CS463 – Natural Language Processing

Lexicons and Morphology

- **Lexicons**
  - **Lexicon**
  - **Lexicon Extraction**
  - **Evaluating Corpora (Zipf's Law)**
- **Morphology**
  - **Templatic Morphology**
  - **Concatenative Morphology**

Lexicon

- Restricted vocabulary of a (NLP) system
  - A list of all expected or allowed valid words.
- backbone of any NLP application.
- Generated:
  - Manually (many people)
  - With Computers (Today's trend)
    - Extract from corpora
    - Reduce (Stem)
    - Synthesized??
- Examples:
  - Bare (...عار)
  - With description

; conjunctions

و وَ Pref-Wa and
<pos>wa/CONJ+</pos>

ف فَ Pref-Wa and;so
<pos>fa/CONJ+</pos>

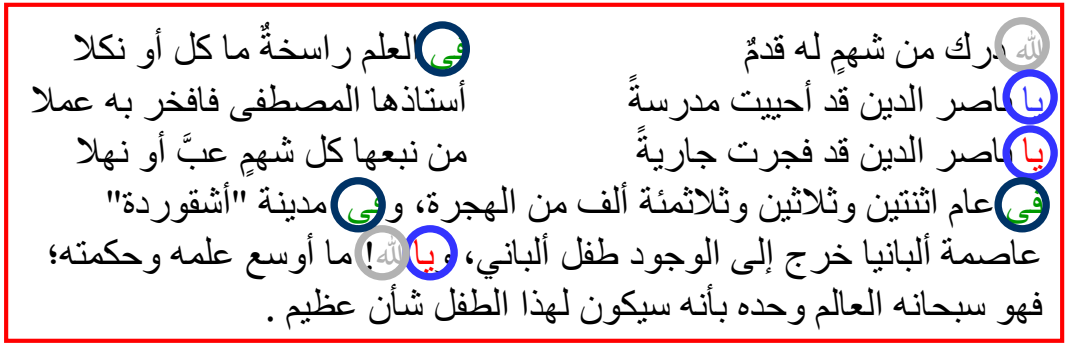
One
Two
Three
Four
Five
Here
Mars
Days
Name
Go

Lexicon Extraction

- Computational-linguistic community is converging to extract the lexicon from naturally used text (newspaper, phone call).
- A large amount of representative text is gathered and processed (*Corpus*).
- Typically involves normalizing surface-words into a common basic form (e.g. roots or stems)
  - Reduce the number of entries.
  - Need Morphology!

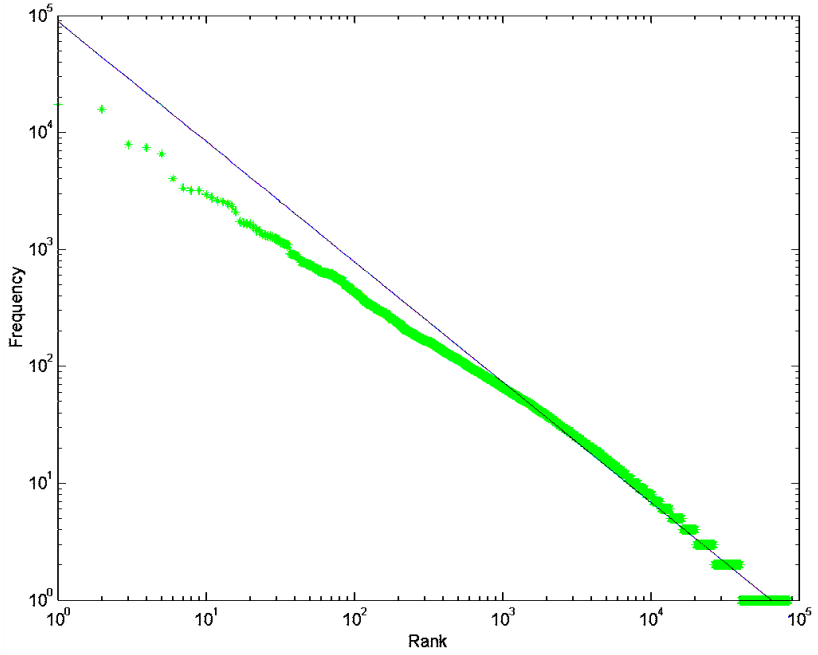
Lexicon Extraction

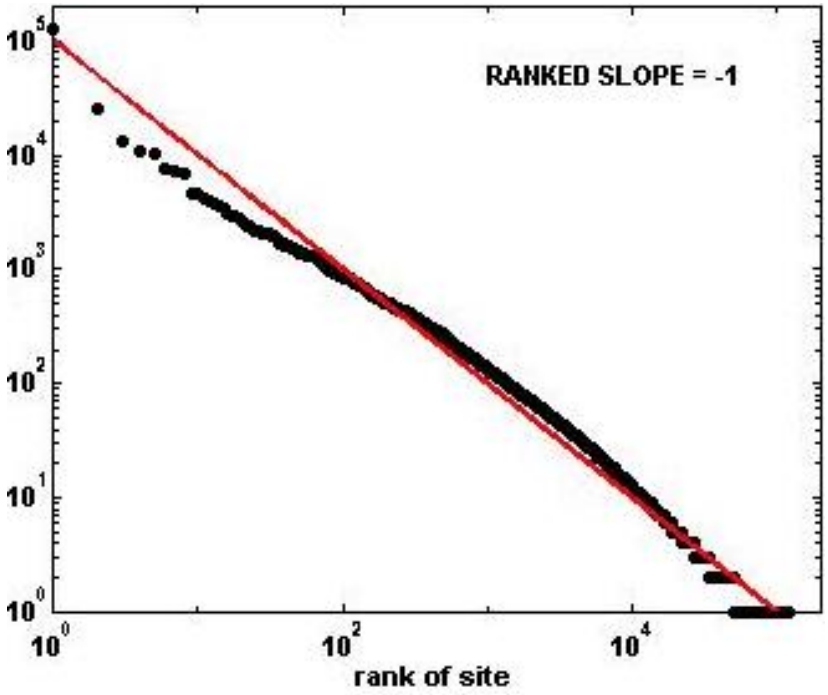
- Corpus:
  - *pl. corpuses or corpora*
  - A very large amount of NL representative text.
  - Typically (but not exclusively) from newspapers.
- Pros:
  - Capture the frequencies of NL. (Utterances.)
- Cons:
  - Never complete.
  - Typos.
- Example
  - KSUCCA (King Saud University Corpus of Classical Arabic)



Evaluating Corpora using Zipf's law

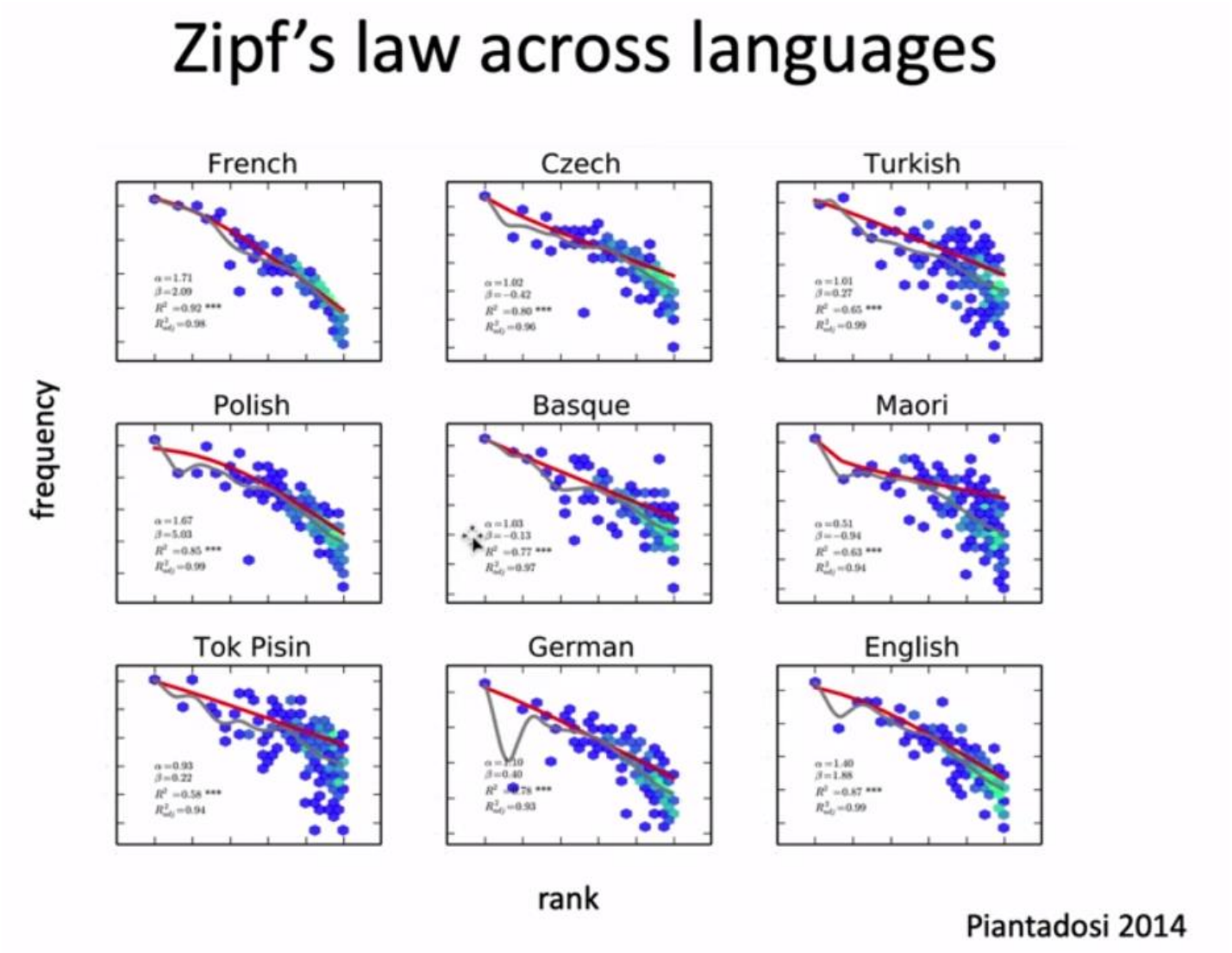
- Empirical law
  - Measures corpus quality
  - Theory
- f × r = k.





•log-log plot

Evaluating Corpora using Zipf's law

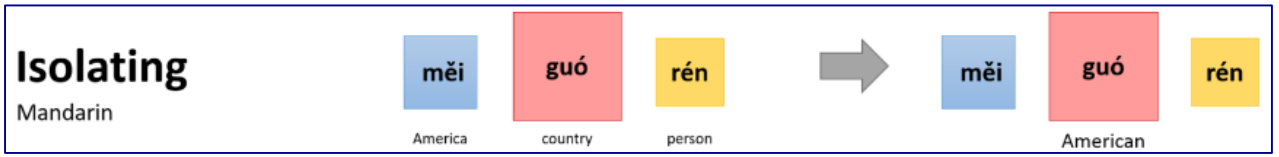


Morphology

- The (grammatical) study of the (internal) structure of words.
- The **internal structure** of words and the **segmentation** into different kinds of **morphemes** are two basic purposes of morphology.
- A **morpheme** is defined as the minimal meaningful unit of a language.

Morphology

- Types of Language Morphology (by August Schleicher):
  1. **Analytic (Isolating) –** Analytic languages have a low ratio of morphemes to words. They are isolating languages because each morpheme represents a word and vice versa. For example:



THOMPSON RIVERS UNIVERSITY

  2. **Concatenative (Agglutinative) –** such languages combine one or more morphemes into one meaningful word to give it different meanings.
     - Prefix: informal , سيذهب
     - Suffix: formalize , ذهبوا
     - Circumfix: informalize , يذهبان
  3. **Templatic (Fusional) –** combine morphemes to modify meanings by fusing them.
     - Root: mouse , ذهب
     - Pattern (infix): mice , ذاهب

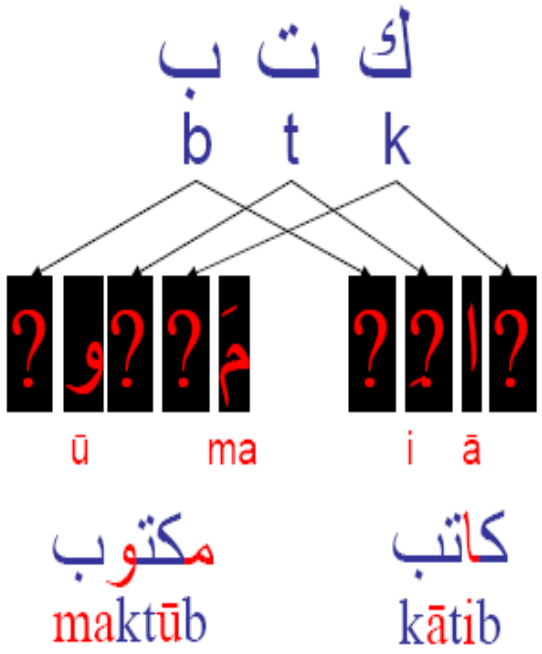
Templatic Morphology

- Root – is primary lexical unit of a word.
  - Carries semantic content.
  - Cannot be reduced.
  - Left when all, including internal, morphologically added structure has been wrung out.
  - In Arabic:
    - An ordered sequence of 3, 4, or 5 letters.
    - bare verb.
- Pattern - a sequence of constant and variable characters.
  - Variable characters: (ف، ع، ل) = (1, 2, 3).
    - To be substituted by the letters of the Arabic root.

Templatic Morphology - Example

- Starts from Roots & Patterns
- Examples:

- Root
- Pattern
- Lexeme

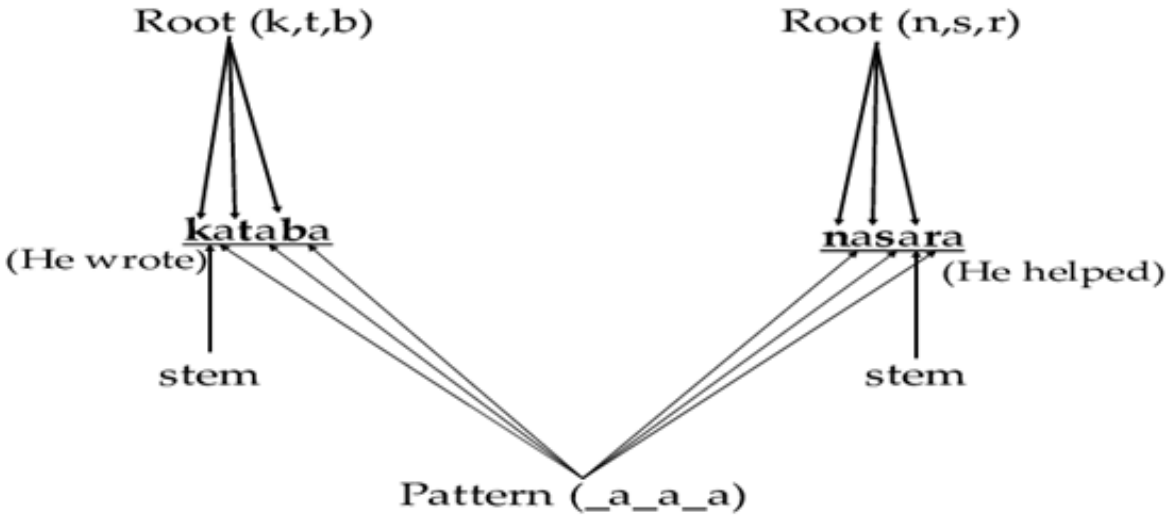


Templatic Morphology – Another example

- Consider the following data.
- 11) *Arabic* (*McCarthy* 1979:244; 1981:374)
- katab ‘wrote’
- kattab ‘causes to write’
- kaatab ‘corresponded’
- ktatab ‘wrote, copied’
- kutib was written (perfective passive)
- All above data share the consonants ***ktb***
- All the active verb forms have the vowel ***a***
- What we find in Arabic is called **templatic** or **root and pattern** morphology.

Templatic Morphology – Another example

Template	Stem	Meaning
CCC – فعل	ktb – كتاب	books, wrote, etc.
mCCwC – مفعول	mktwb – مكتوب	something written
CCAC – فعال	ktAb – كتاب	book
CCACyC – فعاعيل	ktAtyb – كتاتيب	Qur'an school
CACC – كاتب	kAtb – كاتب	writer
CcwC – فعول	ktwb – كتوب	skilled writer

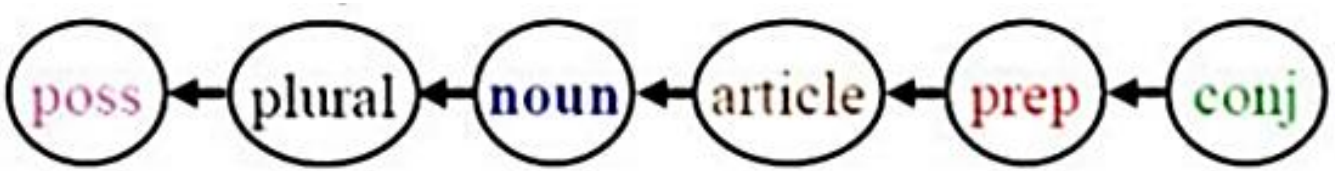


Concatenative Morphology

- Starts from stems.
- Minimal surface-form
  - Nouns, Verbs, & Particles.
  - But not all surface-words are stems.
  - Roots + Patterns
    - `Root.GeneralMeaning + Patten.specificMeaning`
    - Only further *Circum*fixation allowed
      - No further *in*fixation.

Concatenative Morphology - Examples

- Noun Examples:



وللمكتبات	وكبيوتنا
/walilmaktabāt/	/wakabiyūtinā/
و+ل+ال+مكتبة+ات	و+ ك + بيوت + نا
wa+li+al+maktaba+āt	wa+ka+biyūt+nā
and+for+the+library+plural	and+like+houses+our
And for the libraries	*And like our houses*

Concatenative Morphology - Examples

- Verb Examples:

وسنقولها	فقلناها
/wasanaqūluhā/	/faqulnāhā/
و+ س+ ن+ قول + ها	ف+ قال+ نا+ ها
wa+sa+na+qūl+u+hā	fa+qul+na+hā
and+will+we+say+it	so+said+we+it
And we will say it	*So we said it.*

Concatenative vs Non-concatenative Morphology

- **Concatenative** morphology refers to attaching morphemes to one another. In other words, affixes attach to bases. It is a linear process (e.g. un-accept-able).
- **Non-concatenative** morphology refers to a non linear process, where you don't attach different morphemes together. Internal change (drink – drank) and suppletion (go – went) are examples of non-concatenative process.

- Exercise:
  - Compare between Concatenative and Templatic Morphology.
  - How does Non-concatenative morphology relate to Templatic morphology?
  - Give examples for both questions. Consider Buckwalter's AraMorph.