بەرەو API یەک بۆ پرۆسسکردنی تێکستی کوردی

Towards an Application Programming Interface (API) for Processing Kurdish Text

Dr. Abdul-Rahman Mawlood-Yunis

PhD from the School of Computer Science,

Carleton University,

Ottawa, Ont., Canada

armyunis@scs.carleton.ca

Outline

- Motivation
- Environment setup
- Character coding , read and write files
- Kurdish text processing operations
- Applications
- Conclusion
- Future work
- Promising Computer study trends for Kurdistan region

Motivation

- بۆ ئەوەی بەشێوەیەکی سەرکەوتووانە کۆمپیتەر بە زمانی کوردی بەکاربێنن لە ژیانی ڕۆژانە (بۆ نموونە، حکوومەت، بازرگانی، لێکۆلینەوە) ئەوا پێوستمان بە API یەك هەیە بۆ پرۆسسکردنی تێکستی کوردی

- In order to use computers successfully in our daily life (e.g., business, government and research) we need an API for Kurdish text processing

- بە دەربرینێکی تر، هەبوونی API یەك بۆ پرۆسسکردنی تێکستی کوردی دەرگا دەکاتەوە بۆ دروستکردنی کۆمپیتەر ئەپلیکەیشن بە بێ ئەژمار.

- An API for Kurdish text processing will open up doors for unlimited number of applications

- یارمەتی بە ستانداركردن و ڕێخستنی ڕێنماکانی نووسینی کوردی دەکات

- Assists in standardizing Kurdish Language and Kurdish writing

Outline

- Motivation
- Environment setup
- Character coding , read and write files
- Kurdish text processing operations
- Applications
- Conclusion
- Future work
- Promising Computer study trends for Kurdistan region

Eclipse setup for Kurdish text display

بۆ ئەوەی بتوانین بە زمانی کوردی بنووسین پێویستمان بە کۆدینگێکە
(Coding) کە پیتی کوردی پێ بنووسرێ.
UTF-8 دەتوانرێت بۆ ئەم مەبەستە بە کاربێت.

C:\Users\Rahman\workspace>java Slaw

???? ???????? (**کوردی**)

C:\Users\Rahman\workspace>java Slaw

Hello World (**English**)

Eclipse setup

1. Run → Run configuration → common tab → select utf-8 coding
2. Go to Eclipse -> Preferences -> General -> Appearance -> Colors and Fonts -> Debug -> Console font
3. Control Panel\System and Security\System → advance system settings → Environment variable → create new user variable

   JAVA_TOOL_OPTIONS: -Dfile.encoding=UTF8

Eclipse javadoc with utf-8

- *JavaDoc setup (to enter comments: shift-alt-J)*

  project → generate javadoc in configuration choose javadoc.exe

  for example:

  C:\Program Files\Java\jdk1.7.0_04\bin\javadoc.exe

- project-> javadoc -> next -> in extra vm options write **-encoding UTF-8 -charset UTF-8 -docencoding UTF-8**

- //readFileToList("C:\\Users\\Rahman\\workspace\\goran.txt");
- // WriteListToFileToColumn("C:\\Users\\Rahman\\workspace\\goran_out.txt") ;

Redirect console output to separate frame

- PipedInputStream pin=new PipedInputStream()
- PipedOutputStream pout = new PipedOutputStream(this.pin)
- System.*setOut(new PrintStream(pout, true))*
- *Catch Exceptions*

```
//    new RedirectConsoleOutput();
```

Redirect console output to file

- Run Configurations -> Common and in the Standard Input and Output choose File

- *Other integration environments include, NetBean, jEdit*

```
//KurdLangApi.count_words("C:\\Users\\Rahman\\workspace\\hawlati-24-6-2012\\z1.txt");
```

Outline

- Motivation
- Environment setup
- Character coding , read and write files
- Kurdish text processing operations
- Applications
- Future work
- Promising Computer study trends for Kurdistan region

Kurdish character in UTF-8 representation

- The extreme UTF-8 table

- Some special characters

  { 33, 34, 40, 41, 44, 45, 46, 47, 58, 95, 1548, 1563, 1567, 1569,
  1570, 1571, 1572, 1573, 1654, 8211, 8230, 61623, 65279 }

- Can be seen in the program debugging mode

//kurdishUnicodeCharValues() ;

Steps to Read and Write text file written in Kurdish

1. Reader reader = **new InputStreamReader(new FileInputStream(** "C:\\Users\\Rahman\\workspace\\h1.txt"), "**UTF-8**“)
2. fin = **new BufferedReader(reader)**
3. Writer writer = **new OutputStreamWriter(new FileOutputStream(** "C:\\Users\\Rahman\\workspace\\out1.txt"), "**UTF-8**")
4. BufferedWriter fout = **new BufferedWriter(writer)**
5. **while ((s = fin.read()) != -1) {**
   fout.write((**char)s)**
   }
6. fin.close()
   fout.close()

//ReadAndWriteFile();

Outline

- Motivation
- Environment setup
- Character coding , read and write files
- Kurdish text processing operations
- Applications
- Future work
- Promising Computer study trends for Kurdistan region

Kurdish text processing operations

- Counting words
  - *isSpace, isNumeric*
- Sorting words
  - System.*getProperty("line.separator")*
- cleaning words form noise

- The frequency use of ۆ in Kurdish writing

org.apache.commons.lang3.StringUtils jar file

```
// 1. KurdLangApi.count_words("C:\\Users\\Rahman\\workspace\\hawlati-24-6-2012\\z2.txt");  // isSpa
// 2.  readFileToList("C:\\Users\\Rahman\\workspace\\goran.txt");
// WriteListToFileToColumn("C:\\Users\\Rahman\\workspace\\goran_out.txt") ;  // line seprator
// 3.  KurdLangApi.remove_two_letter_words(fin, fout)
```

Outline

- Motivation
- Environment setup
- Character coding , read and write files
- Kurdish text processing operations
- Applications
- Future work
- Promising Computer study trends for Kurdistan region

Application
Most common words in Kurdish

Ex: English common words

Rank	Word	Rank	Word
1	the	11	it
2	be	12	for
3	to	13	not
4	of	14	on
5	and	15	with
6	a	16	he
7	in	17	as
8	that	18	dd
9	have	19	do
10	I	20	at

یەکەم ١٠٠ ووشەی ئینگلیزی

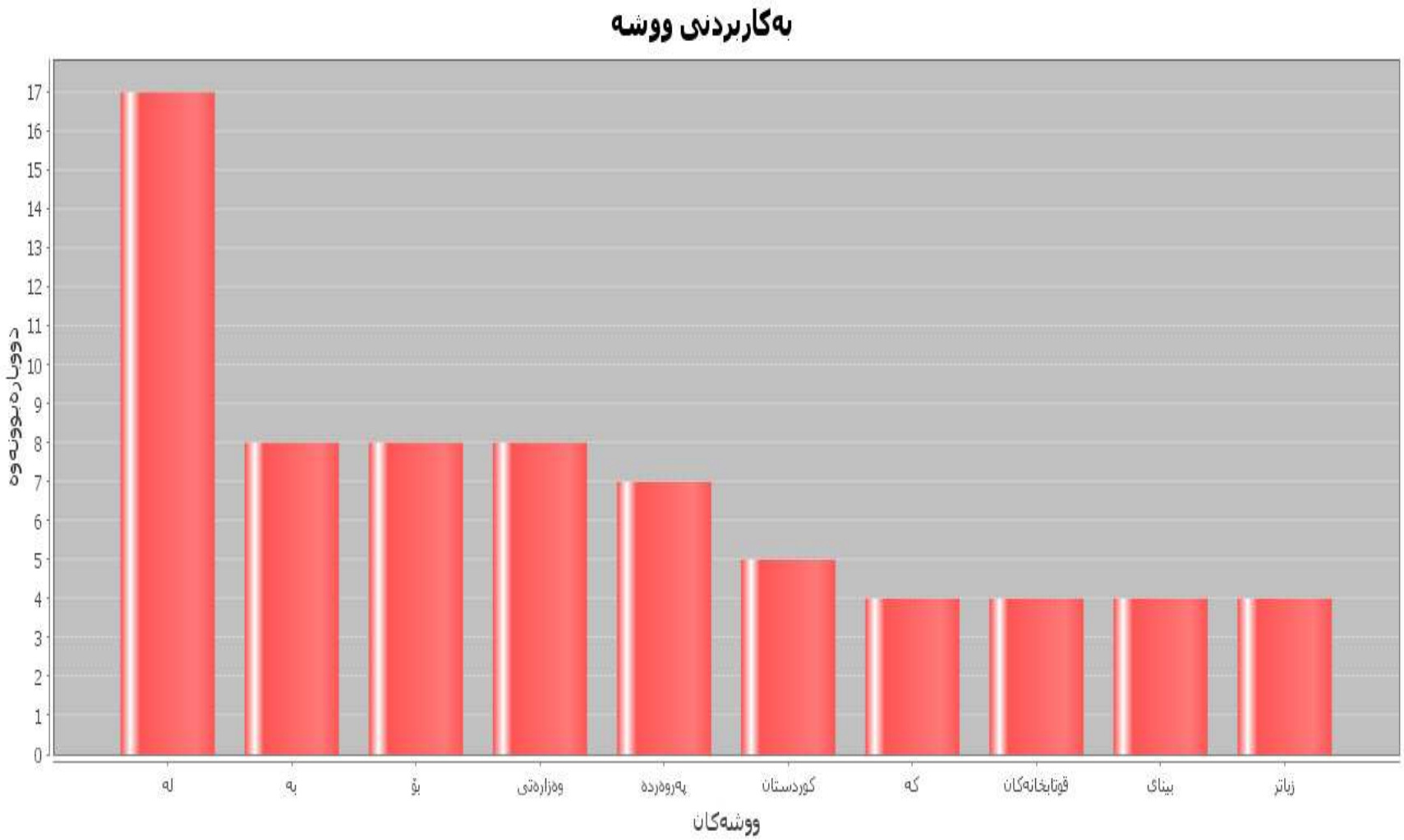
Example of common words continued

The **Teacher's Word Book** is an alphabetical list of the 10,000 words which are found to occur most widely in:

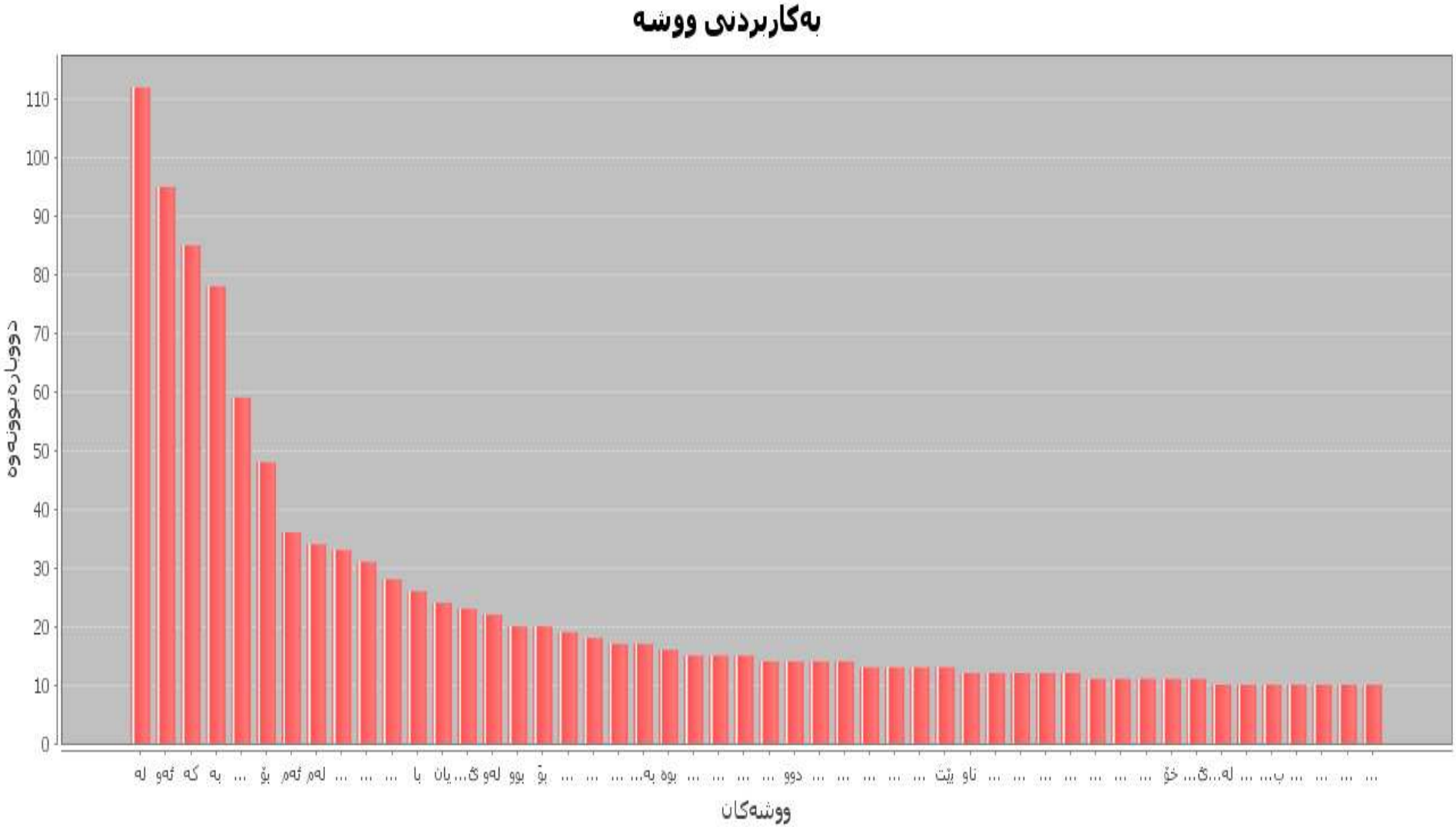
- 625,000 words from literature for children
- 3,000,000 words from the Bible and English classics
- 300,000 words from elementary-school text books
- 50,000 words from books about cooking, sewing, farming, the trades, and the like;
- 90,000 words from the daily newspapers

(Forty-one different sources were used)

نموونەیەک لە بەکاربردنی ووشەکان لە ووتاری کوردی



نموونەیەک لە بەکاربردنی ووشەکان لە ووتاری کوردی



Other application

- Spell checker
- Thesauri (e.g. word web)
- Crossword
- Unlimited application

Future work

- Extend the current work to a comprehensive API
  1. Number of lines in a text
  2. Number of paragraphs
  3. The longest and the shortest line or paragraph
  4. the average length
  5. Remove double space,

A course on natural language processing and Computational Linguistic

- Phonetics and Phonology —knowledge about linguistic sounds
- Morphology —knowledge of the meaningful components of words
- Syntax —knowledge of the structural relationships between words
- Semantics —knowledge of meaning
- Pragmatics — knowledge of the relationship of meaning to the goals and intentions of the speaker
- Discourse —knowledge about linguistic units larger than a single utterance

Thanks