Geneva, July 2nd

XII UNL School

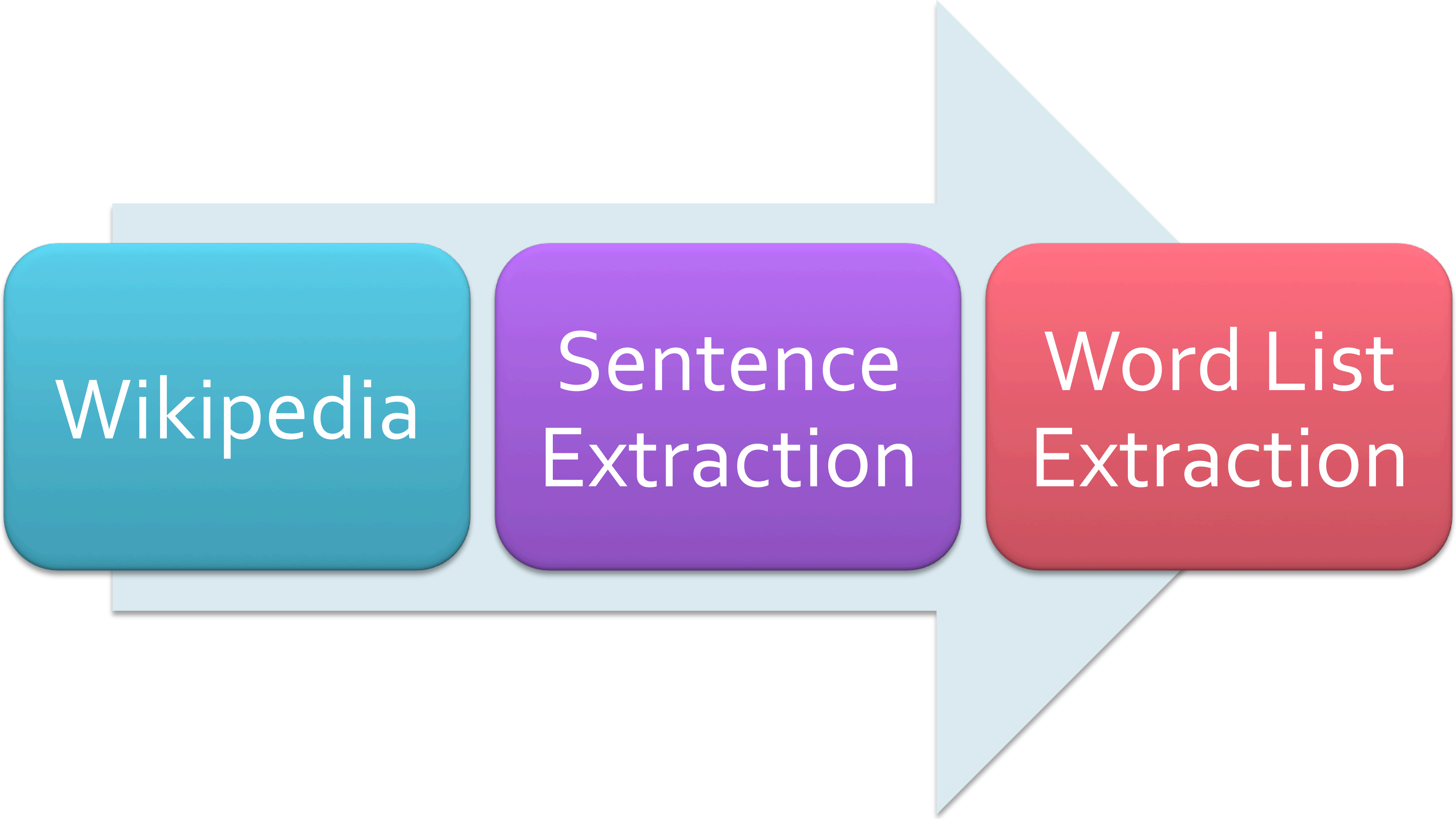
Day #3

Day #2

- ~~Welcome~~
- ~~Context~~
- ~~Normalization Grammar~~
- ~~Closed-Class Dictionary~~
- Open-Class Word List
- Corpus

Open-Class Word List

Open-Class Word List

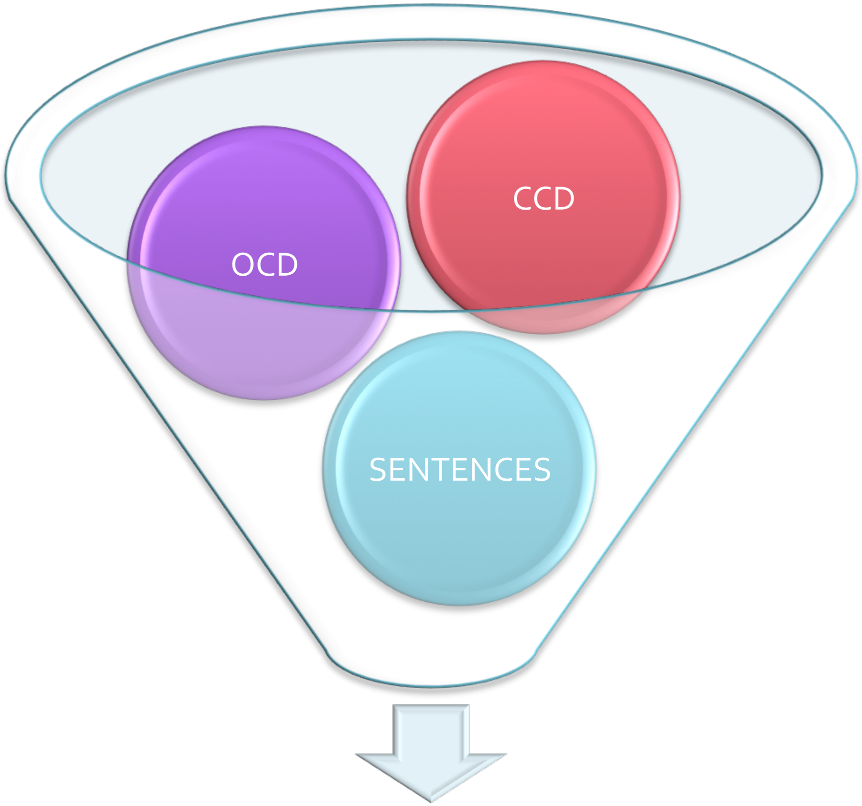


Wikipedia

Sentence Extraction

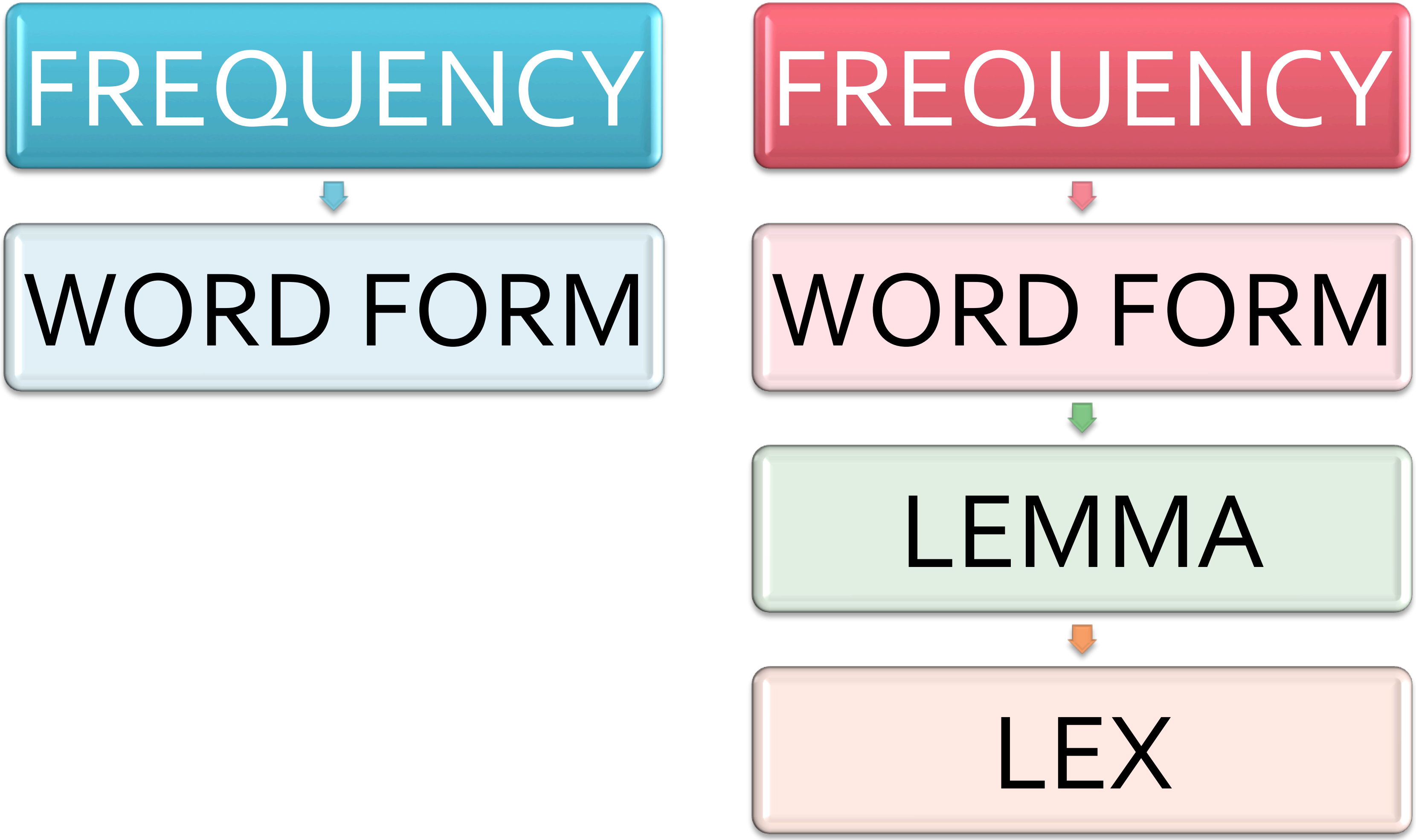
Word List Extraction

Pattern Extraction



SSS

Open-Class Word List



FREQUENCY

WORD FORM

FREQUENCY

WORD FORM

LEMMA

LEX

Exercise #4

- Donwload the word list available at the Wiki
- Provide, for the word forms, the lemma and the LEX
  - Observations
    - Leave the lemma blank in case of lemma = word form
    - Leave the LEX blank in case of wrong entries
    - Duplicate homonyms, but only if they are reasonably frequent
- Send your data to r.martins@undlfoundation.org in order to be included in the BRUNO-A1 project for your language.

Corpus

Exercise #5

1. Select 5-10 articles in the Wikipedia of your native language that refer to culture-specific subjects (plants, animals, dishes, rituals, festivities, etc.). The five 5 articles must correspond to around 500 sentences.
2. Convert the articles to plain text format and upload them to IAN.
3. Use your segmentation grammar to segment them.
4. Export the segmented text.

Exercise #6

1. Extract, from the segmented text, manually, all the noun phrases. Exclude, from your data, any noun phrase including verbs (such as a relative causes), but keep those involving prepositions, adjectives, adverbs and other determiners
2. Send your data to r.martins@undlfoundation.org in order to be included in the NC-A1 project for your language.