Underneath the Hood of arabiCorpus

Dilworth Parkinson
Brigham Young University

What is arabiCorpus?

- Plain text corpus of Modern Arabic
  - not lemmatized
  - not POS tagged
  - Includes some colloquial and medieval material
- A web interface

Contents of arabiCorpus

- Newspapers (99,467,664)
- Modern Literature (1,001,899)
- Non-fiction (579,545)
- Chat/Colloquial (164,457)
- Medieval (912,996)

total words: 102,700,928

Newspapers in arabiCorpus

- Al-Hayat 1996 (21,564,329) (full year)
- Al-Hayat 1997 (19,473,315) (full year)
- Al-Ahram 1999 (16,475,979) (full year)
- Al-Masri Al-Yawm 2010 (13,880,826) (full year)

- Al-Watan (Kuwait) abt half of 2002 (6,454,411)
- Al-Tajdid (Morocco) abt half of 2002 (2,919,782)
- Al-Thawra (Syria) abt a year (16,631,975)
- Al-Shuruq (Egypt) year (editorials and columns) (2,067,137)
- Treebank (598,590)

Intended Users

- Students
- Teachers
- Non-computational researchers

Short History

- Commissioned to create Synonym Book for CUP
- Needed to find citations and examples for book
- Obtained 1996-1997 Hayat on CD
- Downloaded Ahram 1999

History (continued)

- Used simple perl scripts with regular expressions to find these examples from plain text files of the newspaper content

```
#! /usr/bin/perl
$f = "NAME OF TEXT FILE”;
open RF, "<$f" or die "Can't open $f\n";
while (<RF>) {
                print if m/REGULAR EXPRESSION HERE/;
}
```

History (continued)

- Started by searching for exact form

Example:

Need sentences illustrating درس 'lesson' (drs)

Regular Expression: m/drs/

-Finds every example of that string of letters

-*Overfinds*, provides many examples of things I don't want:

مدرسة، يدرس، اندرسون، هدرسفيلد، ساندرسون، مدرسون

History (continued)

- Added word boundary (\b) to cut out undesired items

Example:

Need sentences illustrating درس 'lesson' (drs)

Regular Expression: m/\bdrs\b/

-Finds ONLY the bare form

-*Underfinds*, not finding many forms I DO want:

الدرس، بالدرس، بدرس، بدرسها، درسا، للدرس، لدرسه، ودرسه،

History (continued)

- Used Regular Expression to specify all 'permitted' forms

Example:

Need sentences illustrating درس 'lesson' (drs) (as a noun)

Regular Expression:

m/(wf)?(bAl|kAl|ll|Al|b|k|l)?drs(h|hA|k|y|hm|hn|km|kn|nA|hmA|kmA)?/

-Finds most of what we want, not much of what we don't want

History (continued)

- Saved ‘templates’ for regular expressions finding nouns, verbs, adjectives, etc. in a template file

- Copied desired template into perl program, and replaced the ‘core’ with the word I was looking for

History (continued)

- Wanted to allow my students to have access to this corpus
- Found that no matter how easy I made it, students normally would not do it if it involved using templates and perl programs
- Decided that the basic programming could be the basis of a nice site that would be appealing enough that students might actually use it

Structure of arabiCorpus

- Perl core (or 'engine')
- Php for web interface
- MySQL for saving immediate search results so user can access different parts quickly

Perl Core

- Strip vowels, kashidas, Latin letter words
- Search for every example of the string
- Part of Speech Filters: Omit from the results those that do not fit the morphological patterns for the part of speech chosen (using regular expressions)

POS Filters

- Noun: allow initial conjunctions (و، ف), prefix prepositions (ل، ك، ب), the definite article, and pronoun suffixes
- Adjective: allow initial conjunctions and the definite article only
- Adverb: allow initial conjunctions only
- Verb: allow initial conjunctions, initial verbal particles, imperfect prefixes, perfect suffixes, and pronoun endings
- String: any occurrence of the string, no restrictions

POS Filter Example

- ktb as a string finds:

كتب، الكتب، كتبه، بكتبه، يكتب، استكتب، كتبرعات

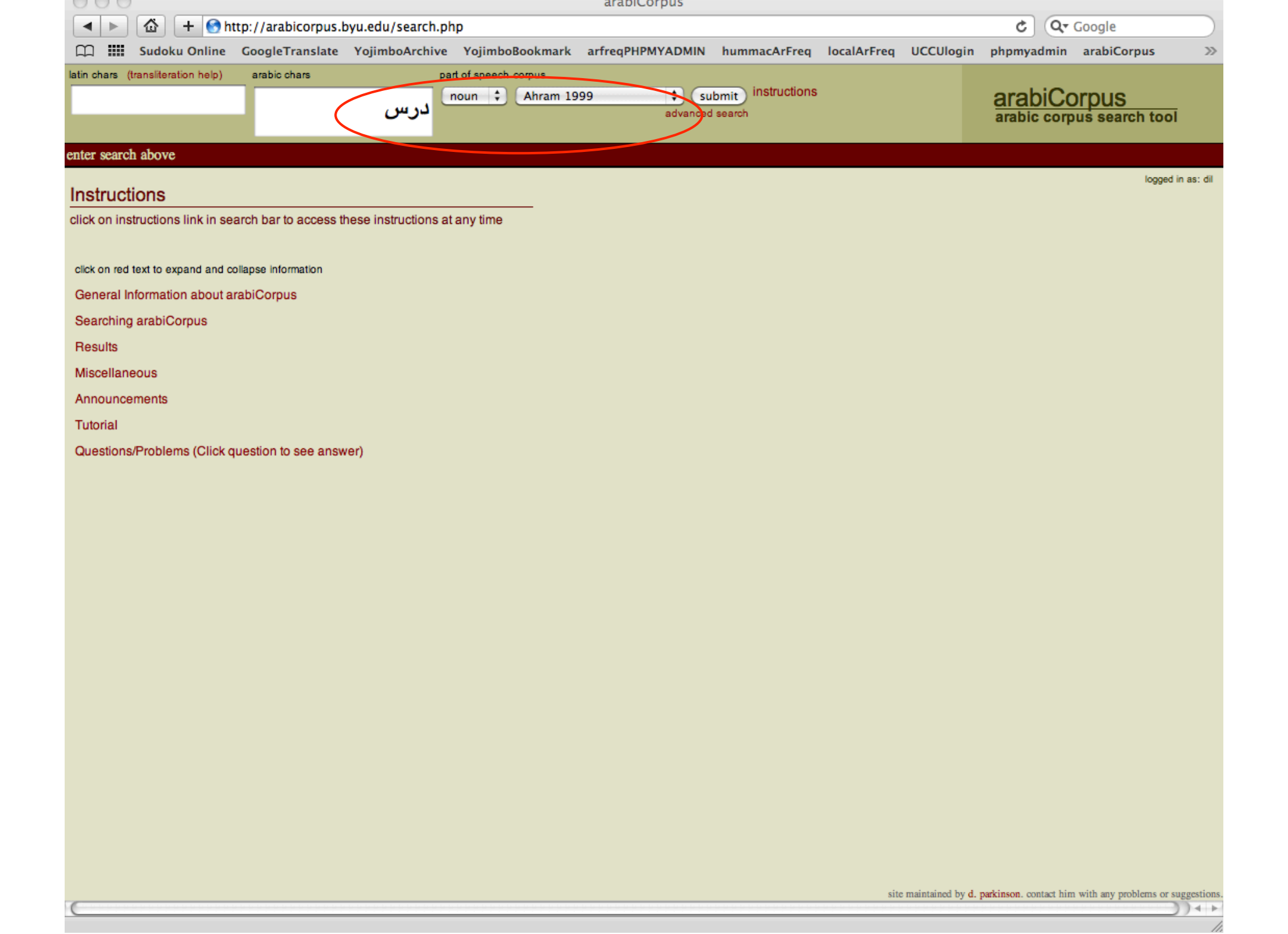
- ktAb as string finds:

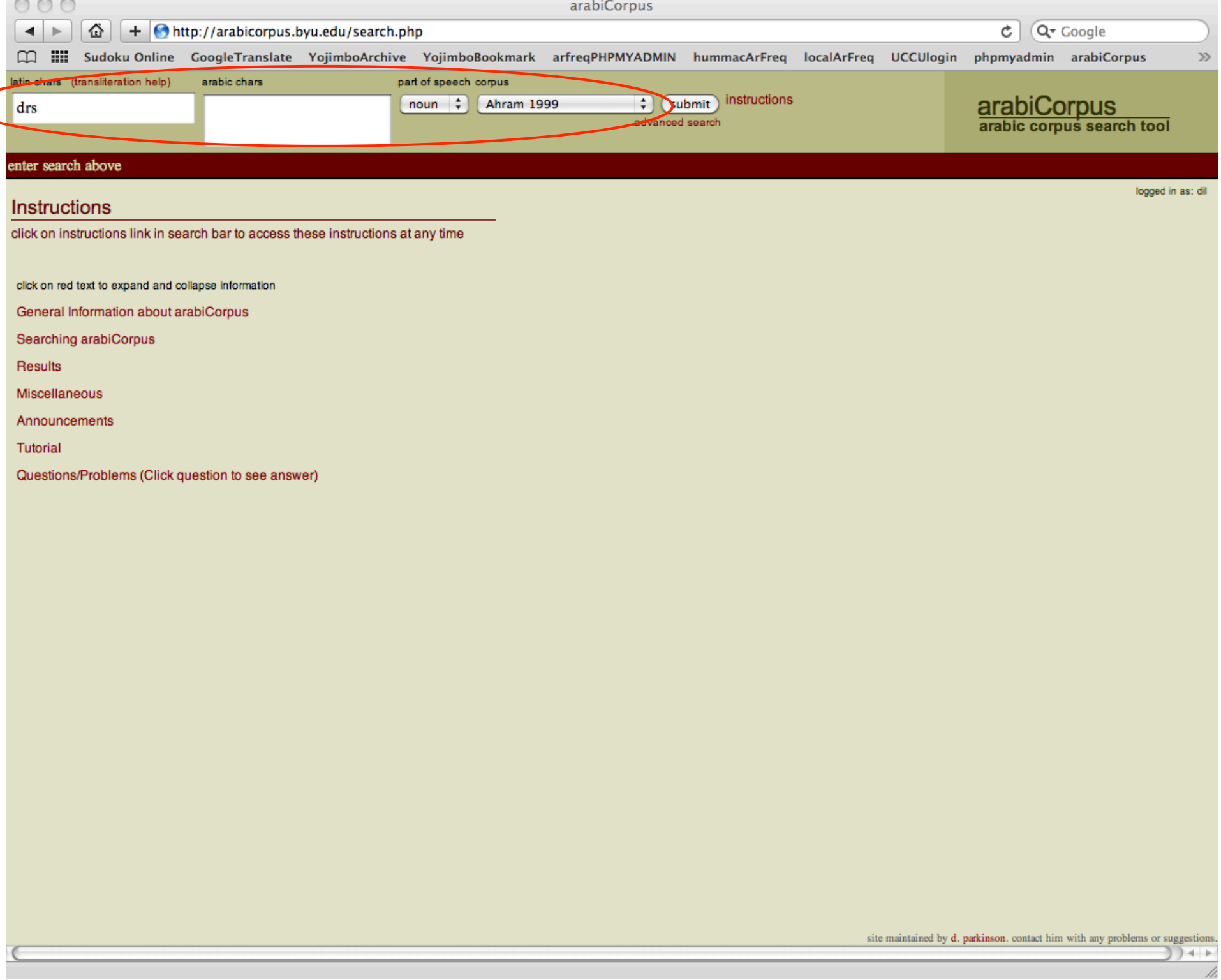
كتاب، الكتاب، بالكتاب، كتابة، استكتاب

Frequency Information

- Gives # per 100,000 of text
- Compares frequency over genres and sections
- Comparing use over one full year of a newspaper helps students get a feel for frequency

PHP User Interface


arabiCorpus
http://arabicorpus.byu.edu/search.php
Google
Sudoku Online
GoogleTranslate
YojimboArchive
YojimboBookmark
arfreqPHPMYADMIN
hummacArFreq
localArFreq
UCCUlogin
phpmyadmin
arabiCorpus
latin chars (transliteration help)
arabic chars
درس
part of speech corpus
noun
Ahram 1999
submit
instructions
advanced search
arabiCorpus
arabic corpus search tool
enter search above
logged in as: dil
Instructions
click on instructions link in search bar to access these instructions at any time
click on red text to expand and collapse information
General Information about arabiCorpus
Searching arabiCorpus
Results
Miscellaneous
Announcements
Tutorial
Questions/Problems (Click question to see answer)
site maintained by d. parkinson. contact him with any problems or suggestions.


arabiCorpus
http://arabicorpus.byu.edu/search.php
Google
Sudoku Online
GoogleTranslate
YojimboArchive
YojimboBookmark
arfreqPHPMYADMIN
hummacArFreq
localArFreq
UCCUlogin
phpmyadmin
arabiCorpus
latin chars (transliteration help)
drs
arabic chars
part of speech
noun
corpus
Ahram 1999
submit
instructions
advanced search
arabiCorpus
arabic corpus search tool
enter search above
logged in as: dil
Instructions
click on instructions link in search bar to access these instructions at any time
click on red text to expand and collapse information
General Information about arabiCorpus
Searching arabiCorpus
Results
Miscellaneous
Announcements
Tutorial
Questions/Problems (Click question to see answer)
site maintained by d. parkinson. contact him with any problems or suggestions.

summary of search results

word: drs
search string: درس — drs
database: Ahram 1999
search time: 4 seconds
part of speech: noun
search part of speech: noun
total number of occurences: 880

5.34 instances of درس per 100,000 words in Ahram 1999.

search results for drs | درس in Ahram 1999 summary citations subsections word forms words before/after

264	الفني	الفني الذي لقنه له جون ايف الخبير الفرنسي وقام بتقديم	الدرس	وفي اليوم الثاني استوعب	SPOR
265	الفيزياء	الفيزياء والرياضيات في الجامعة العبرية وجامعة ستانفورد بالولايات المتحدة الامريكية.	درس	عسكرية عديدة، في الاستخبارات والوحدات الخاصة والعمليات الخاصة بهيئة الاركان،	FILE
266	القاسي	القاسي الذي لقنه له جون ريف المدير الفني للمنتخب الوطني	الدرس	اليوم الثاني استفاد البطل السكندري من أخطاء اليوم الاول واستوعب	SPOR
267	القاسي	القاسي لاستمرار احتلالها لجنوب لبنان.	الدرس	تعتبر منطقة آمنة كما تدعي وأن المقاومة اللبنانية ستعطي لإسرائيل	FRON
268	القاسي	القاسي الذي تعلمته إسرائيل من خلالها.	والدرس	ونظرية مختلفة كانت حرب اكتوبر 73 هي الدافع الأول لها	FILE
269	القاسي	القاسي الذي تلقته في اكتوبر 1973 والزلزال الذي اجتاح المجتمع	الدرس	اسرائيل ان تفرض مفهومها الجامح عن السلام.. فعليها ان تتذكر	FILE
270	القاسي	القاسي الذي تعلمته هذه الدولة (النمر) بعد ان	الدرس	هناك منذ تفجر الأزمة المالية في عام 97 وتلخص ايضا	REPO
271	القاسي	القاسي الذي تعلمته هذه الدولة (النمر) بعد ان	الدرس	هناك منذ تفجر الأزمة المالية في عام 97 وتلخص ايضا	REPO
272	القاسي	القاسي الذي اعطاه لهم الزمالك في اليوم السابق علي اللقاء،	الدرس	أمامهم سوي هدف واحد هو تحقيق الفوز، وقد استفادوا من	SPOR
273	القاسي	القاسي الذي تلقاه منتخبنا الاوليمبي امام الاردن 5/1 في تصفيات	الدرس	المنتخبين ولن نفرط بالتفاؤل والثقة الزائدة وسنلعب للفوز مستفيدين من	SPOR
274	القاسي	القاسي، وتتخذ من هذه التجربة التي راح ضحيتها آلاف القتلي	الدرس	المهم بعد ذلك أن تعي تركيا هذا	OPIN
275	القاسي	القاسي، حتي نفيق مرة أخري علي لاعب آخر يسلك نفس	الدرس	كل الفرق بمن فيها من المدربين واللاعبين والإداريين علي هذا	AMOD
276	القاسي	القاسي من حربه الأولي هناك التي استغرقت عامين ما بين	الدرس	الروسي في اندفاعه الجنوني في الشيشان قد نسي أو تناسي	AMOD
277	القاسي	القاسي، لكنها رفضت مقابلتي وذكرني والدها حين فاتحته برغبتي في	درسها	اليها آملا في ان تكون الأيام قد علمتنا نحن الاثنين	POST
278	القانون	القانون في لندن وسافر الي جميع اركان العالم كما فعل	درس	يقفو امام الاغراءات التي تقدم اليهم. بطل الرواية امامو محامي	WRIT
279	القانون	القانون بجامعة القاهرة والتي حصل منها علي درجة الدكتوراه عامه	درس	الدكتور وليم سليمان في فوه بمحافظة كفر الشيخ عام 1924،	OPIN
280	القانون	القانون الدولي.. وانه كتب بعقلية غربية، الي الغربيين. فهو قد	درس	ومن أهم مزايا الكتاب ان المؤلف غربي	AMOD
281	القديم	القديم الذي تعلمه كل من قدر له الجلوس في مقعد	بالدرس	الجيش وكأنه يهش ذبابة، ارتكب أكبر اخطاء حياته عندما استهان	AMOD
282	القرآن	القرآن الكريم وتعاليم الدين الحنيف علي يد الفقهاء والعلماء وبعد	درس		FRON
283	القرآن	القرآن واستجلاء معانيه السامية أو بممارسة القراءة المتعمقة في الدين	درس	به، ولبدأ بتوجيه قدر اكبر من الوقت والجهد للعبادة ومحاولة	POST
284	القضية	القضية بعمق واجال النظر في جميع عناصرها وزواياها، ولقد خسر	درس	انه مسلم مثلهم، انما كان يؤيدهم نتيجة اقتناعه بعد ان	AMOD
285	القضية	القضية جيدا، والدفاع عن الحق، واحترام القضاء. ومن هنا لم	درس	بليغ باشا في لجنة الامتحان: ما هي واجبات المحامي؟ قال	OPIN
286	القضية	القضية من جميع جوانبها القانونية وفعلا حينما رجعت لمراجعة قانون	درس	نقلت كلام الوزير السابق بالحرف وهذا الكلام صادر من مسئول	OPIN
287	القمة	القمة الإفريقية المنتظمة إلي القمة العربية المنتظرة	درس		ARAB
288	القمة	القمة 84 حيث سيدفع بمشير حنفي في خط الظهر مع	درس	المنتظر ان يغير تسويبل من خطته وتشكيله في المباراة بعد	SPOR
289	القمة	القمة 84 حيث سيدفع بمشير حنفي في خط الظهر مع	درس	المنتظر ان يغير تسويبل من خطته وتشكيله في المباراة بعد	SPOR

subsections

subsection	occurences	frequency
OPIN	193	7.92 per 100,000
WRIT	142	10.63 per 100,000
AMOD	100	11.3 per 100,000
SPOR	91	5.59 per 100,000
FILE	70	5.68 per 100,000
INVE	62	5.28 per 100,000
ARTS	60	9.43 per 100,000
REPO	51	4.68 per 100,000
POST	34	10.23 per 100,000
FRON	32	1.73 per 100,000
ARAB	17	1.87 per 100,000
EGYP	13	1.59 per 100,000
WORL	9	0.99 per 100,000
ECON	6	0.49 per 100,000

word forms

23 word forms found

word form	occurences	word form	occurences	word form	occurences
درس	324	لدرس	7	وللدرس	1
الدرس	268	للدرس	7	درسهم	1
درسا	142	ودرسا	6	وبدرس	1
ودرس	32	بدرس	6	فالدرس	1
والدرس	31	فدرس	4	درسان	1
درسنا	19	بالدرس	3	ودرسه	1
درسه	11	لدرسها	1	ودرسنا	1
درسها	11	درسين	1		

words before and after

lists of before and after words occuring at least twice
click on the word for citations including that word before or after

word before	occurences	word after	occurences
هذا	39	في	53
من	30	الذي	30
هو	25	من	24
وهو	13	التاريخ	19
وهذا	12	المستفاد	16
حيث	11	خصوصي	15
وقد	10	القاسي	12
الذي	10	جيدا	11
في	10	لكل	10
أنه	10	قاسيا	10
أن	9	الثاني	8
ان	9	الأول	7
يكون	9	التجربة	7
كان	8	هو	6
قد	8	الخصوصي	5
أول	7	فيها	5
ثم	7	عمليا	5
إنه	6	لن	5
هناك	6	مهما	5
	6	يمكن	5
التى	6	الحماة	5

MySQL

- Each user allowed one search at a time
- Results temporarily stored for easy retrieval of citations, word forms, subsections, etc.

User-defined Regexs

- For more precise control, user can search with regular expressions
- User can also define a regular expression which will cut out any results that match it

Increasing Accuracy

- Define your own regular expressions
- Use the word form list and the before/after word list to cut out unlikely items
- Limit your search to unambiguous sub-sets of lemmas (like the ya- form of the imperfect verb)
- Go through results by hand to delete bad items
- Use a random number generator to pick out a representative sample of the resulting uses, and go through those by hand

Ambiguity

- The program can do nothing to limit inherently ambiguous forms in Arabic like:

ابني
كتب
والى

Common Uses

- Finding examples of specific words and forms
- Finding examples of morphological and syntactic structures
- Finding word senses
- Finding collocations and idiomatic usages
- Finding overall frequencies and frequencies of different senses, different forms, etc.

Usage

- Over 22,000 distinct queries
- Over 3300 distinct users
- Over 200 of users have accessed system over 10 times
- 25 users have accessed system over 100 times
- At least one dissertation and one scholarly book based on the corpus, with others in progress

Future Plans

- Add more literature and non-fiction
- Add historical depth
- Add newspapers from other countries
- Add ability to access surrounding collocations (not just the word before and the word after)

Thank you!

Dilworth B. Parkinson
Brigham Young University
dil@byu.edu