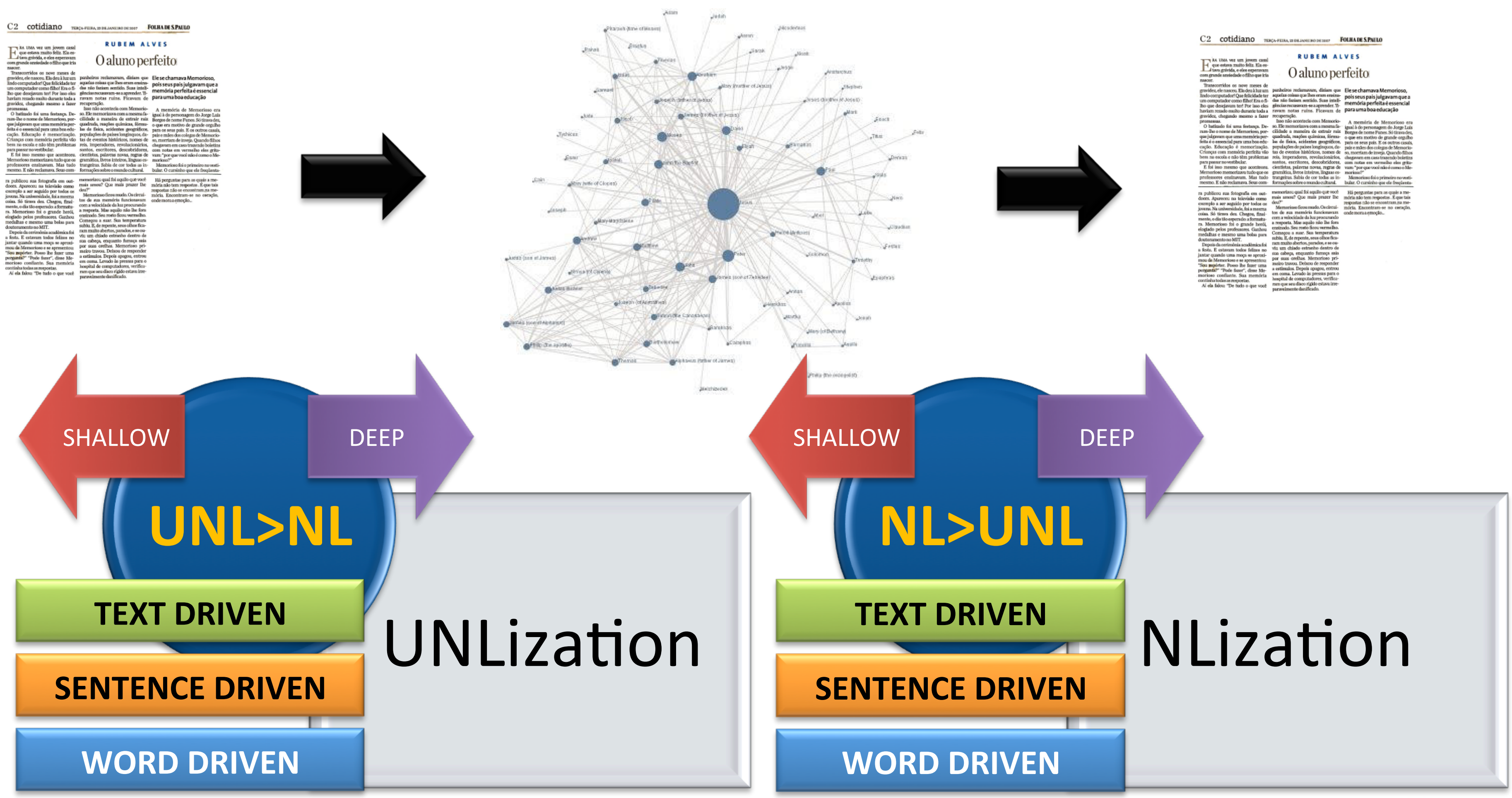
Peter kissed Mary.

Pierre a embrassé Marie.

Pedro besó a María.

The UNL System

The UNL System



UNDL FOUNDATION

The UNDL Foundation Road Map

FRONT-END APPLICATIONS

keys

information retrieval and extraction

tut

summarization and simplification

lily

translation

BACK-END APPLICATIONS

UNL dev

UNL editor

IAN

SEAN

EUGENE

NORMA

EDGES

RESOURCES

UNL arium

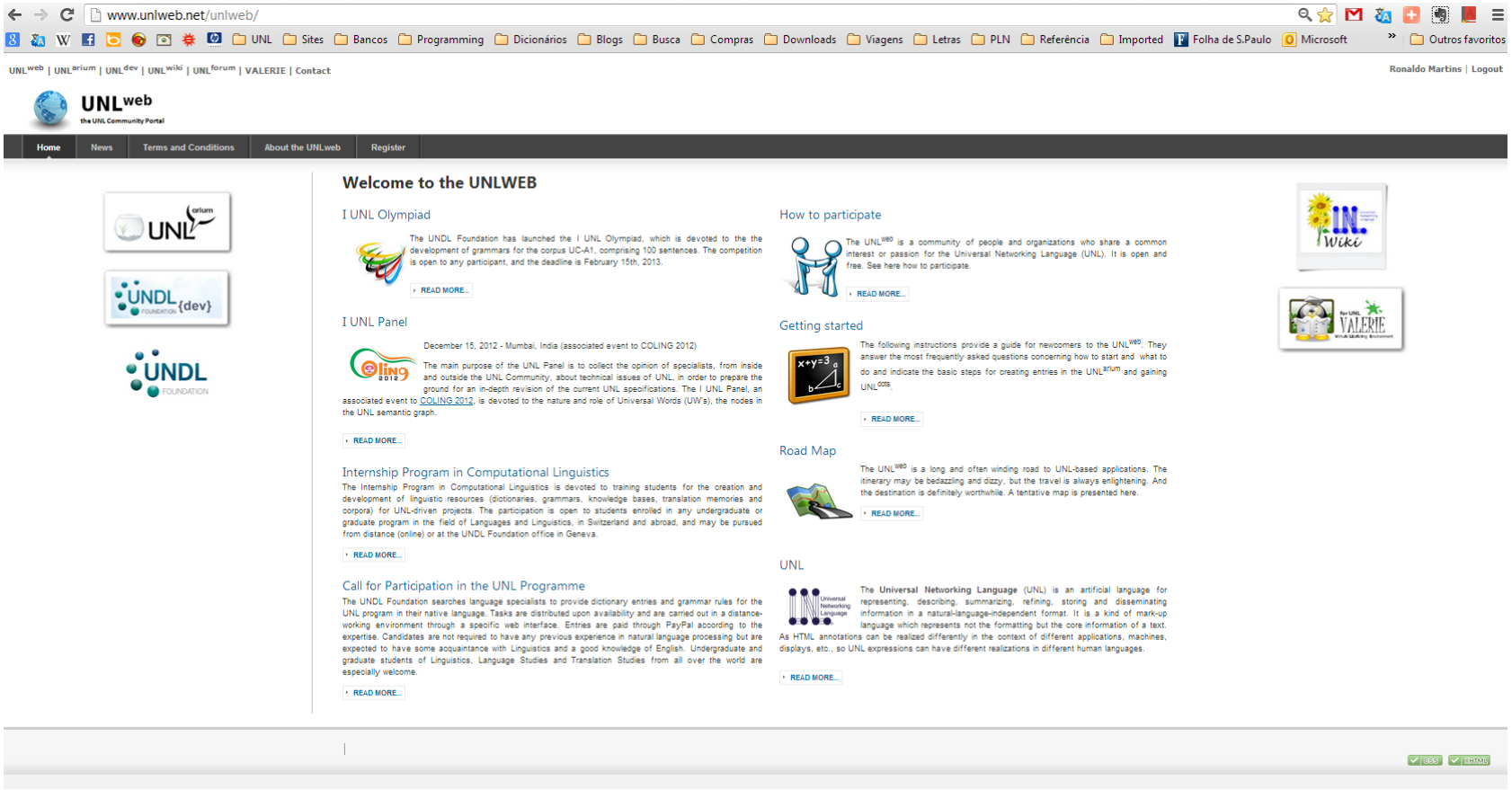
TRAINING & RESEARCH

VALERIE

UNLversity

UNL forum

www.unlweb.net



FoR-UNL

LEVEL	UNL-NL DIC	NL-UNL DIC	UNL-NL Grammar	NL-UNL Grammar
A1	MIR-A1	BRUNO-A1	UC-A1	NC-A1
A2	MIR-A2	BRUNO-A2	UC-A2	NC-A2
B1	MIR-B1	BRUNO-B1	UC-B1	NC-B1
B2	MIR-B2	BRUNO-B2	UC-B2	NC-B2
C1	MIR-C1	BRUNO-C1	UC-C1	NC-C1
C2	MIR-C2	BRUNO-C2	UC-C2	NC-C2

UNDL FOUNDATION

The UNL^versity

The UNL^versity

- Training and Research Program
  - Internship Program in Computational Linguistics
  - UNL Panel
    - I UNL Panel (Mumbai, COLING 2012)
    - II UNL Panel (Sofia, ACL 2013)
  - UNL School
    - 2012
      - VIII UNL School (Geneva, February 2012)
      - IX UNL School (Mumbai, June 2012)
      - X UNL School (Alexandria, October, 2012)
    - 2013
      - XI UNL School (Macau, March 2013)
      - XII UNL School (Geneva, June 2013
      - XIII UNL School (Yerevan, September 2013)
      - XIV UNL School (Kuwait, November 2013)

UNDL FOUNDATION

XI UNL School

Goals

- To build the basic modules of a NL-UNL (analysis) grammar for the corpus UCA1
- To build the basic modules of a UNL-UNL (generation) grammar for the corpus UCA1

Schedule

- March 11th
  - Introduction
  - Corpus
  - Dictionaries
- March 12th
  - Grammars
- March 13th
  - UNLization (with IAN)
- March 14th
  - NLization (with EUGENE)
- March 15th
  - Evaluation & Discussion

Warnings

- Doubts are allowed: don't be afraid or shy.
- This is an ongoing initiative: we don't have all the answers yet.
- This is not a competition.
- All the material will be available at www.unlweb.net/wiki/XI_UNL_School
- You may also use the forum (macau2013@unlweb.net) and the UNLCommunity to discuss problems and issues

Corpus

Corpus

- NC (NL>UNL)
- UC (UNL>NL)
  - A1 (100 sentences)
  - A2 (+300 sentences)
  - B1 (+ 5 texts)

Exercise #1 (10 min)

1. Translate the file uca1_eng.txt to your locale and name the resulting file uca1_<your locale>.txt. (pre workshop task)
2. Upload the file uca1_<your locale>.txt to:
    1. the UNL Community (Macau Group)
    2. to IAN (UNLWEB>UNLDEV>IAN>NL Files)

Exercise #2 (60 min)

1. Download the corpus uca1_unl_exp.txt from www.unlweb.net/resources/corpus/UCA1/uca1_unl_exp.txt.
2. Generate the corpus uca1_unl_exp.txt to your locale by hand (in a separate .txt file, one sentence per line, only the results). Name it uca1_<your locale>_exp.txt.
3. Upload the file uca1_<your locale>_exp.txt to the UNL Community (Macau group).

Exercise #3

1. Compare the results of:
    1. uca1_<your locale>.txt
    2. uca1_<your locale>_exp.txt
2. Present and explain the differences to the group

Dictionaries

Universal Words

- nodes in the UNL graph
- temporary (not to be included in the UNL Dictionary)
  - 3.14159
  - www.undlfoundation.org
  - H_2O
- permanent (to be included in the UNL Dictionary)

TYPE	GRANULARITY	UNL (simplified)	ENGLISH
SIMPLE	NODE	big	big
COMPOUND	NODE + ATTRIBUTE	big.@more	bigger
COMPLEX	HYPER-NODE	obj(affix,stamp)	to stamp

UNDL FOUNDATION

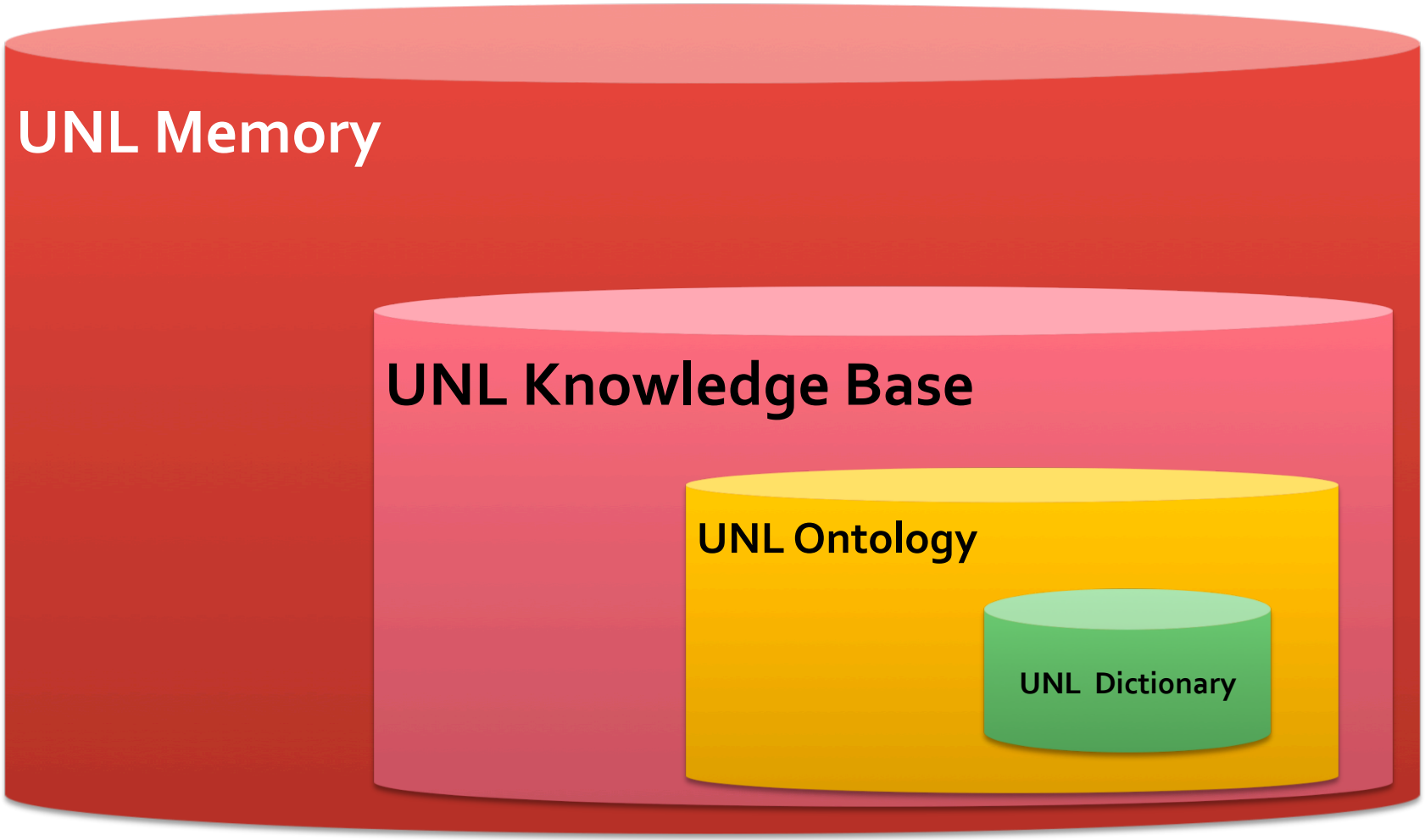
Semantic Accessibility

UCI (Uniform Concept Identifier)

- UCL (Uniform Concept Locator)
  - ucl://<AUTHORITY>/<ID>
    - ucl://unlkb.unlweb.net/104379964
- UCN (Uniform Concept Name)
  - ucn:<LID>:<NSS>
    - ucn:eng:table(icl>furniture)
    - ucn:fra:table(icl>mobilier)
    - ucn:esp:mesa(icl>mobiliario)
    - ucn:deu:Tisch(icl>Möbel)
    - ucn:rus:стол(icl>мебель)

UNDL FOUNDATION

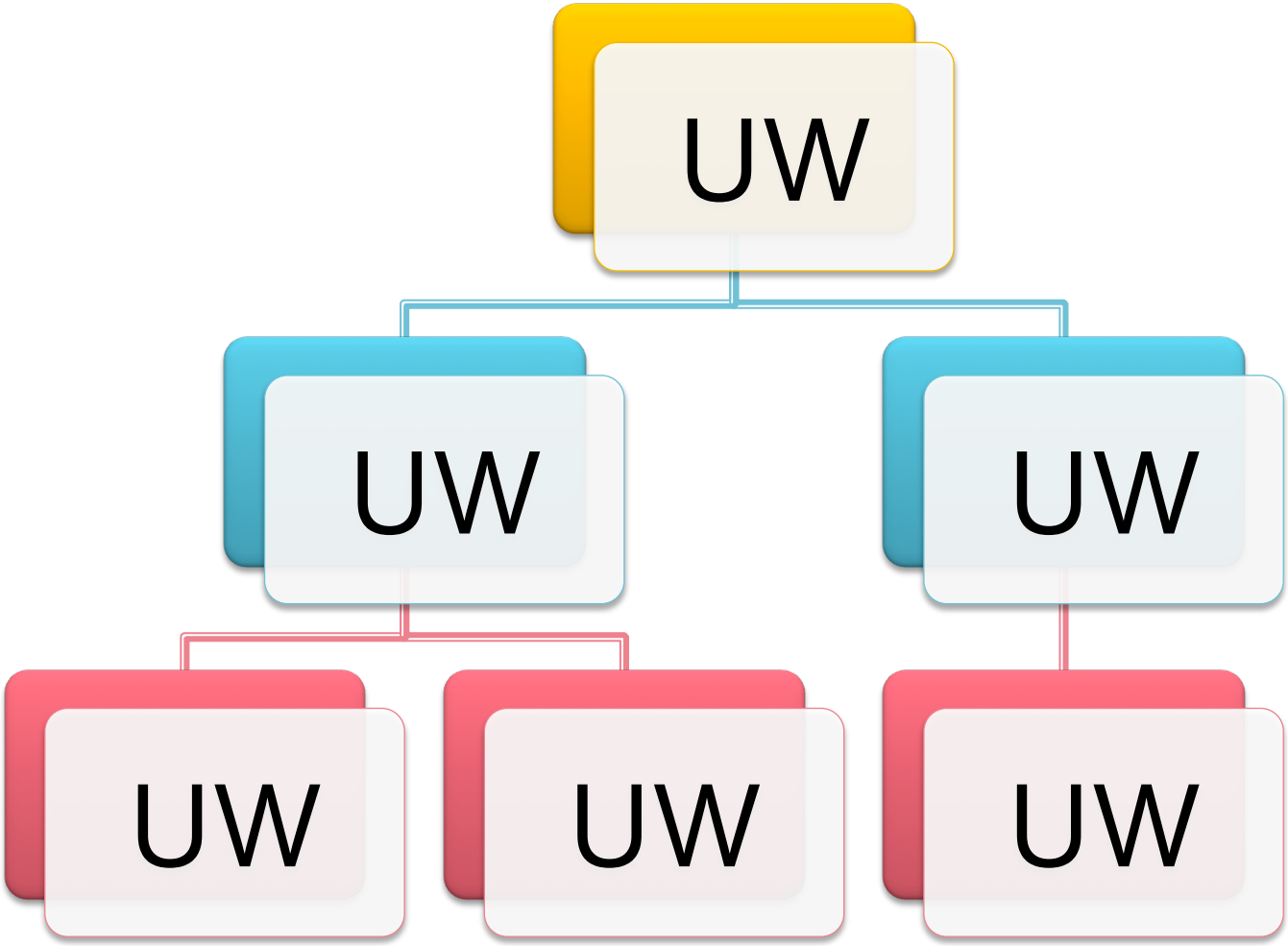
UNL Lexical Resources



UNL Dictionary

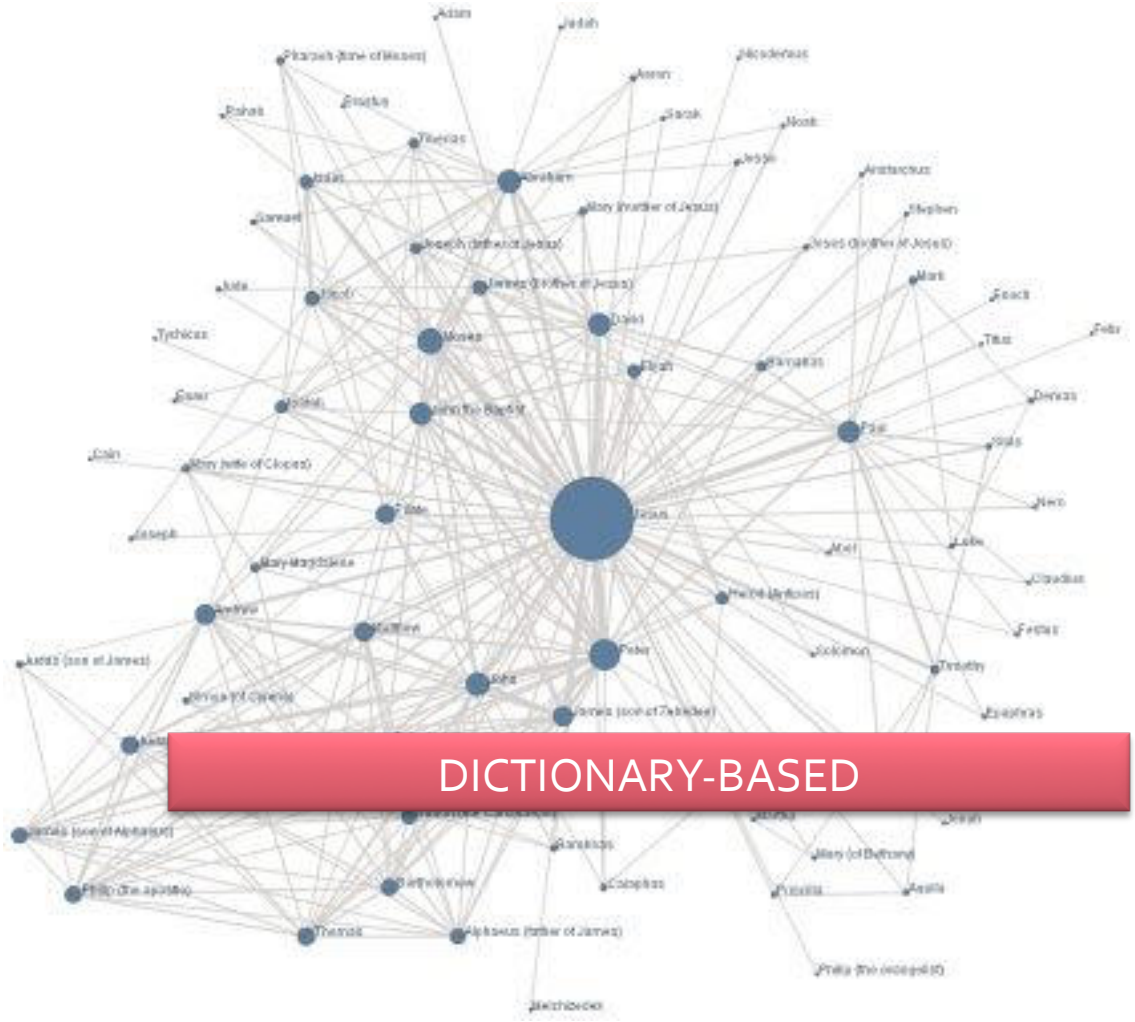
UNDL FOUNDATION

UNL Ontology



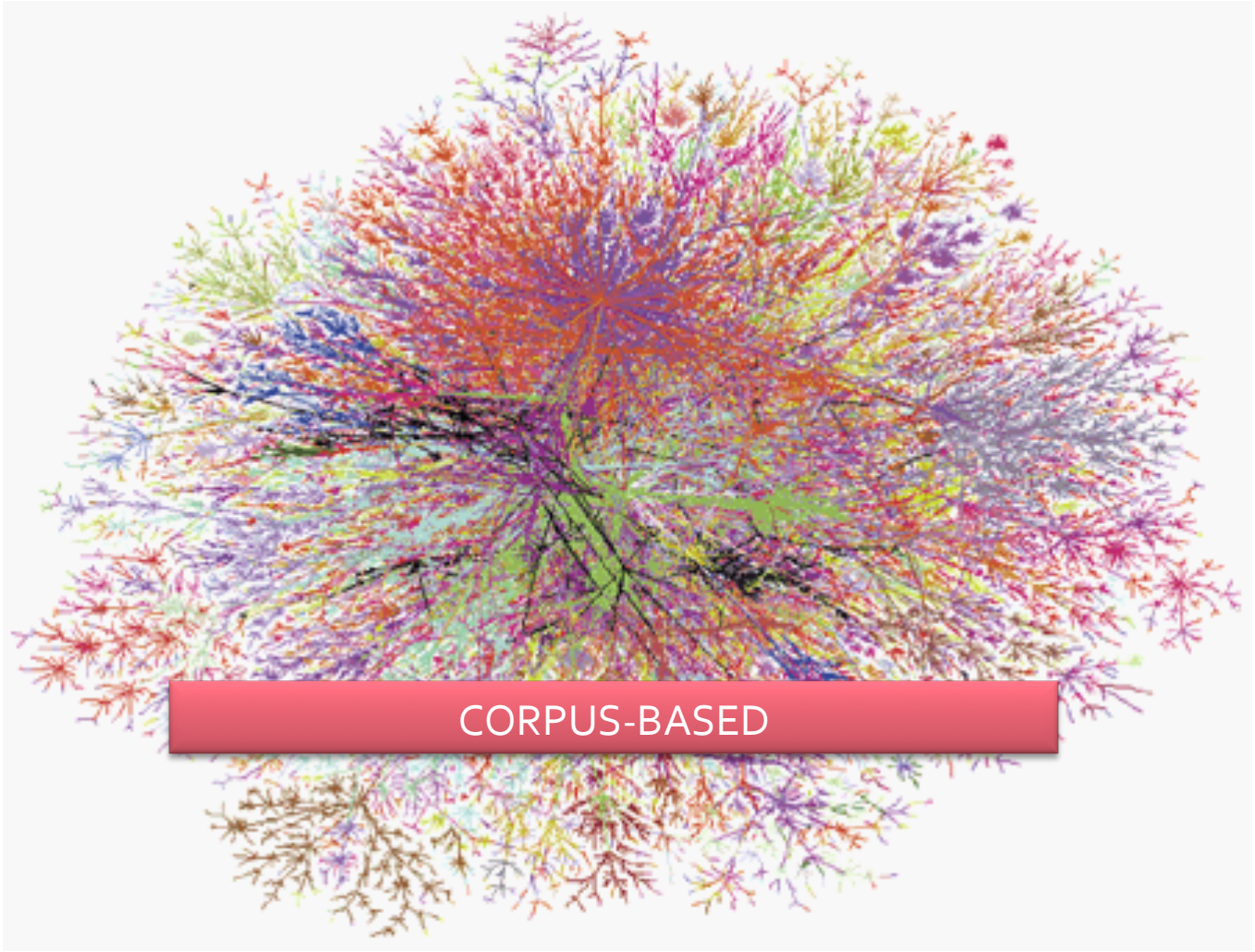
UNDL FOUNDATION

UNL Knowledge Base



UNDL FOUNDATION

UNL Memory

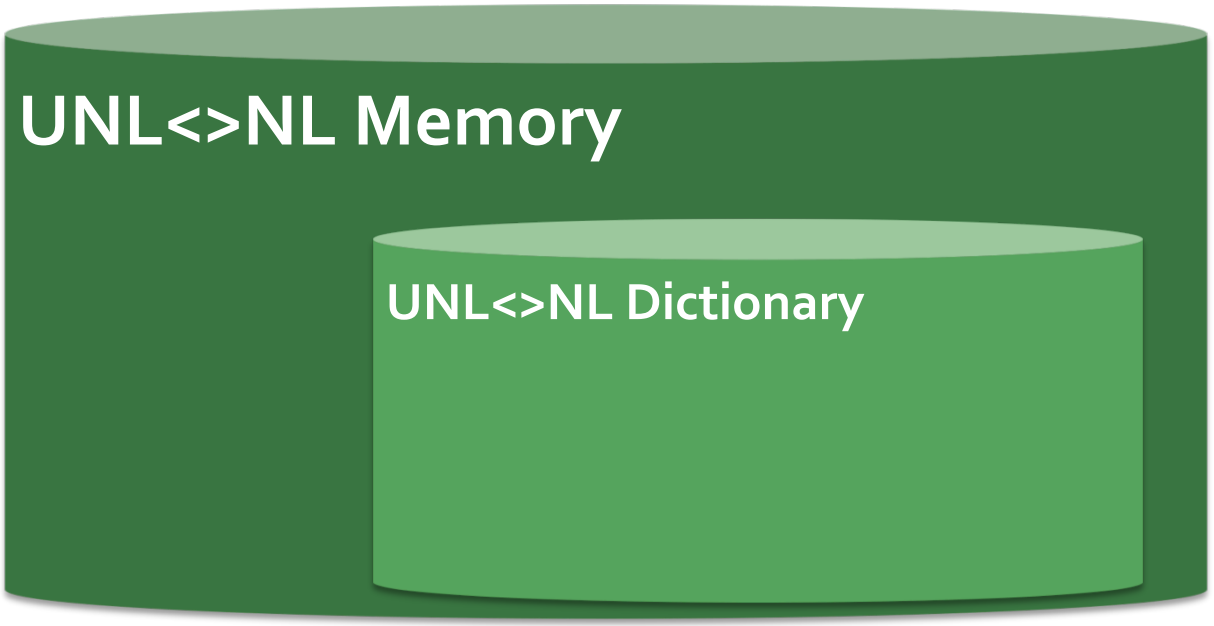


UNDL FOUNDATION

NL Lexical Resources

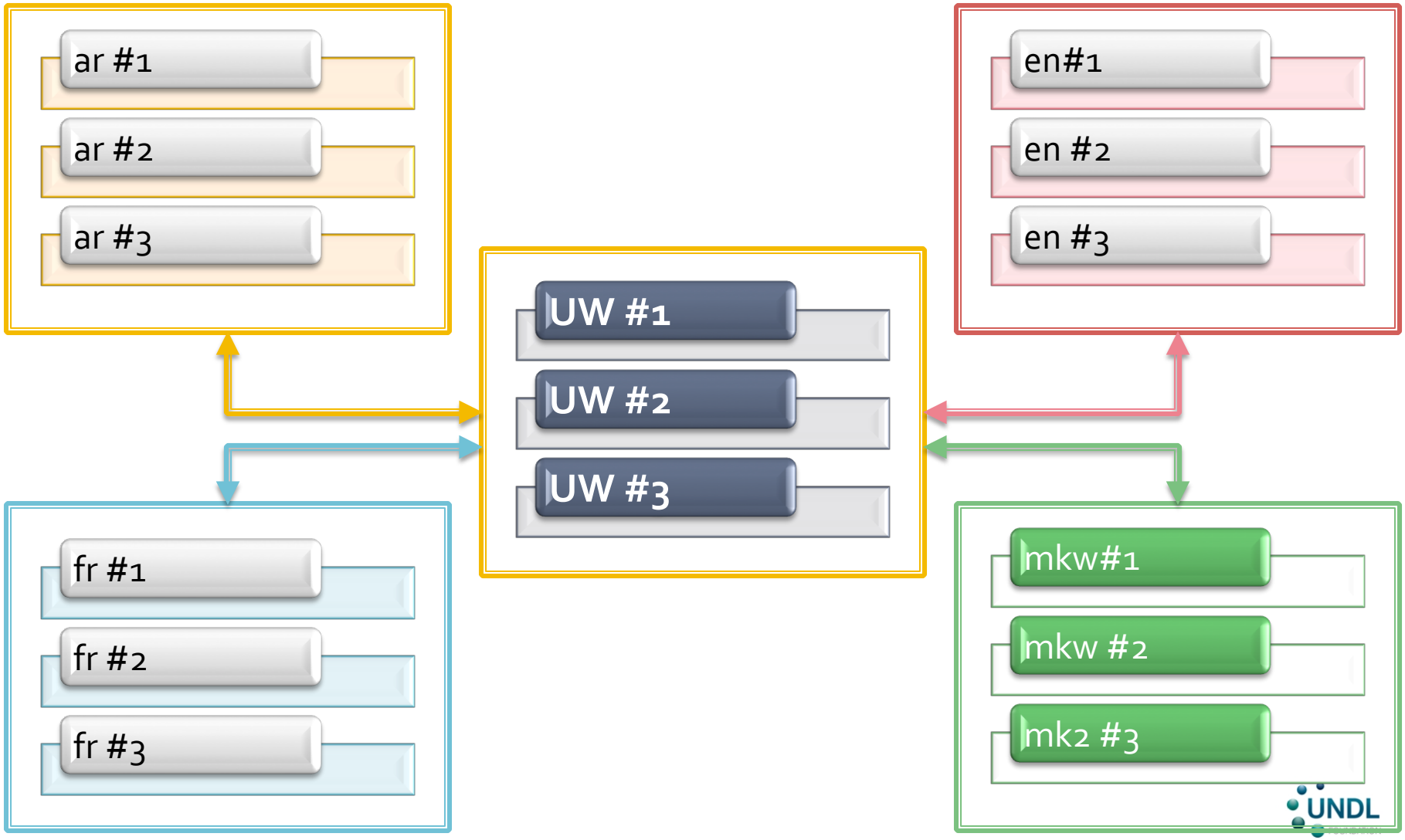
UNDL FOUNDATION

UNL<>NL Lexical Resources



UNL<>NL Dictionaries

UNL<->NL Dictionaries

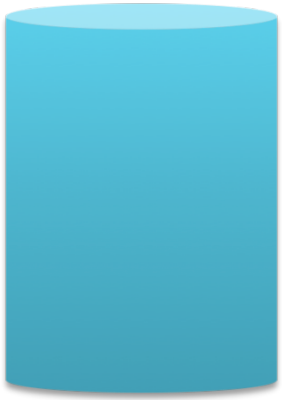


UNL<->NL Dictionaries

- NL-UNL Dictionary (Analysis)
  - Enumerative (word forms)

    [table] {2883} "table" (POS=NOU,**NUM=SNG**) <eng,0,0>;
    [tables] {2883} "table" (POS=NOU,**NUM=PLR**) <eng,0,0>;

    [foot] {2883} "foot" (POS=NOU,NUM=**SNG**) <eng,0,0>;
    [feet] {2883} "foot" (POS=NOU,NUM=**PLR**) <eng,0,0>;



- UNL-NL Dictionary (Generation)
  - Generative (base forms)
    - [table] {2883} "table" (POS=NOU,NUM=SNG,**PAR=M2**) <eng,0,0>;
    - [foot] {2883} "100284665" (POS=NOU,PAR=M1,**FLX(PLR:="feet";)**) <eng,0,0>;

Dictionary Specs
www.unlweb.net/wiki/Dictionary_Specs

- Dictionary Specs
  - Dictionary structure
    - a plain text file (.txt)
    - one entry per line
    - entries must have the following format:

[NLW] {ID} "UW" (ATTR , ...) < LG , FRE , PRI >; COMMENTS

[NLW]

[NLW] {ID} "UW" (ATTR , …) < LG , FRE , PRI >; COMMENTS

- a multiword expression: [United States of America]
- a compound: [hot-dog]
- a simple word: [happiness]
- a simple morpheme: [happ]
- a complex structure: [[bring] [back]]
- a non-motivated linguistic entity: [g]

{ID}

[NLW] {ID} "UW" (ATTR , ...) < LG , FRE , PRI >; COMMENTS

- The unique identifier (primary-key) of the entry.

"UW"

[NLW] {ID} "UW" (ATTR , …) < LG , FRE , PRI >; COMMENTS

- The Universal Word of UNL. This field can be empty if a word does not need a UW.

(ATTR, …)

[NLW] {ID} "UW" (ATTR , …) < LG , FRE , PRI >; COMMENTS

- The list of features of the NLW.
- Attributes should be separated by ",".
- It can be:
  - a list of simple features: (NOU, MCL, SNG)
  - a list of attribute-value pairs: (pos=NOU, gen=MCL, num=SNG)
  - a list of transformation rules : (plural:="oo":"ee")
    - Replacement
      - <ATTRIBUTE>":="<SOURCE>":"<TARGET>
      - plural:="oo":"ee
    - Left appending
      - <ATTRIBUTE>":="<LEFT DELETION>"<"<LEFT ADDITION>
      - not:=<"un"
    - Right appending
      - <ATTRIBUTE>":="<RIGHT ADDITION>">"<RIGHT DELETION>
      - plural:=y>ies

<LG, FRE, PRI>

[NLW] {ID} "UW" (ATTR , ...) < LG , FRE , PRI >; COMMENTS

- FLG
  - The three-character language code according to ISO 639-2.
- FRE
  - The frequency of NLW in natural texts. Used for natural language analysis (NL-UNL). It can range from 0 (less frequent) to 255 (most frequent).
- PRI
  - The priority of the NLW. Used for natural language generation (UNL-NL). It can range from 0 to 255.

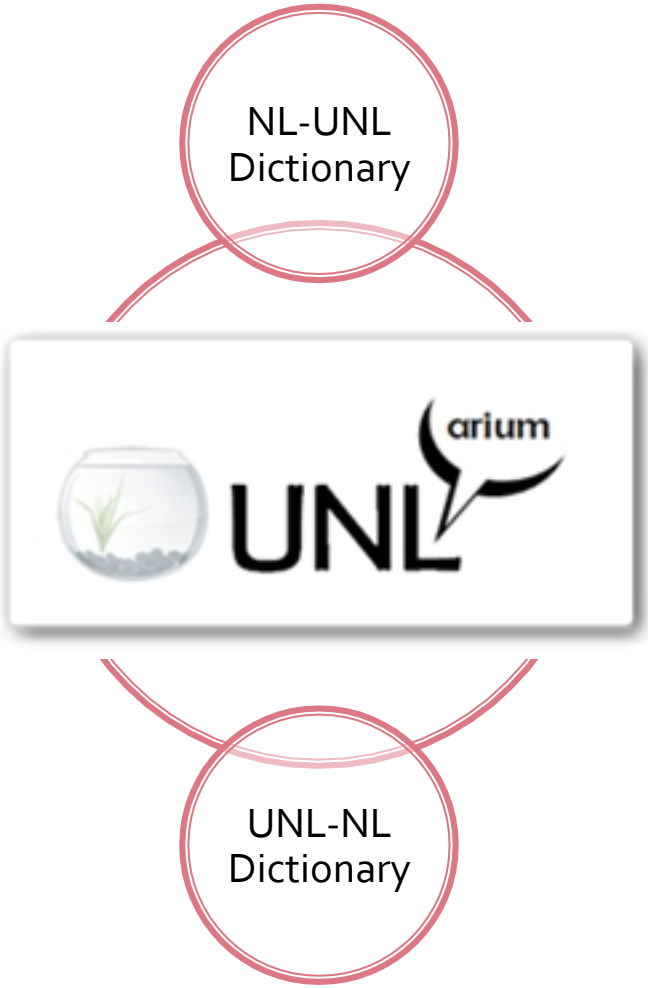
Examples

- [China]{24} "China(iof>Asian country)" (NOU, WRD, SNG, P0, F0) <eng,0,0>;
- [choose]{106} "to choose(icl>to decide)" (POS=VER, LEX=WRD, PAR=M1, FRA=Y76, FLX(3PS&PRS&IND:=0>"s"; PAS:="chose"; PTP:="chosen"; GER:="choosing";)) <eng, 0,0>;
- [clear-eyed]{25} "clear-eyed(icl>discerning)" (POS=ADJ, LEX=WRD, PAR=M0, FRA=Y0) <en,0,0>;
- [Peter]{177}"Peter(iof>person)"(NOU)<eng,10,30>;
- [kill]{5987}"kill(icl>do)"(FLX(PAS:=0>"ed";))<eng,70,80>;
- [[bring] [back]]{2345}"bring back"(POS=VER,VA(01>02),#01(POS=VER,FLX(PAS:=3>"ought";)),#02(POS=PRE))<en g,50,34>;
- [/br(ing|ought)/] "bring(icl>do)" (POS=VER) <eng,0,0>;
- [[/br(ing|ought)/] [back]]{2345} "bring back(icl>do)" (POS=VER,#01(POS=VER),#02(POS=PRE))<eng,50,34>;
- [/colo(u)?r/] "color" (POS=NOU) <eng,0,0>; (NLW = {color, colour})
- [/cit(y|ies)/] "city" (POS=NOU) <eng,0,0>; (NLW = {city, cities})
- [/(\d){4}/] "" (ENT=YEAR) <eng,0,0>; (NLW = any sequence of four digits)
- [city] "/city(.)*/" (POS=NOU) <eng,0,0>; (UW = any UW that starts by the string "city")
- [city] "/(.)+\(iof\>city\)/" (POS=NOU) <eng,0,0>; (UW = any UW that ends by the string "(iof>city)"

Special issues (Null UW)

- [extremely]{}"extremely"(LEX=A,POS=SAV)<eng,0,0>;
  - an extremely beautiful car
- [very]{}"" (LEX=A,POS=SAV,att=@plus)<eng,255,0>;
  - very beautiful
- [of]{}"" (LEX=P,POS=PRE,rel=mod)<eng,255,0>;
  - the book of John
- [after]{}"" (LEX=P,POS=PRE,rel=tim,att=@after)<eng,255,0>;
  - few minutes after noon

Building dictionaries



Exercise #4 (10 min)

- Create the UNLization dictionary for the corpus UCA1_<your locale>txt. (pre-workshop task)
- Upload the dictionary to IAN
  - (UNLWEB>UNLDEV>IAN>DICTIONARY)
- Run IAN for the dictionary and corpus (without any rule) and analyze the results

Exercise #5 (10 min)

- Create the NLization dictionary for the corpus UCA1_unl.txt (pre-workshop task)
- Upload the dictionary to EUGENE
  - (UNLWEB>UNLDEV>EUGENE>DICTIONARY)
- Run EUGENE for the dictionary and corpus (without any rule) and analyze the results