XIV UNL School

Patras, March 10-14, 2014

Day #4

Morning

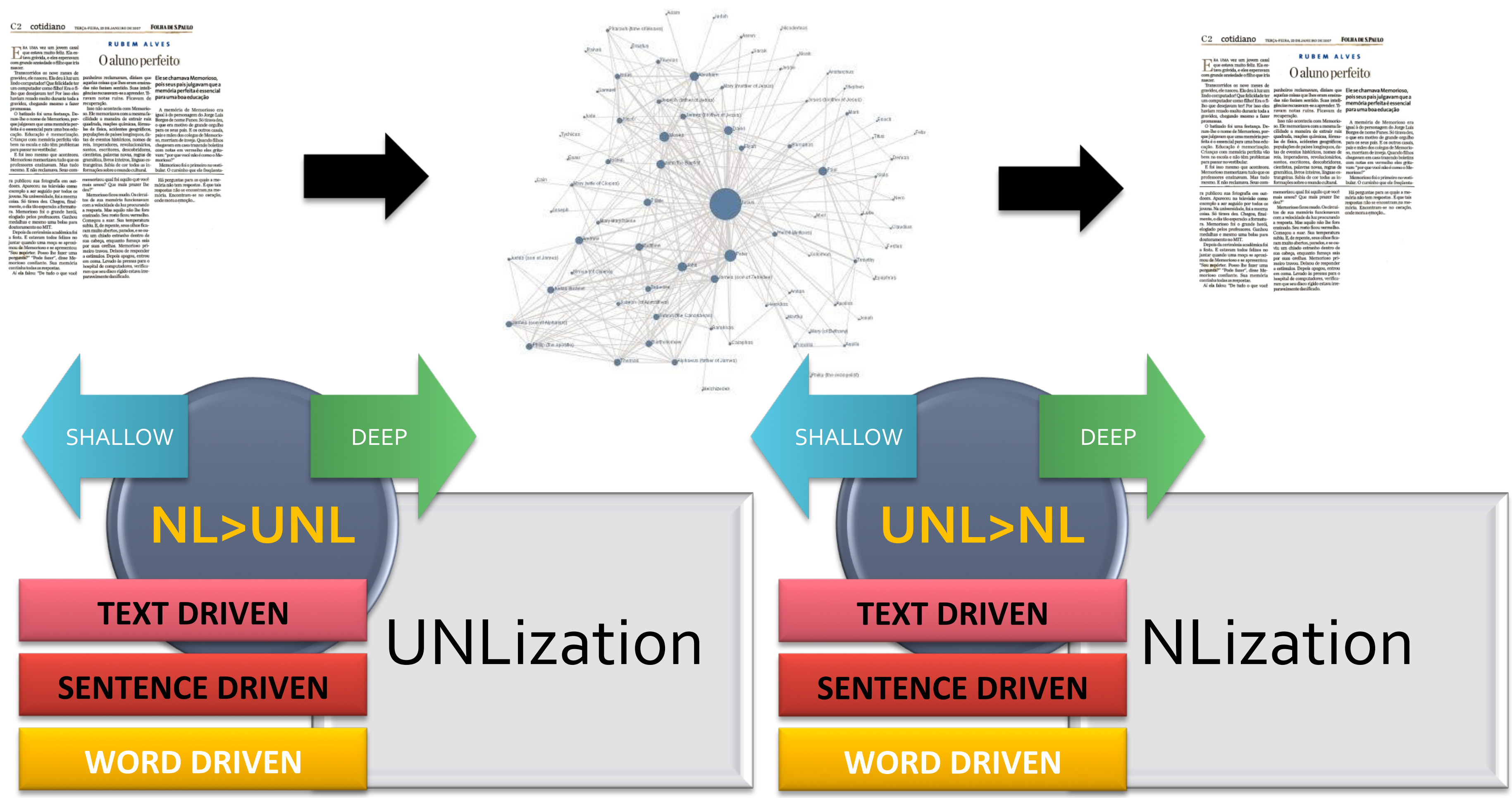
- UNLization

Lunch break

Afternoon

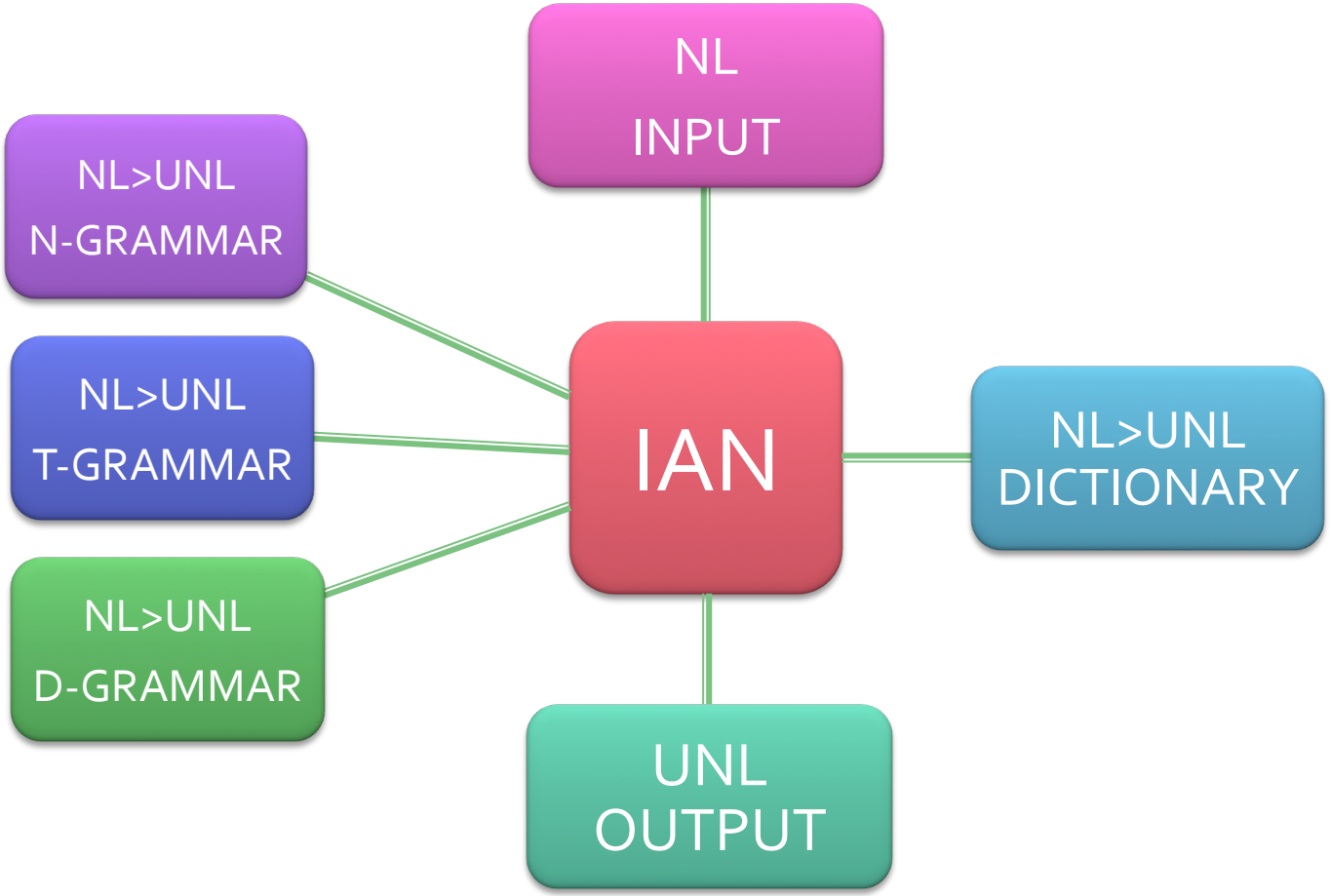
- NLization

The UNL System



UNLization

IAN (Interactive ANalizer)



Exercise #1

- Upload the original version of the corpus, available at the wiki, to IAN (at www.unlweb.net)

Normalization

Normalization
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
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 ");

Indexes

- Deletion
  - ~~(w1)(w2):=(w2);~~
  - (w1,%a)(w2,%b):=(%b);
- Inversion
  - ~~(w1)(w2):=(w2)(w1);~~
  - (w1,%a)(w2,%b):=(%b)(%a);

Normalization
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):=(%a)(%b,+STAIL);
  - (".",%a)(" ",%b)("/[A-Z]/",%c):=(%a)(%d,+SHEAD)(%c);

Exercise #1

- Create a N-grammar to normalize the following English text
  - He isn't coming tomorrow. He'll be in another meeting in N.Y. I'll check with him another date asap. Would u be available next week?
- Test it in IAN
  - Upload the text
  - Upload the grammar

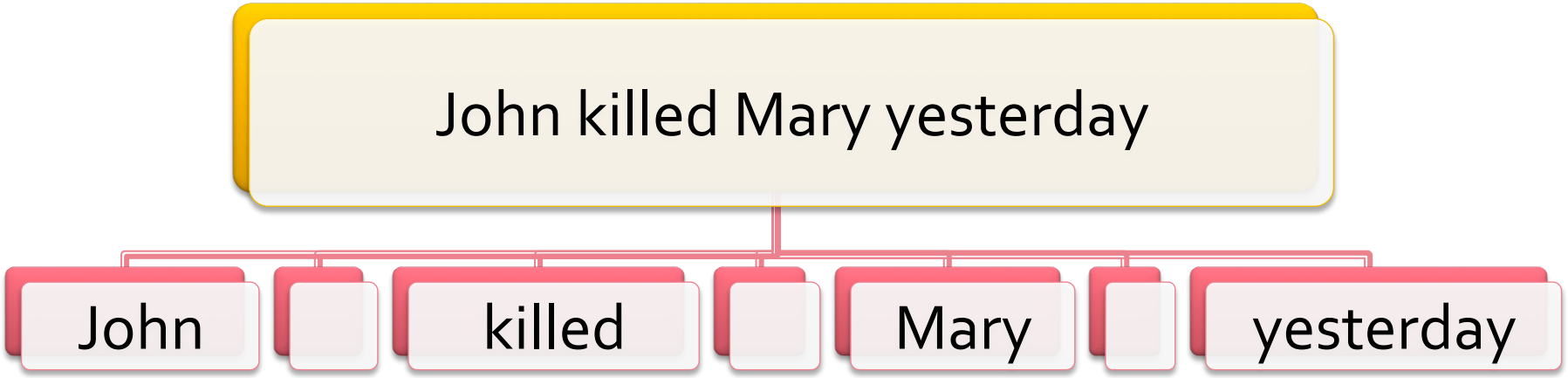
Exercise #2

- Create a N-grammar to normalize the corpus AESOP-3, available at the wiki.

Tokenization

Tokenization

- Longest first
- From left to right
- Can be controlled by d-rules



Example

- Dictionary
  - [a] {} "1" (A) <eng,100,0>;
  - [aa] {} "2" (B) <eng,50,0>;
  - [aaa] {} "3" (C) <eng,10,0>;
- Input
  - aaaaaaaa
- Result #1
  - [aaa][aaa][aa]
- D-Grammar
  - (C)(C)=0;
- Result #2
  - [aaa][aa][aaa]

Exercise #3

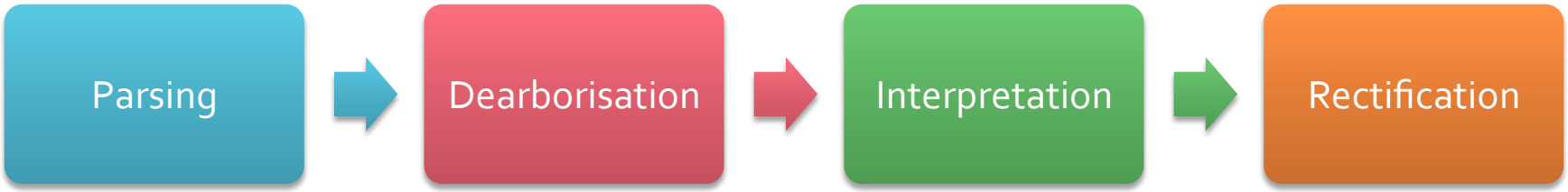
- Create a simplified dictionary to tokenize the sentence "The Hare and the Tortoise"

Exercise #4

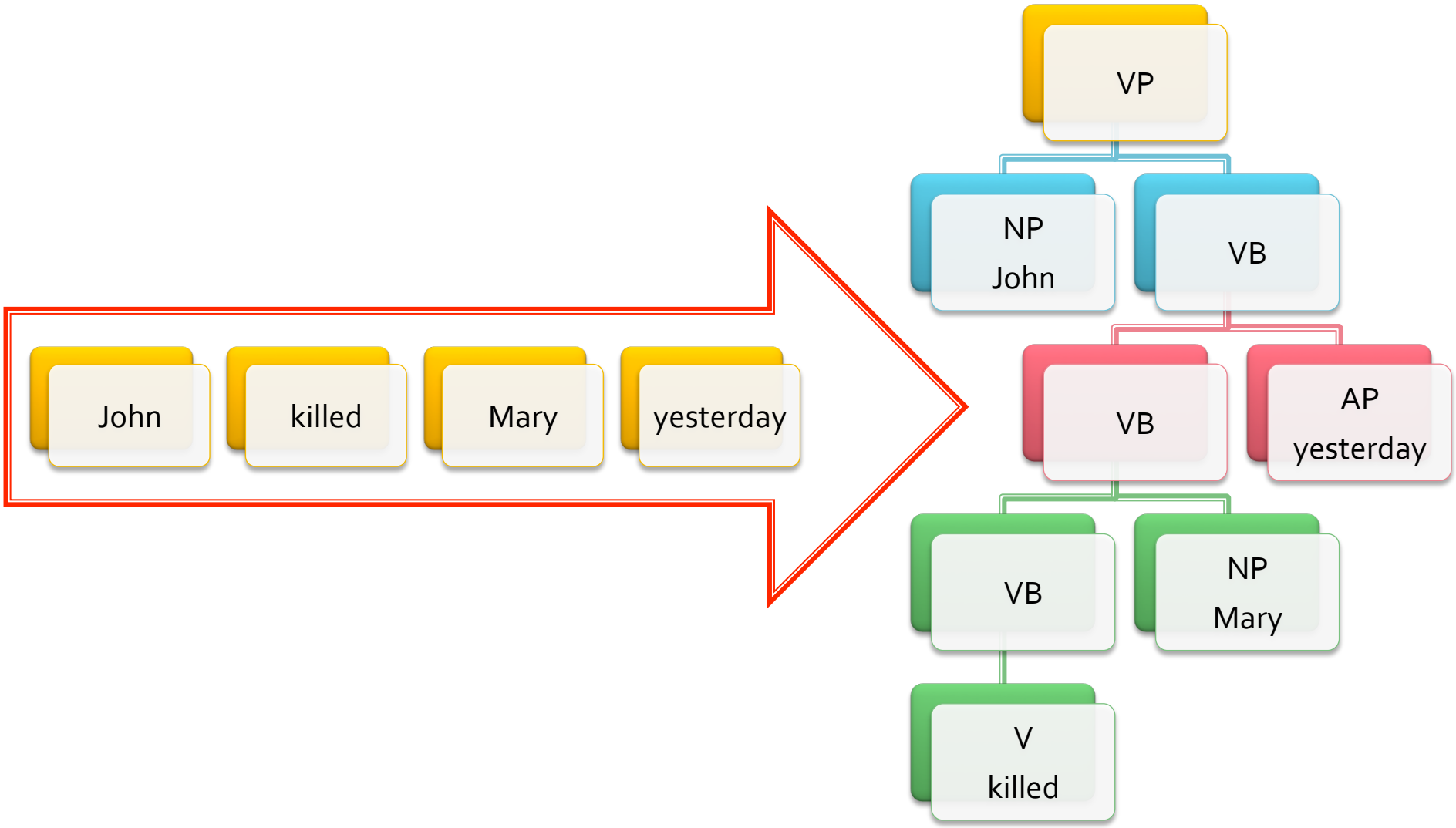
- Create a simplified dictionary to tokenize the sentence "The Hare one day ridiculed the short feet and slow pace of the Tortoise"

Transformation

Transformation



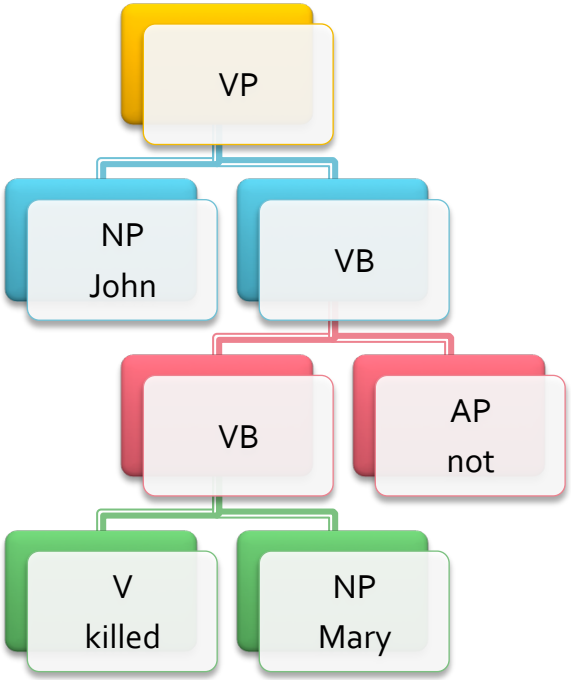
Parsing
list > tree



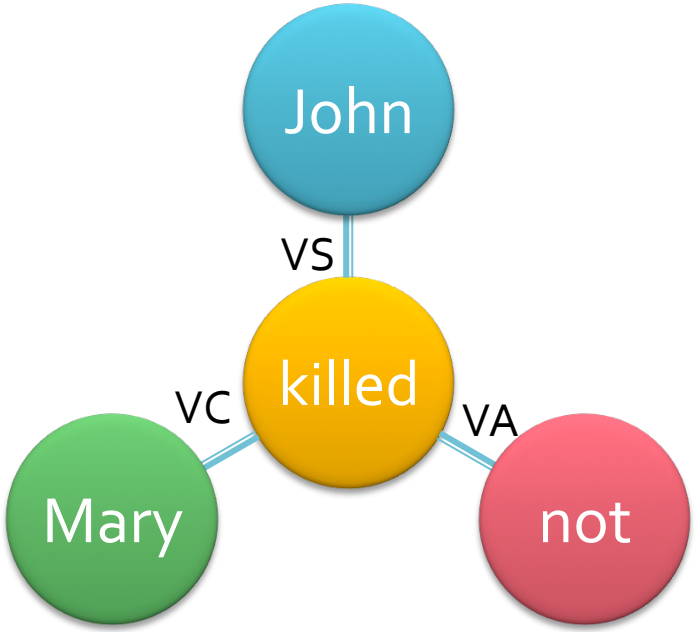
Dearborisation

syntactic tree > syntactic network

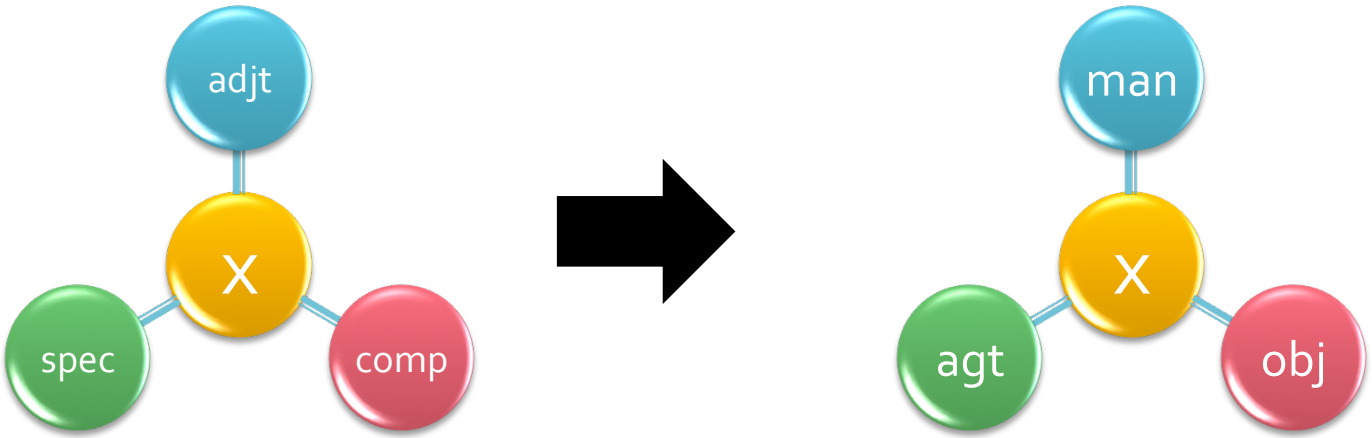
TREE STRUCTURE



NETWORK STRUCTURE



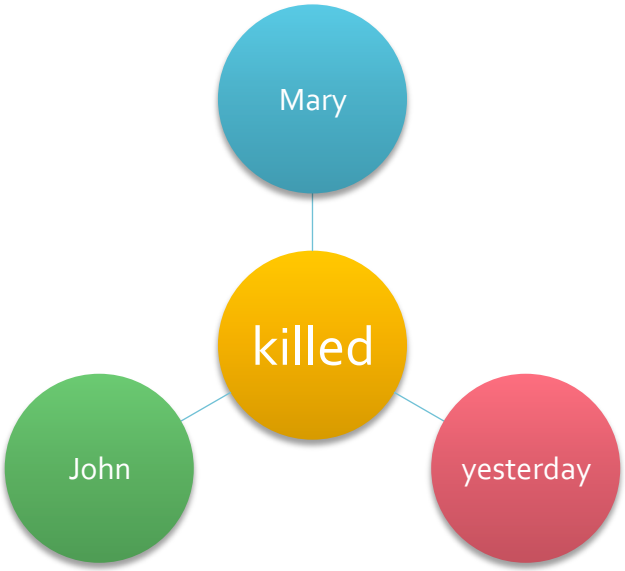
Interpretation



Rectification

- Post-processing

Shallow Processing



Example #1

- The Hare
  - ([the],%x)([hare],%b):=(%b,+@def);
    - the hare > hare.@def
  - (D,%x)(N,%b):=(%b,+@def);
    - [the]{}""(D)<eng,o,o>;
    - [hare]{}"hare(icl>mammal)"(D)<eng,o,o>;
    - ~~a hare > hare.@def~~
  - (D,DEF,%x)(N,%b):=(%b,+@def)
    - [the]{}""(D,DEF)<eng,o,o>;
- HOWEVER
  - [the][hare] != [the][ ][hare]
    - ([ ]):=;
    - (BLK):=;
      - if [ ]{}""(BLK)<eng,o,o>;

Example #2

- slow progress
  - [slow]{}"slow(aoj>speed)"(J)<eng,0,0>;
  - [progress]{}"progress(icl>motion)"(N)<eng,0,0>;
  - (J,%x)(N,%y):=mod(%y;%x);

Example #3

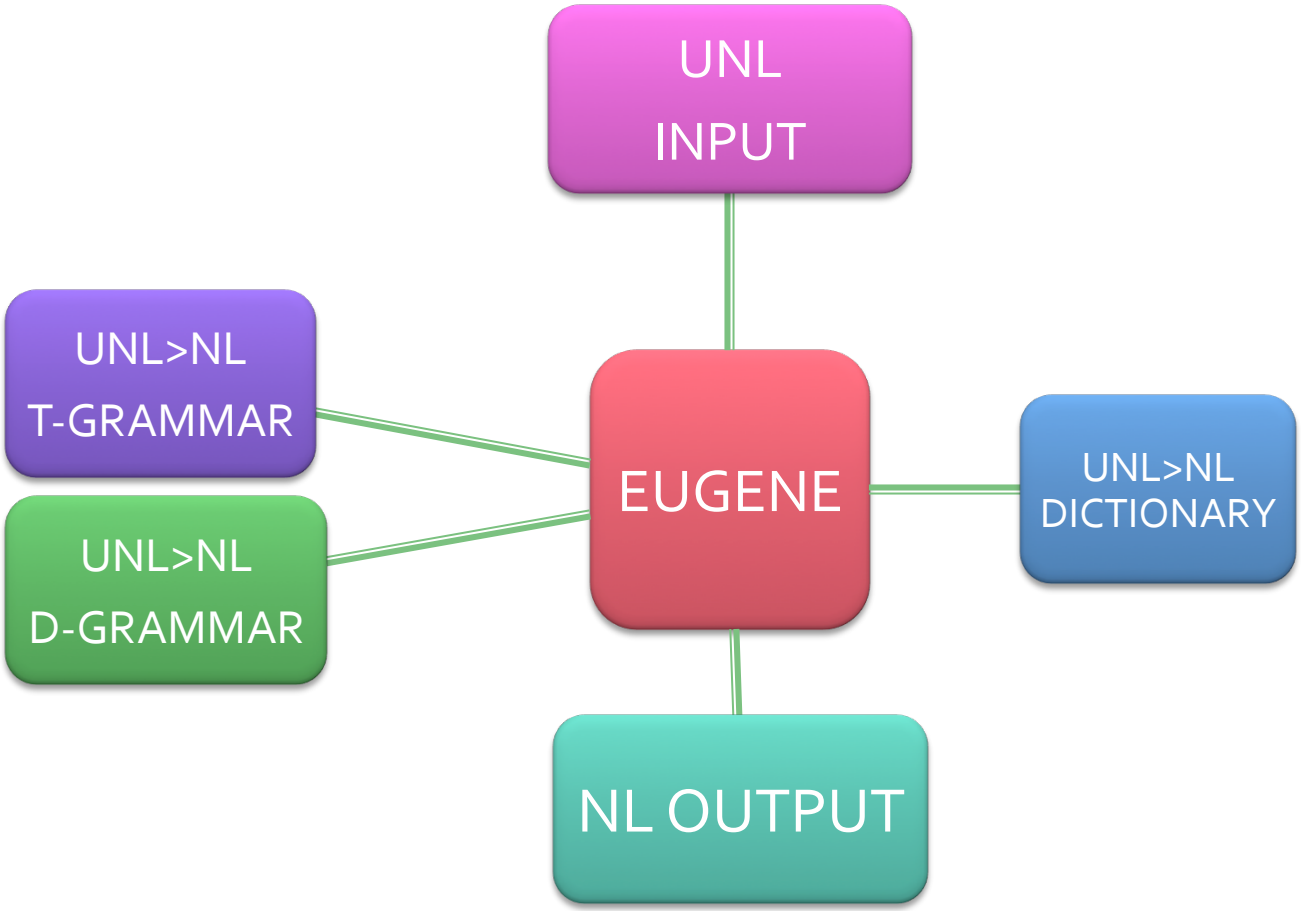
- The short feet
  - DICTIONARY
    - [ ]{}""(BLK)<eng,o,o>;
    - [the]{}""(D,DEF)<eng,o,o>;
    - [short]{}"short(aoj>length)"(J)<eng,o,o>;
    - [feet]{}"foot(pof>animal)"(N,PLR)<eng,o,o>;
  - GRAMMAR
    - (BLK):=;
    - (J,%x)(N,%y):=mod(%y;%x);
    - (D,@def,%x)(N,%y):=(%y,+@def);

Exercise #5

- Create the grammar to UNLize the corpus Aesop-3, available at the wiki.
  - Upload the corpus AESOP-3 to IAN
  - Provide the necessary dictionary entries, if they are not available yet.

NLization

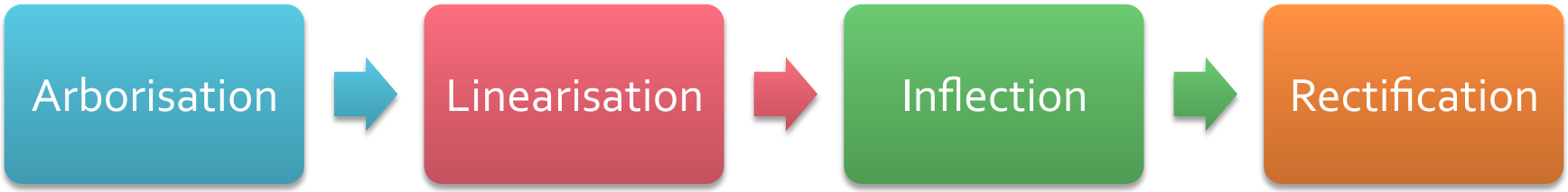
EUGENE
(dEep-to-sUrface GENErator)



Exercise #6

- Upload the corpus Aesop-6, available at the wiki, to EUGENE (at www.unlweb.net)

Transformation

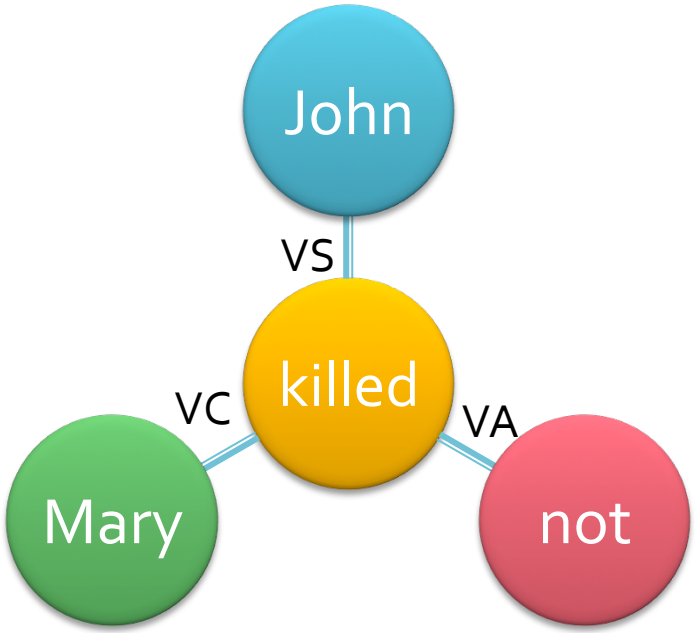


Arborisation (I)

semantic network > syntactic network

SEMANTIC NETWORK

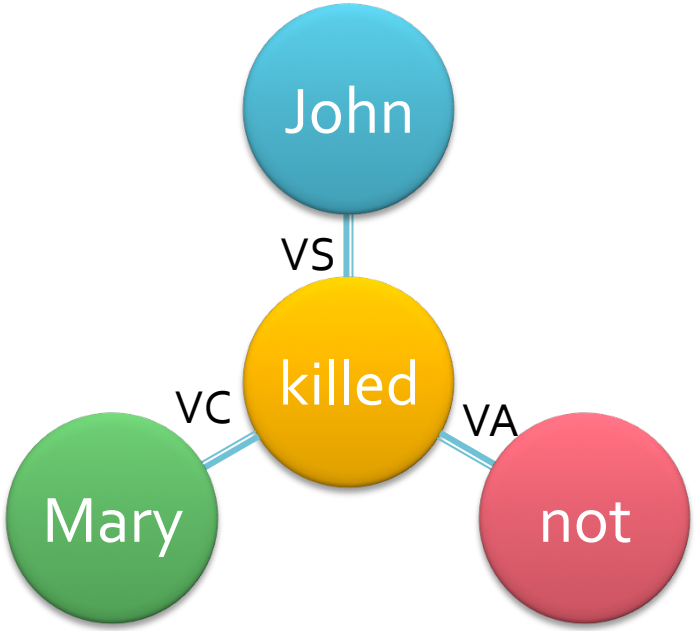
SYNTACTIC NETWORK



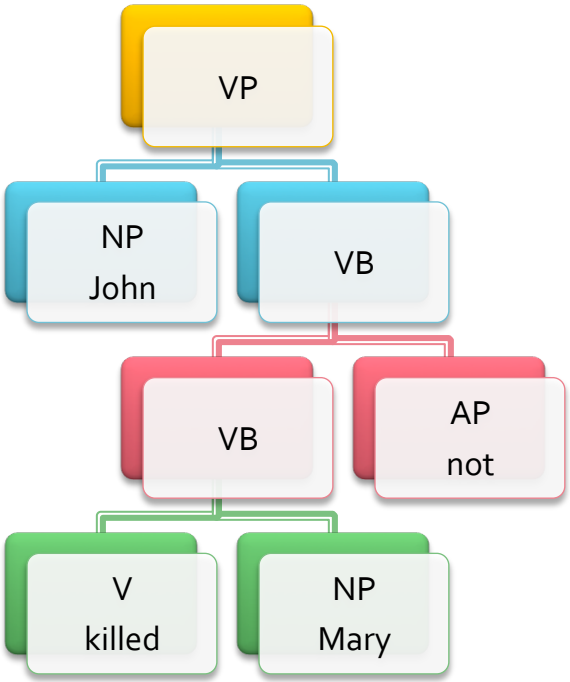
Arborisation (II)

syntactic network > syntactic tree

NETWORK STRUCTURE



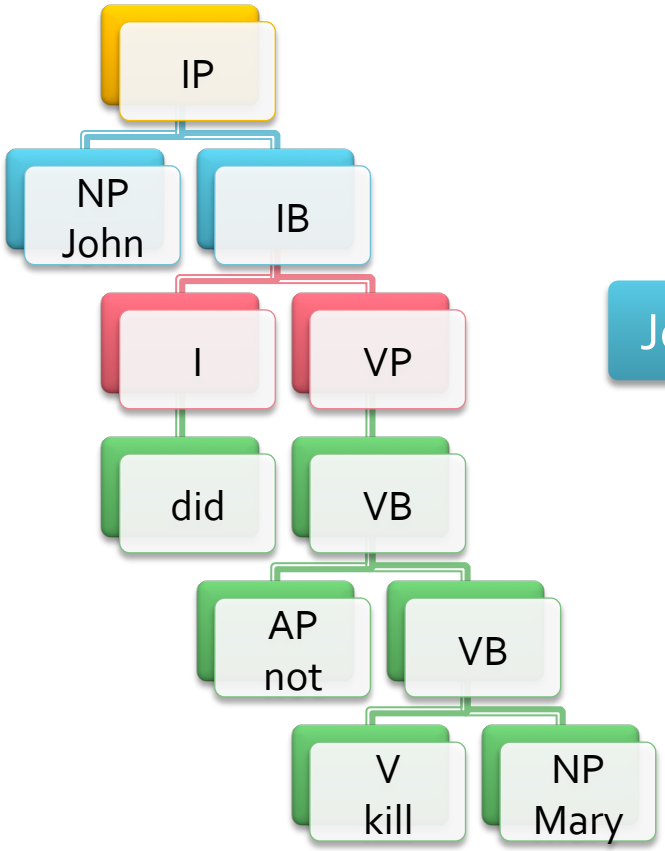
TREE STRUCTURE



Linearization

tree > list

TREE STRUCTURE



LIST STRUCTURE

Inflection

- Inflectional Grammar
  - (%x,M2):=(%x,-M2,+FLX(SNG:=o>"";
    PLR:=o>"s";));
- Triggering inflectional rules
  - (%x,^inflected,FLX):=(%x,!FLX,+inflected);

Rectification

- Eliminate excessive blank spaces
  - (BLK,%x)(BLK,%y):=(%x);
- Punctuation
  - (^PUT,%x)(STAIL,%y):=(%x)([.])(%y);
- Contraction
  - ([de])(BLK)([le]):=([du]);
  - ([a])(BLK,%x)("/[aeiou].*",%y):=([an])(%x)(%y);

Example 1

- hare.@def
  - (%x,@def):=([the])(" ")(%x);
    - [the][ ][the][ ][the][ ]…
  - (%x,@def):=([the])(" ")(%x,-@def);
    - hare.@def > the hare
  - Agreement
    - (%x,N,MCL,@def):=([le])(" ")(%x,-@def);
    - (%x,N,FEM,@def):=([la])(" ")(%x,-@def);
    - …
    - (%x,N,GEN):=(?D,?DEF,?[ ],?NUM=%x)(" ")(%x,-@def);

Example 2

- mod(pace, slow)
  - mod(%x;%y):=(%y)(" ")(%x);

Example 3

- mod(pace.@def, slow)
  1. mod(%x;%y):=(%y)(" ")(%x);
  2. (%x,@def):=([the])(" ")(%x);
     - [slow][ ][pace]
     - [slow][ ][the][ ][pace]

Exercise #6

- Create the grammar to NLize the corpus AESOP-4, available at the wiki.