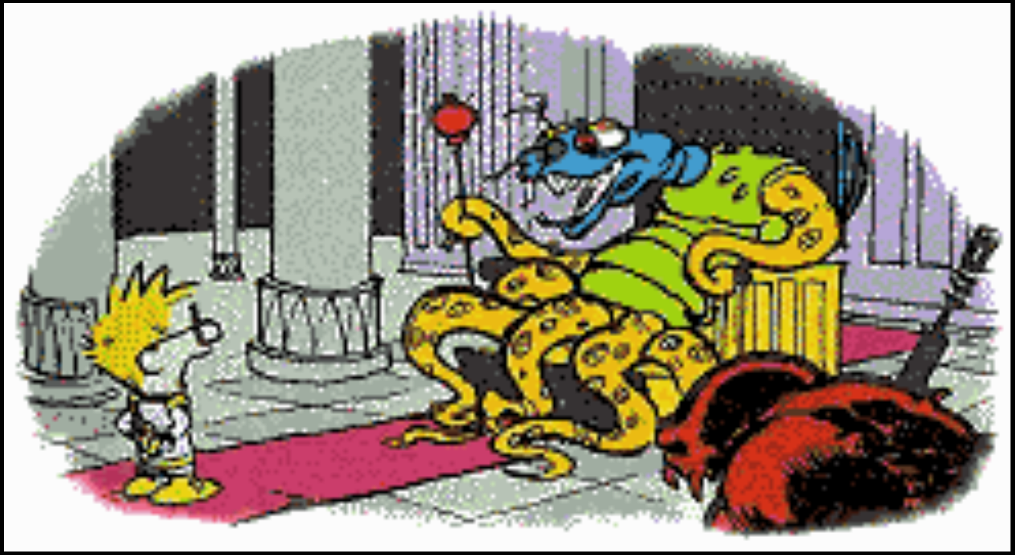


XV UNL School

Geneva, July 21-25, 2014

Welcome

Participants

- Ali Safari (Azerbaijani)
- Anna Loukopoulou (Greek)
- Eirini Kouvara (Latin)
- Georgia-Rengina Loukatou (Greek)
- Martin Luts (Estonian)
- Maryam Faal Hamedanchi (Persian)
- Parteek Kumar (Panjabi)
- Sergiy Prots (Ukrainian)
- Somdev Kar (Bengali)
- Yordanka Stancheva (Bulgarian)
- Ana Luisa Varani Leal (University of Macau)
- Athina Papachrysostomou (University of Patras)
- Monica Gallo (University of Geneva)
- Sameh Alansary (University of Alexandria)

Goals

- To teach you how to build the basic infrastructure for UNLization and NLization

Program

- July 21st
  - Introduction
  - Normalization
- July 22nd
  - Tokenization
- July 23rd
  - UNLization
- July 24th
  - NLization
- March 14th
  - Evaluation & Discussion

Timetable

Monday 21st	Tuesday 22nd	Wednesday 23rd	Thursday 24th	Friday 25th
9-12	9-12	9-12	9-12	9-12
14-17	14-17	14-17	14-17	

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/XV_UNL_School

Day #1

Morning

- Introduction

Lunch break

Afternoon

- Corpus
- N-grammar
- Word list

Introduction

Universal
Networking
Language®

The Universal Networking Language (UNL)

UNL

translation

knowledge representation

☼1996

UNITED NATIONS
UNIVERSITY

UNDL
FOUNDATION

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$_j$

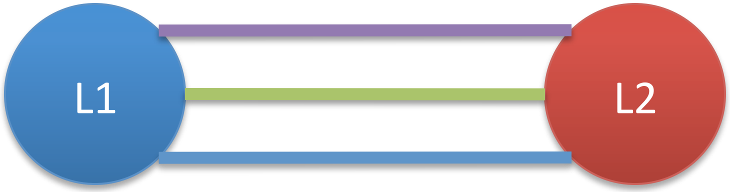
Structure

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

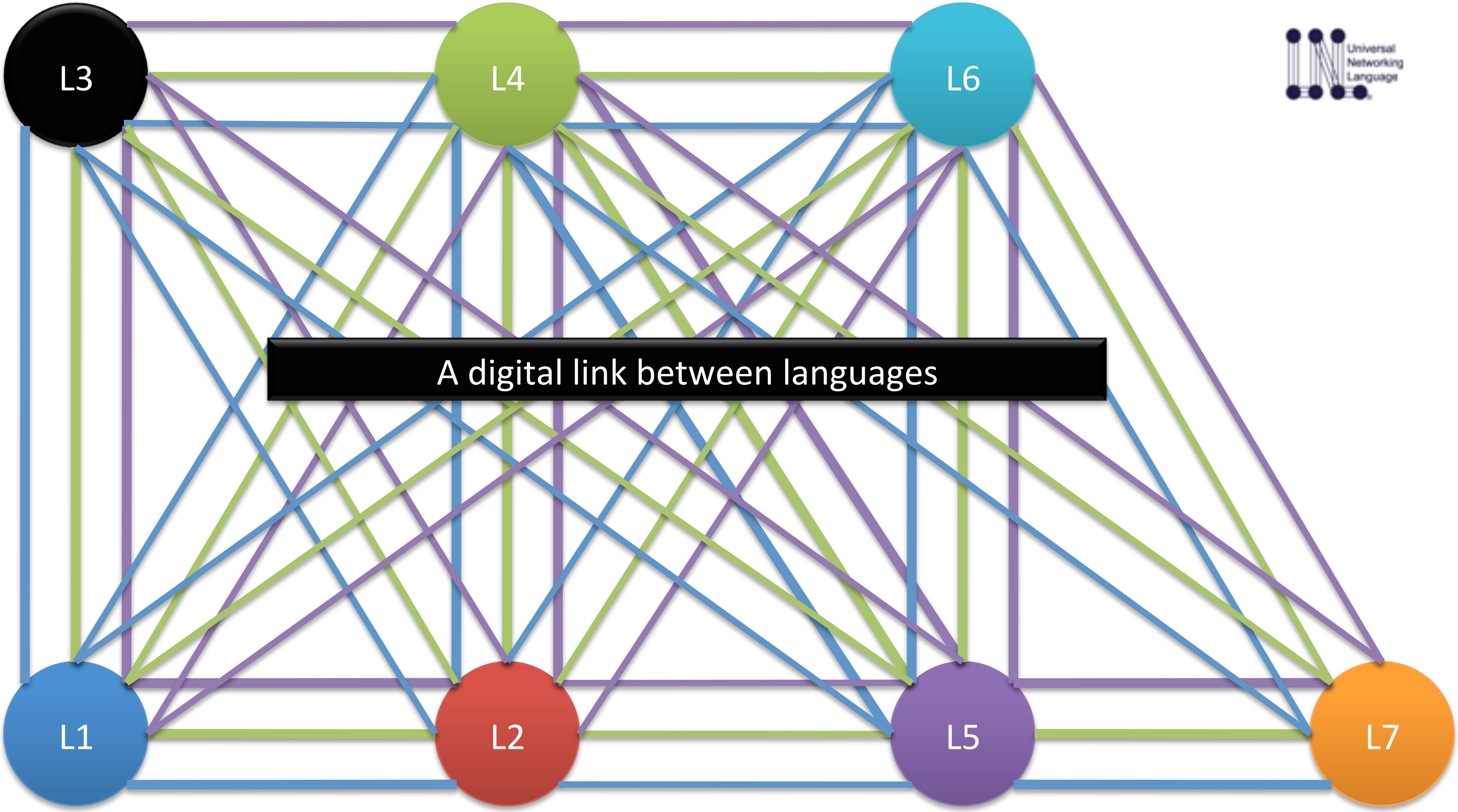
CONCEPTS = Universal Words (UWs)

CONCEPT SPECIFIERS = Universal Attributes

RELATIONS BETWEEN CONCEPTS = Universal Relations



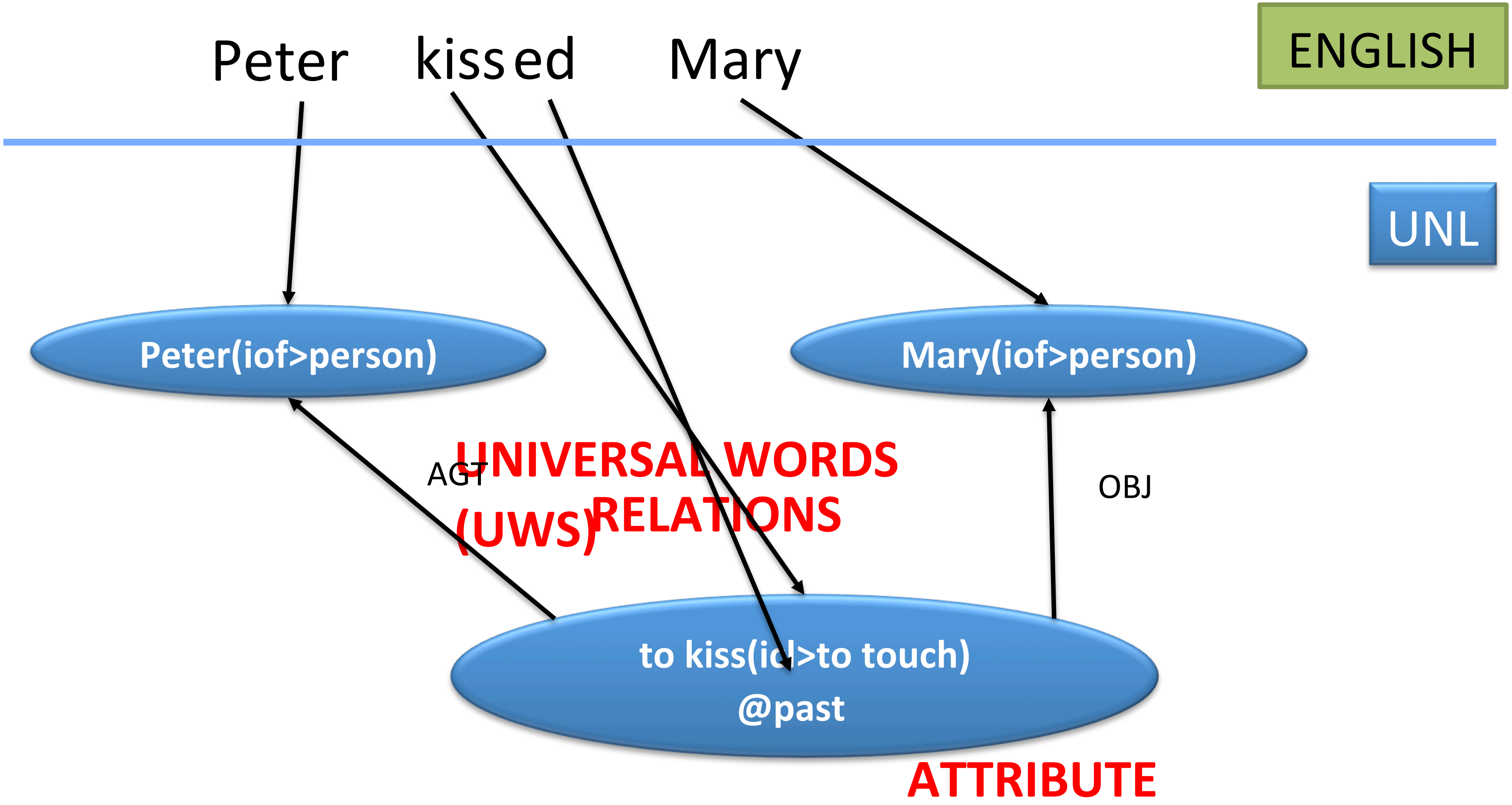
The Universal NETWORKING Language



A digital link between languages

Universal Networking Language

Natural Language-to-UNL (UNL-ization)



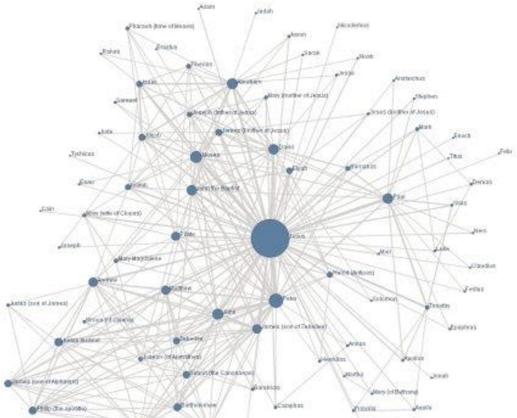
The UNL System

The UNL System

C2 cotidiano TERÇA-FEIRA, 23 DE JANEIRO DE 2007 FOLHA DE S.PAULO

RUBEM ALVES

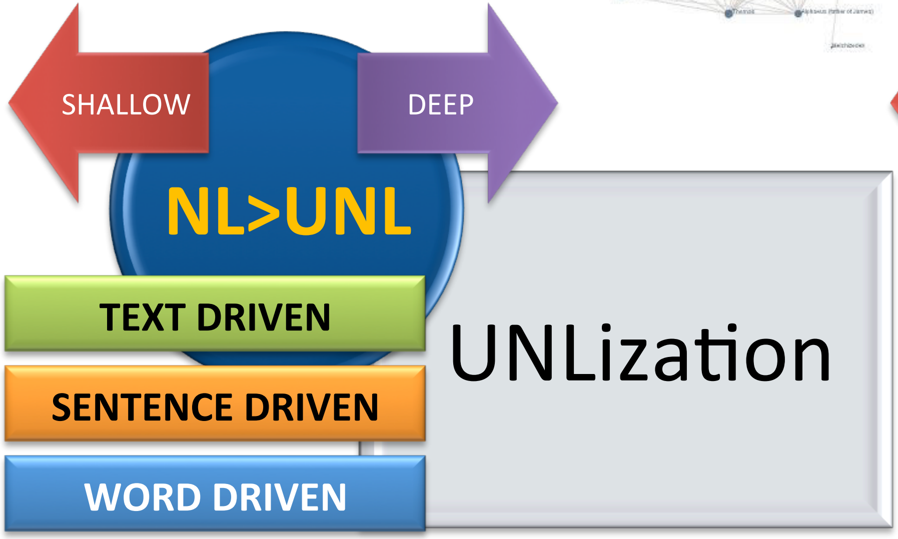
O aluno perfeito



C2 cotidiano TERÇA-FEIRA, 23 DE JANEIRO DE 2007 FOLHA DE S.PAULO

RUBEM ALVES

O aluno perfeito



SHALLOW

DEEP

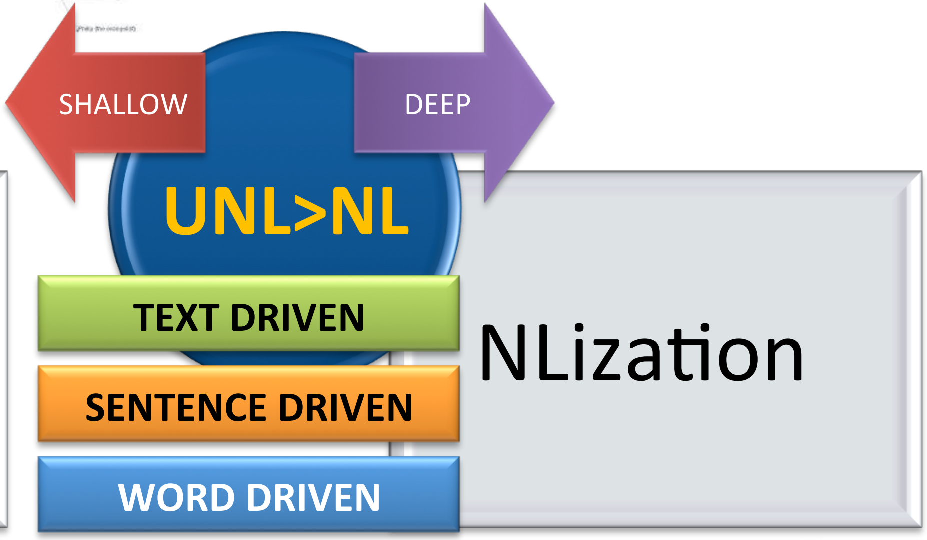
NL>UNL

TEXT DRIVEN

SENTENCE DRIVEN

WORD DRIVEN

UNLization



SHALLOW

DEEP

UNL>NL

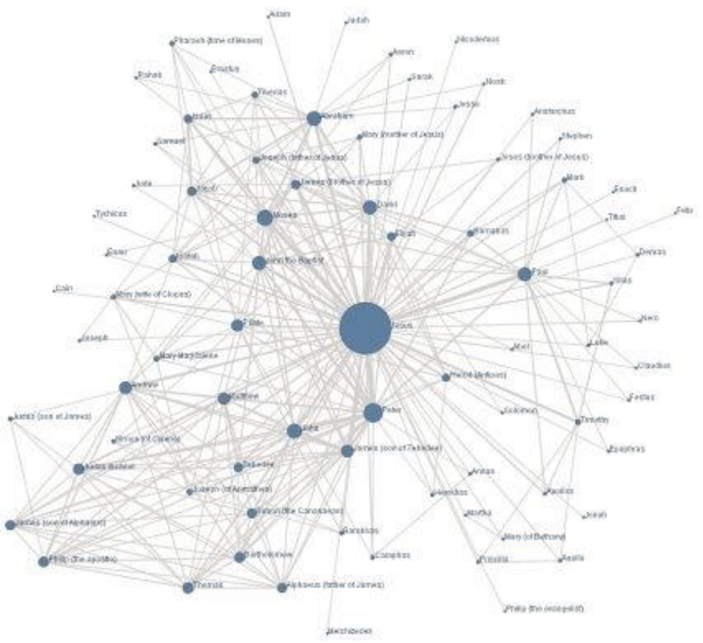
TEXT DRIVEN

SENTENCE DRIVEN

WORD DRIVEN

NLization

Uses of UNL



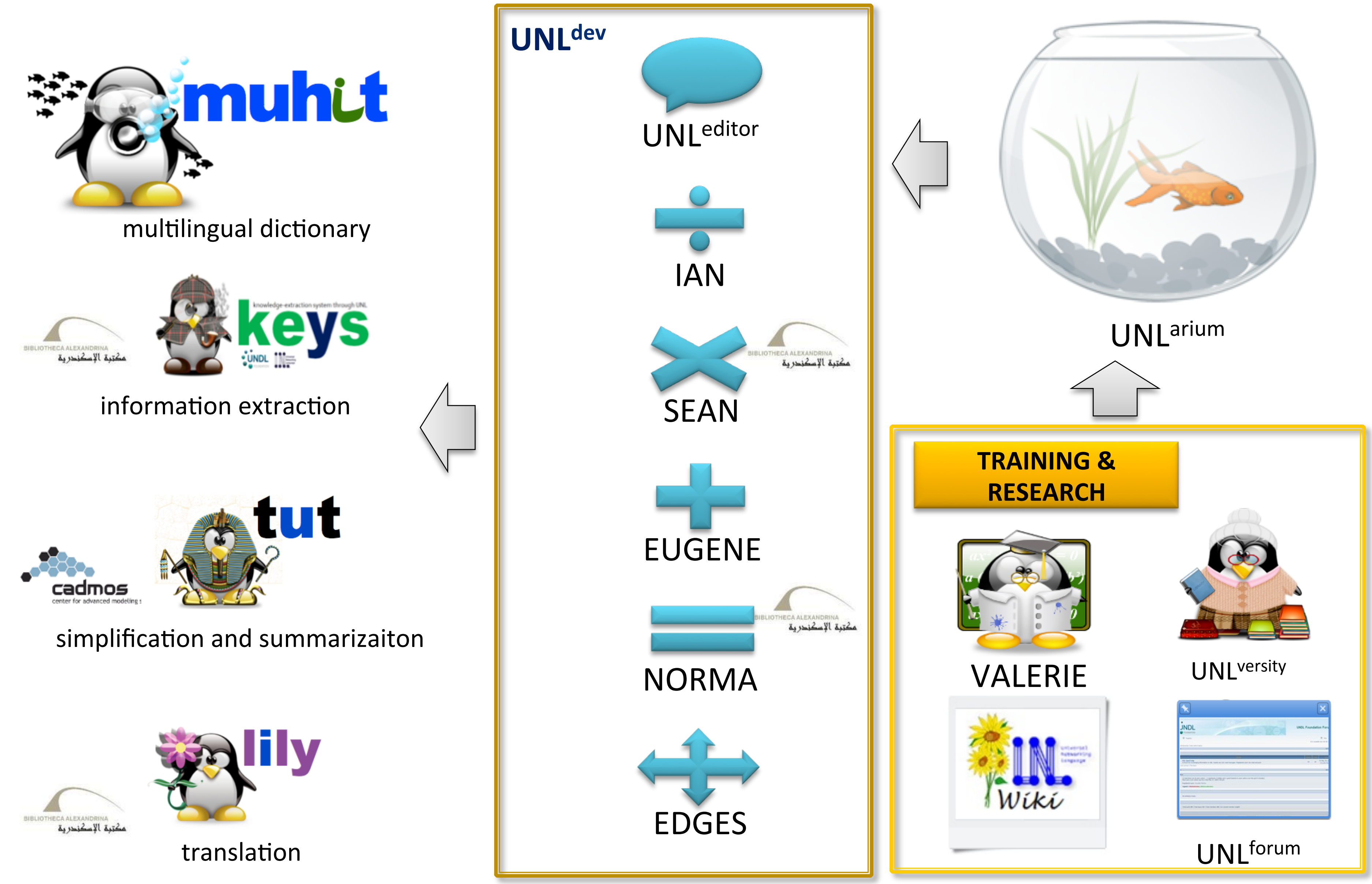
- Search
- Sentiment analysis
- Information extraction
- Generation
- Normalization
- Summarization
- Simplification

The UNDL Foundation Road Map

FRONT-END APPLICATIONS

BACK-END APPLICATIONS

RESOURCES



muhit

multilingual dictionary

keys

information extraction

tut

simplification and summarizaiton

lily

translation

UNL^dev

UNL^editor

IAN

SEAN

EUGENE

NORMA

EDGES

UNL^arium

TRAINING & RESEARCH

VALERIE

UNL^versity

UNL Wiki

UNL^forum

FoR-UNL

LEVEL	DICTIONARY	GRAMMAR
A1	5,000	Morphology: N
A2	10,000	Morphology: J, A, V
B1	20,000	Syntax: NP
B2	40,000	Syntax: VP
C1	70,000	Syntax: IP
C2	100,000	Syntax: CP

Current Status

LANGUAGE	DICTIONARY	GRAMMAR
Arabic	C2	B2
Azerbaijani	A1	A1
Bengali	A1	A1
Bulgarian	A2	A2
Estonian	A2	A2
Greek (Ancient)	A0	A0
Greek (Modern)	A2	A2
Italian	B1	A2
Latin	C1	A2
Panjabi	A1	A1
Persian	B1	A2
Portuguese	B2	A2
Ukrainian	A2	A2

UNDL FOUNDATION

Any questions?

Corpus

The Universal Declaration of Human Rights

THE UNIVERSAL DECLARATION OF Human Rights

UNITED NATIONS

Article 1
All human beings are born free and equal in dignity and rights. They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood.

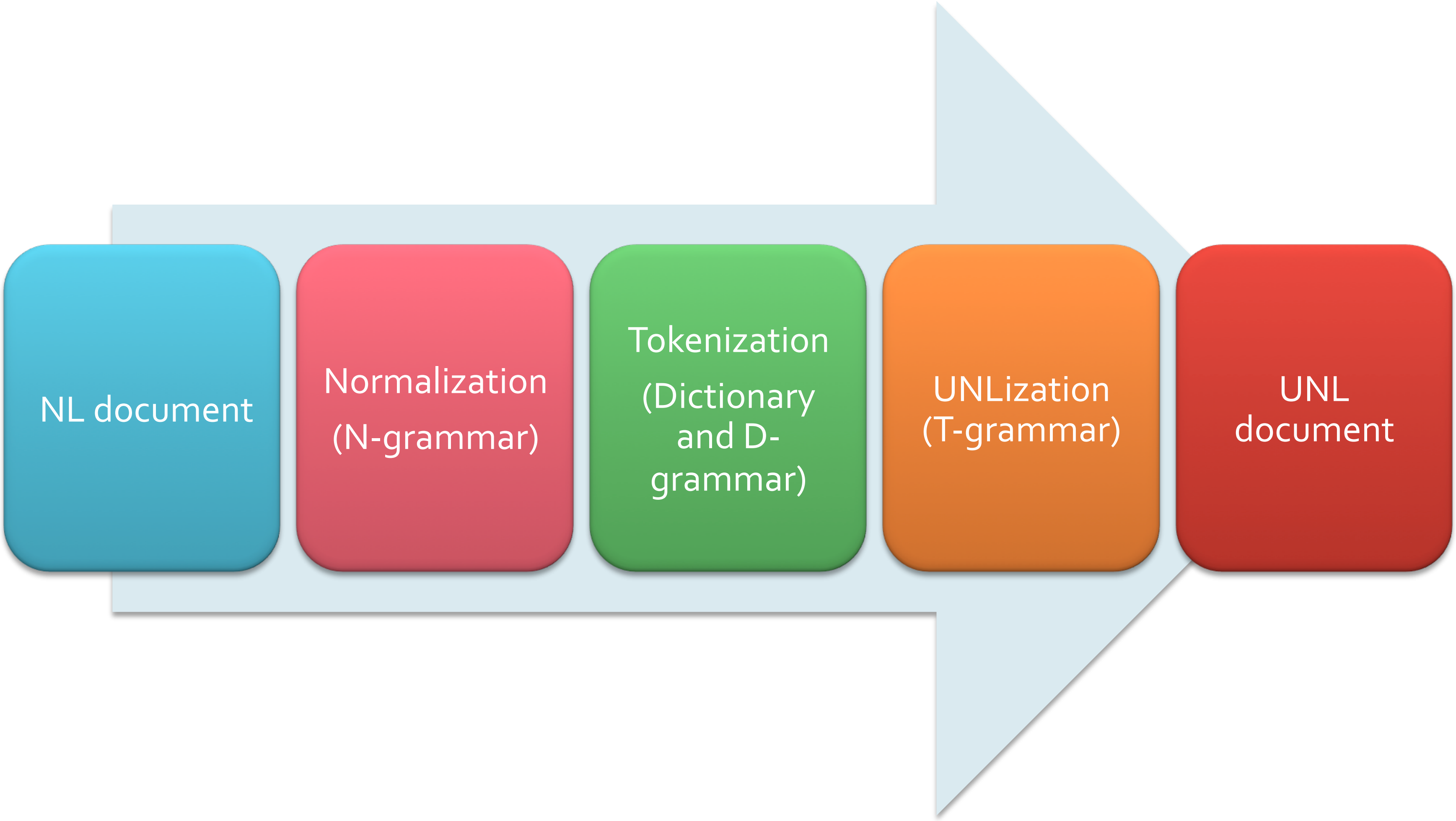
Article 2
Everyone is entitled to all the rights and freedoms set forth in this Declaration, without distinction of any kind, such as race, color, sex, language, religion, political or other opinion, national or social origin, property, birth or other status. Furthermore, no distinction shall be made on the basis of the political, jurisdictional or international status of the country or territory to which a person belongs, whether it be independent, trust, non-self-governing or under any other limitation of sovereignty.

Article 3
Everyone has the right to life, liberty and security of person.

Article 4
No one shall be held in slavery or servitude; slavery and the slave trade shall be prohibited in all their forms.

Article 5
No one shall be subjected to torture or to cruel, inhuman or degrading treatment or punishment.

Goal



NL document

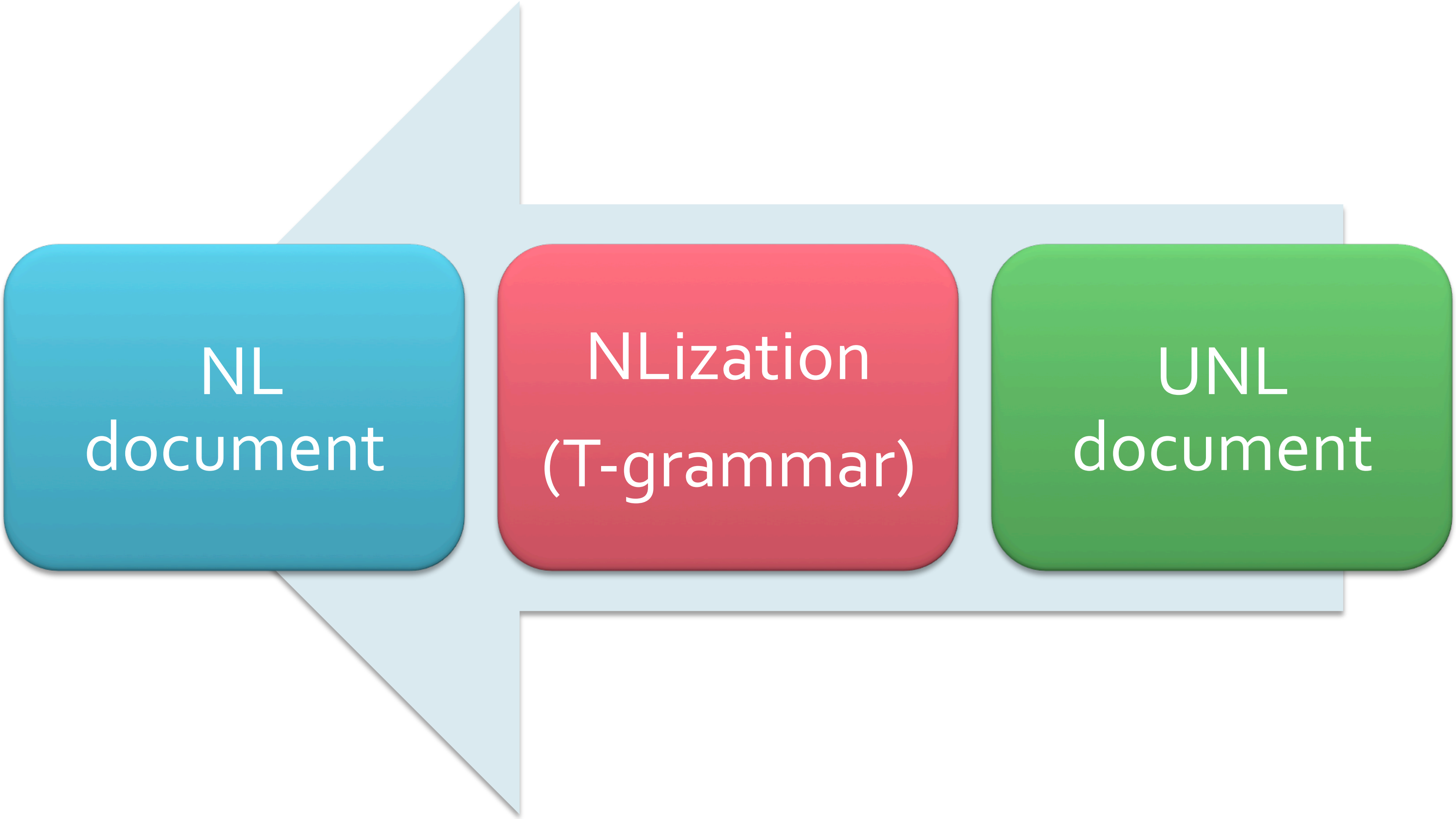
Normalization (N-grammar)

Tokenization (Dictionary and D-grammar)

UNLization (T-grammar)

UNL document

Goal (cont'd)



NL document

NLization (T-grammar)

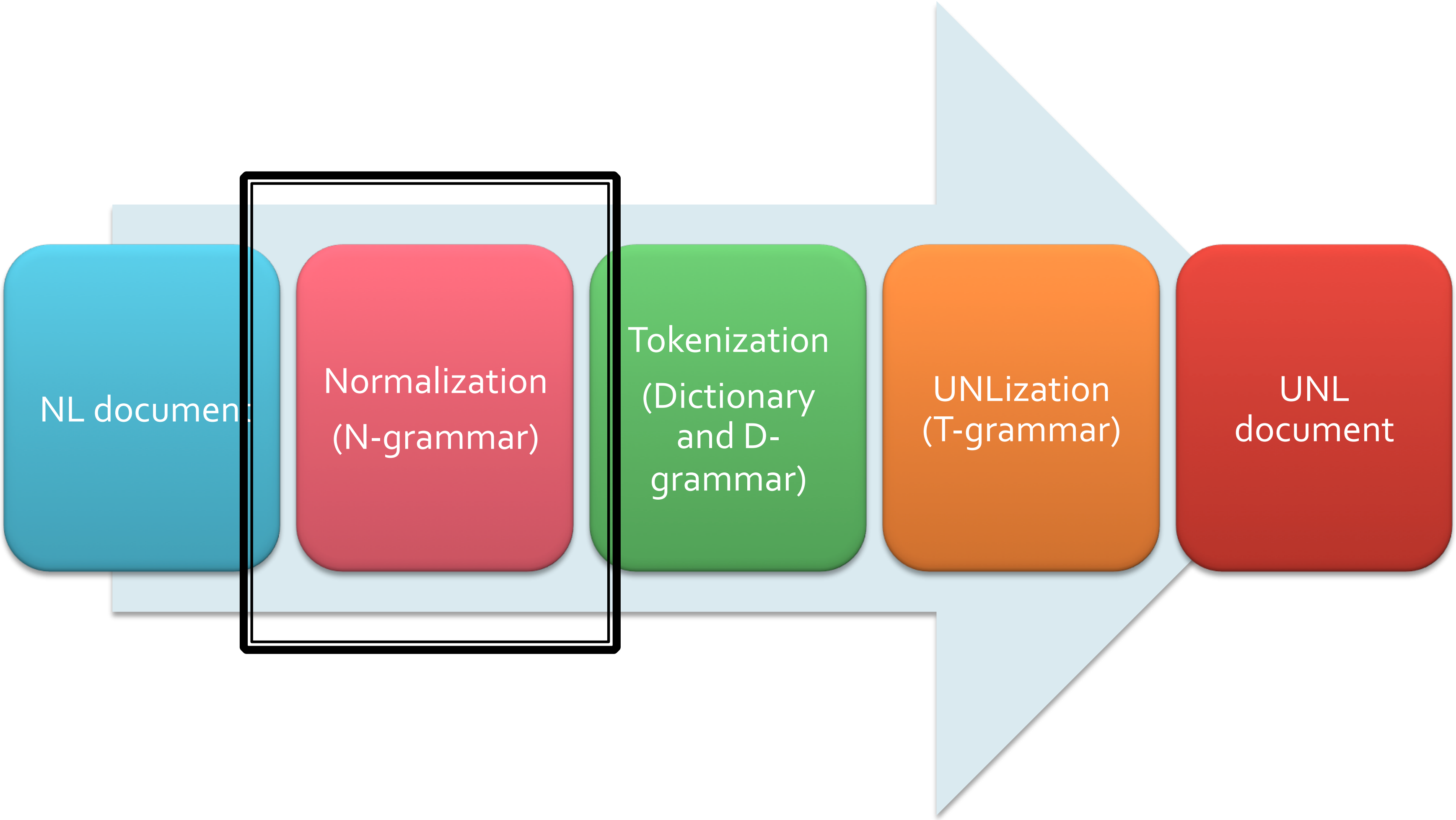
UNL document

Exercise #1 (30 min)

- Goal: To prepare the training corpus
- Deliverable: *corpus_org_<ID>.txt*
- Activities:
  - www.unlweb.net/wiki/XV_UNL_School

Normalization

Goal



NL document

Normalization

(N-grammar)

Tokenization

(Dictionary and D-grammar)

UNLization (T-grammar)

UNL document

Normalization (I)

Why is this necessary?

ORIGINAL

- Dr. Peter H. Smith isn’t coming on July 1st. He’ll be in another meeting in N.Y. I’ll check with him another date asap. Would u be available next week, say, around 2 PM?

NORMALIZED

- Doctor Peter H Smith is not coming on 01/07.
- He will be in other meeting in New York.
- I will check with him other date as soon as possible.
- You would be available in next week around 14:00:00?

Normalization (II)
What does it mean?

- Replacing contractions
  - don’t > do not, he’ll > he will (eng)
  - du > de le, aux > à les (fra)
- Replacing abbreviations
  - Dr. > doctor, N.Y. > New York, asap > as soon as possible
- Replacing variants and non-standard language
  - u > you, an > a
- Reordering
  - Would you > you would
- Filling gaps and ellipses
  - next week > in the next week
- Removing extra content
  - , say, > ∅
- Segmenting
  - He is not coming. He will be elsewhere > He is not coming.//He will be elsewhere.

Normalization (III)
How is this done?

- N-rules
  - (%a)(%b)…(%n):=(%a)(%b)…(%n);
  - Where:
    - left side (condition): % is a string or a regular expression
    - right side (action): % is coindexed to the left side
  - Examples:
    - ("don't"):=("do not");
    - ("dr."):=("doctor");
    - ("an "):=("a ");

Normalization (IV)
Segmentation

- Segmentation is done by assigning the features:
  - SHEAD (to the beginning of the new sentence) or
  - STAIL (to the end of the sentence)
    - There is no need to assign SHEAD and STAIL simultaneously
    - SHEAD and STAIL are automatically assigned to new line or line breaks
- Examples:
    - ("?",%a)(^STAIL,%b):=(%a)(%new,+STAIL)(%b);
    - (".",%a)(" ")("/[A-Z]/",%b):=(%a)(%new,+SHEAD)(%b);

Normalization (V)

Issues

- N-rules can only manipulate strings or regular expressions.
  - Features (such as N, NOU, MCL, etc.) cannot be used in N-rules.
    - ("Mr."):=("Mister"); (string manipulation)
    - ("/[A-Z]/",%x)(".",%y):=(%x); (regular expression manipulation)
    - ("Mr.",ABB):=("Mister"); (this is not a N-rule, because it involves a non-string element, i.e., ABB)
- Regular expressions may only be used in the left side.
  - ("/[A-Z]/",%x)(".",%y):=(%x);
  - ("/[A-Z]/")("."):=("/[A-Z]/");
- N-rules are recursive: rules will apply while conditions are true:
  - The rule "(" "):=("-");" will transform "a b c d e" into "a-b-c-d-e" (and not only in "a-b c d e")
- N-rules manipulate any strings meeting the conditions
  - ("art"):=("article"); provides "**art** 20">"**article** 20", but also "My name is B**art**">"My name is B**article**", "I love S**art**re">"I love S**article**re“
  - ({SHEAD|" "})("art")({STAIL|" "}):=()("article")(); (i.e., replace "art" by "article" if inbetween blank spaces or sentence boundaries
- The symbol ^ is used for negation and may be used to prevent infinite loops:
  - The rule (".",%x):=(%x)(+STAIL,%y); contains a loop, and will lead to (".")(STAIL)(STAIL)(STAIL)(STAIL)….
  - In order to prevent that, we have to indicate that STAIL must be added if it does not exist yet, i.e.: (".",%x)(^STAIL,%z):=(%x)(+STAIL,%y)(%z);

Normalization (VI)
Issues (cont'd)

- In the right side, changes may be expressed by the right side of A-rules inside each form. The default is replacement.
  - The rule "("a")(" ")("/[aeiou].+/"):=("an")( )( );"

  could also be expressed as
  - "("a")(" ")("/[aeiou].+/"):=(o>"n")( )( );"
- Rules apply only if all conditions are true.
  - The rule "("a")(" ")("/[aeiou].+/"):=("an")( )( );" will apply only in case of "a" before a blank and a node starting with "a", "e", "i", "o" or "u".
- Nodes may be deleted through replacement by zero:
  - (" "):=; (deletes all the blank spaces)
- Nodes in the left side that are not coindexed to nodes in the right side are deleted
  - (" ")("don't")(" "):=("do not"); provides "I don't know">"Ido notknow"
  - (" ")("don't")(" "):=()("do not")(); provides "I don't know">I do not know"

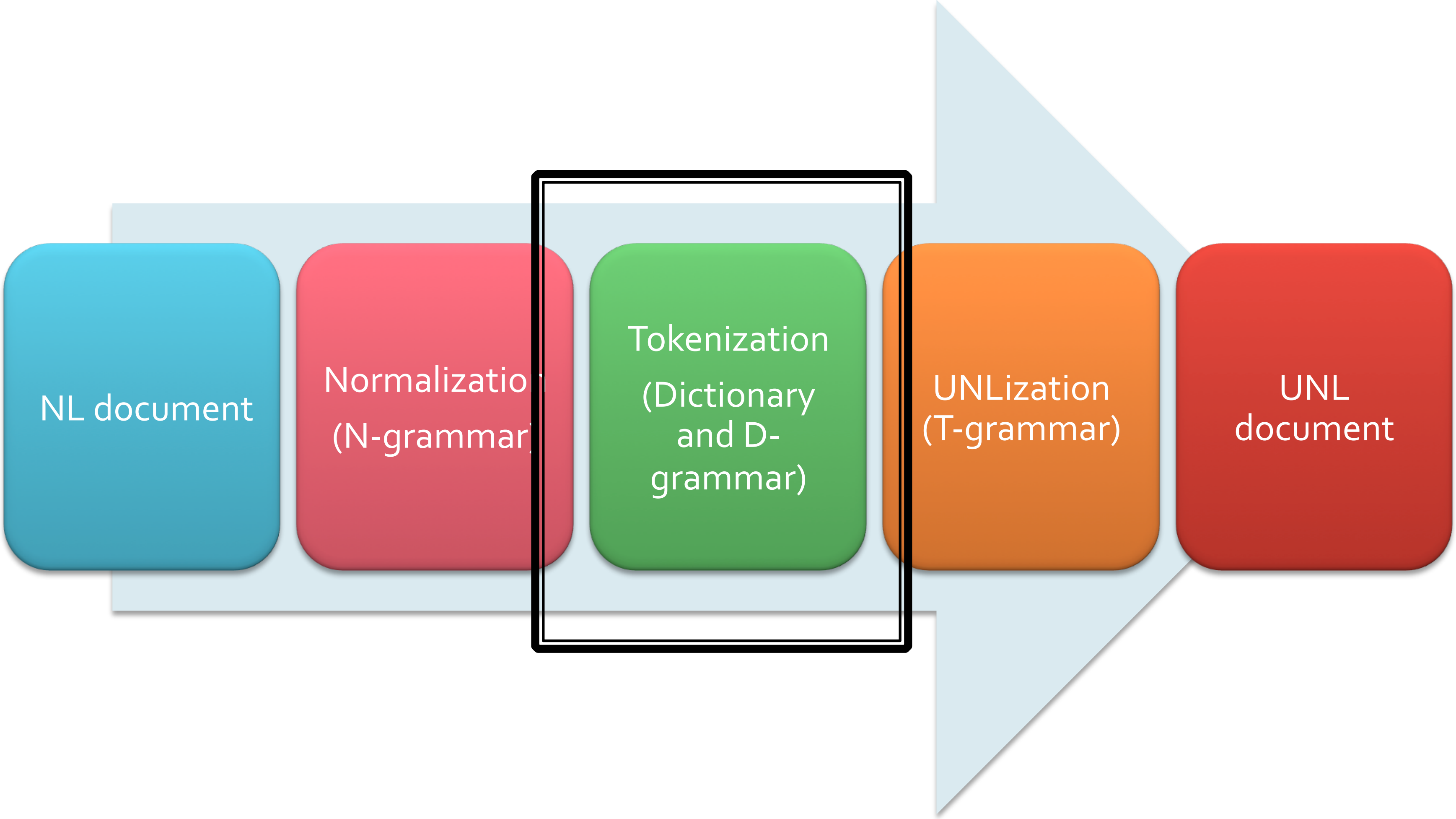
Exercise #2 (60 min)

- Goal: To prepare the N-grammar for the training corpus
- Deliverables:
  - *ngrammar_<ID>.txt*
  - *corpus_seg_<ID>.txt*
- Activities:
  - www.unlweb.net/wiki/XV_UNL_School

Lunch Break

Tokenization

Goal


NL document
Normalization
(N-grammar)
Tokenization
(Dictionary and D-grammar)
UNLization
(T-grammar)
UNL document

Tokenization
What does it mean?

- All human beings are born free and equal in dignity and rights. They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood.

- [All] [ ] [human beings] [ ] [are born] [ ] [free] [ ] [and] [ ] [equal] [ ] [in] [ ] [dignity] [ ] [and] [ ] [rights] [.] [ ] [They] [ ] [are] [ ] [endowed] [ ] [with] [ ] [reason] [ ] [and] [ ] [conscience] [ ] [and] [ ] [should] [ ] [act] [ ] [towards] [ ] [one another] [ ] [in] [ ] [a] [ ] [spirit] [ ] [of] [ ] [brotherhood] [.]

Tokenization

How is this done?

1. The system matches first the longest entry in the dictionary, from left to right
2. The highest frequent entry comes first in case of entries with the same length
3. The first to appear in the dictionary comes first in case of entries with the same length and same frequency
4. The feature TEMP (temporary) is assigned to the strings that were not found in the dictionary
5. The feature DIGIT is assigned to the strings exclusively formed by digits
6. The feature SHEAD is automatically assigned to the beginning of the paragraph, and the featuer STAIL is automatically assigned to the end of the paragraph
7. No other tokenization or segmentation is done by the system (e.g.: blank spaces and punctuation signs are not automatically recognized)

Tokenization example

- INPUT STRING: aaaaaaa
- Example #1:
  - DICTIONARY = [aaaa], [aaa], [aa], [a]
  - TOKENIZATION = [aaaa][aaa]
- Example #2:
  - DICTIONARY = [aaa], [aa], [a]
  - TOKENIZATION = [aaa][aaa][aa]
- Example #3:
  - DICTIONARY = [aa]
  - TOKENIZATION = [aa][aa][aa]{a}

Exercise #3 (30 min)

- Goal: To extract the word list from the training corpus
- Deliverable:
  - *wordlist_<ID>.txt*
- Activities:
  - www.unlweb.net/wiki/XV_UNL_School

Dictionary

Dictionary Specs

www.unlweb.net/wiki/dictionary

- 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 : (PLR:=”oo”:”ee”)
    - Replacement
      - <ATTRIBUTE>”:=”<SOURCE>”:”<TARGET> ;
      - PLR:="oo":"ee";
    - Left appending
      - <ATTRIBUTE>”:=<LEFT ADDITION>"<”<LEFT DELETION>;
      - NOT:=<"un";
    - Right appending
      - <ATTRIBUTE>”:=<RIGHT DELETION">"<RIGHT ADDITION>;
      - PLR:="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.

Example

English Dictionary

```
;DETERMINERS

;ARTICLES (POS=ART)
[a]{}""(LEMMA=a,LEX=D,POS=ART,att=@indef)<eng,0,0>; in a spirit of brotherhood
[the]{}""(LEMMA=the,LEX=D,POS=ART,att=@def)<eng,0,0>; all the rights

;DEMONSTRATIVE DETERMINERS (POS=DEM)(not to be confounded with DEMONSTRATIVE PRONOUNS)
[other]{}""(LEMMA=other,LEX=D,POS=DEM,att=@other)<eng,0,0>; other opinion, other status
[this]{}""(LEMMA=this,LEX=D,POS=DEM,NUM=SNG,att=@proximal)<eng,0,0>; this declaration

;POSSESSIVE DETERMINERS (POS=POD)(not to be confounded with POSSESSIVE PRONOUNS)
[their]{}"00"(LEMMA=their,LEX=D,POS=POD,att=@3;@pl)<eng,0,0>; in all their forms

;QUANTIFIERS (POS=QUA)(not to be confounded with INDEFINITE PRONOUNS)
[all]{}""(LEMMA=all,LEX=D,POS=QUA,att=@all)<eng,0,0>; all human beings, all the rights
[any]{}""(LEMMA=any,LEX=D,POS=QUA,att=@any)<eng,0,0>; distinction of any
[no]{}""(LEMMA=no,LEX=D,POS=QUA,att=@no)<eng,0,0>; no distinction
```

Example (cont'd)

English Dictionary

```
;PRONOUNS (LEX=R)

;PERSONAL PRONOUNS (POS=PPR)
[it]{}"00"(LEMMA=it,LEX=R,POS=PPR,CAS=NOM,PER=3PS,att=@3)<eng,0,0>; whether it be independent
[they]{}"00"(LEMMA=they,LEX=R,POS=PPR,CAS=NOM,PER=3PP,att=@3;@pl)<eng,0,0>; they are endowed

;RECIPROCAL PRONOUNS (POS=CPR)
[one another]{}"00"(LEMMA=one another,LEX=R,POS=CPR,att=@reciprocal)<eng,0,0>; act towards one another

;INDEFINITE PRONOUNS (POS=NPR)
[everyone]{}"00"(LEMMA=everyone,LEX=R,POS=NPR,att=@every;@person)<eng,0,0>; act towards one another
[no one]{}"00"(LEMMA=no one,LEX=R,POS=NPR,att=@no;@person)<eng,0,0>; no one shall be subjected

;RELATIVE PRONOUNS (POS=RPR)
[which]{}"00"(LEMMA=which,LEX=R,POS=RPR)<eng,0,0>; territory to which a person belong
```

Example (cont'd)

English Dictionary

```
;NOUNS
[article]{}"article"(LEMMA=article,LEX=N,NUM=SNG,POS=NOU)<eng,0,0>; article 1
[birth]{}"birth"(LEMMA=birth,LEX=N,NUM=SNG,POS=NOU)<eng,0,0>; birth status
[brotherhood]{}"brotherhood"(LEMMA=brotherhood,LEX=N,NUM=SNG,POS=NOU)<eng,0,0>; spirit of brotherhood
[colour]{}"colour"(LEMMA=colour,LEX=N,NUM=SNG,POS=NOU)<eng,0,0>; such as colour
[conscience]{}"conscience"(LEMMA=conscience,LEX=N,NUM=SNG,POS=NOU)<eng,0,0>; reason and conscience
[country]{}"country"(LEMMA=country,LEX=N,NUM=SNG,POS=NOU)<eng,0,0>; country or territory
[declaration]{}"declaration"(LEMMA=declaration,LEX=N,NUM=SNG,POS=NOU)<eng,0,0>; this declaration
```

Any questions?

Exercise #4 (90 min)

- Goal: To prepare the dictionary for the training corpus
- Deliverable:
  - *dic_<ID>.txt*
- Activities:
  - www.unlweb.net/wiki/XV_UNL_School

That’s all, Folks!