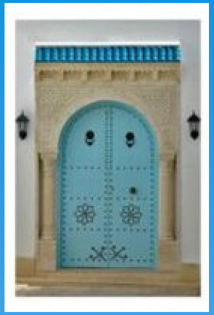
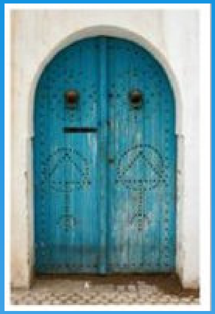
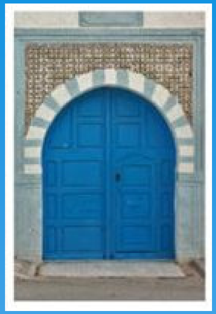




Tunisian Arabic Corpus

Design and Progress

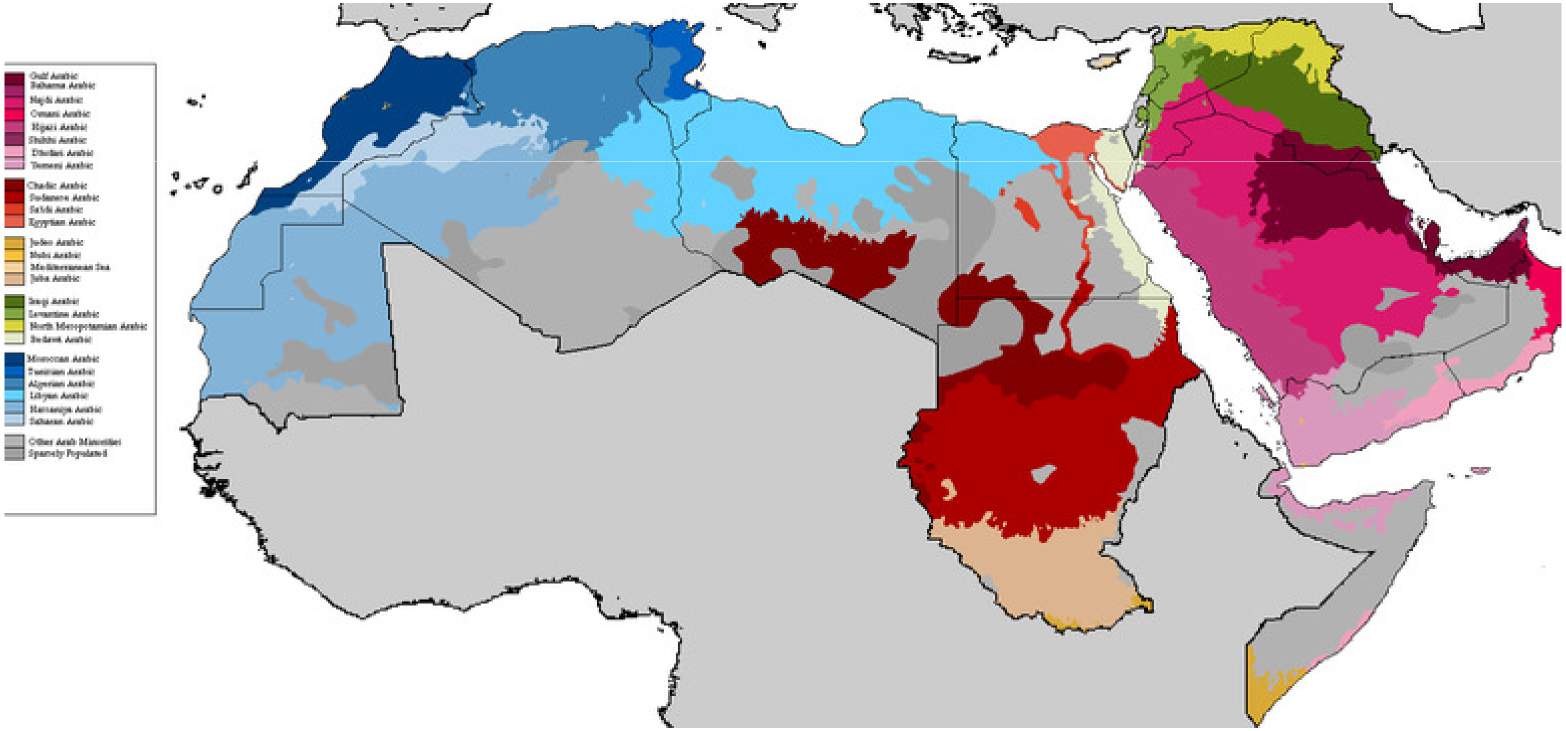
Karen McNeil
Georgetown University

Miled Faiza
University of Virginia

Why Corpora?

- Language Resources
  - Lexicography
  - Grammar
  - Authentic Teaching Materials
- Natural Language Processing
  - Automatic Speech Recognition (ASR)
  - Language ID
  - Machine Translation

A Linguistic Map of Arabic



Why Spoken Arabic?

Percentage of Overlap	Conv Egyptian vs. Media MSA	Conv British vs. Media American
Unigrams	10.3 %	44.5 %
Bigrams	1 %	19.2 %
Trigrams	< 1 %	5.3 %

Kirchhoff 2005:41

Error Rate	Egyptian Only	Egyptian + MSA
Development Set	56.1 %	54.8 %
Evaluation Set	42.7 %	41.4 %

Kirchhoff 2005:49

Spoken Arabic Corpora

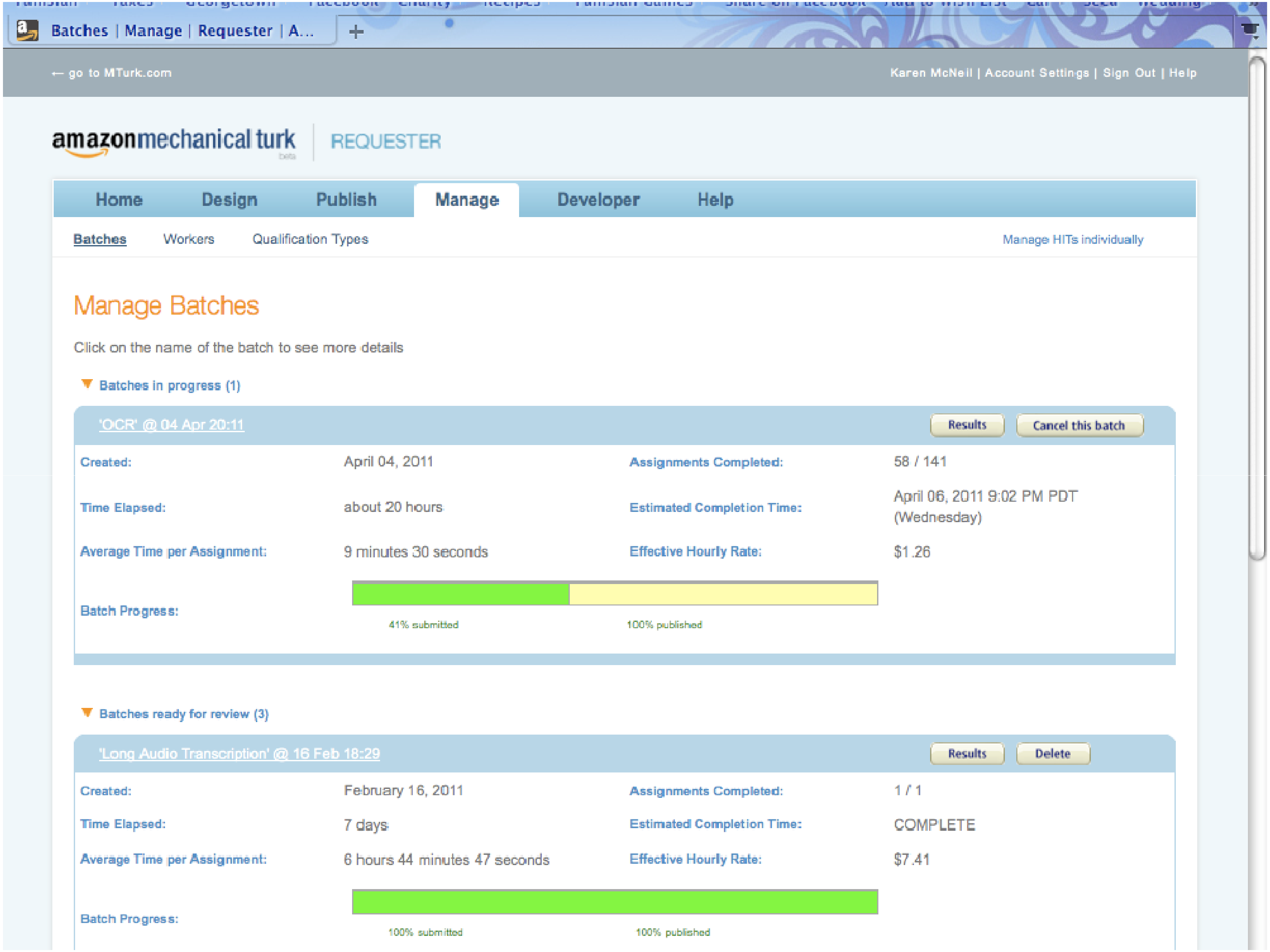
- CALLFRIEND Egyptian Arabic (1996)
- CALLHOME Egyptian Arabic Speech (1996)
- Fisher Levantine Arabic Conversational Telephone Speech (2005)
- Gulf Arabic Conversational Telephone Speech (2006)
- Iraqi Arabic Conversational Telephone Speech (2006)
- Levantine Arabic Conversational Telephone Speech (2006)
- BBN/AUB DARPA Babylon Levantine Arabic Speech and Transcripts (2005)

Why Tunisian?

- So-far neglected dialect
  - No resources available in English
- Linguistically interesting
  - Sociolinguistics / Codeswitching
- North African

Sources

- Traditional Written Sources
  - Folklore
  - Songs / Folk Poetry
  - Proverb collections
  - Screenplays
- New Written Sources
  - Blogs
  - Email
  - Facebook
- Transcribed Audio
  - Radio



Manage Batches

Click on the name of the batch to see more details

▼ Batches in progress (1)

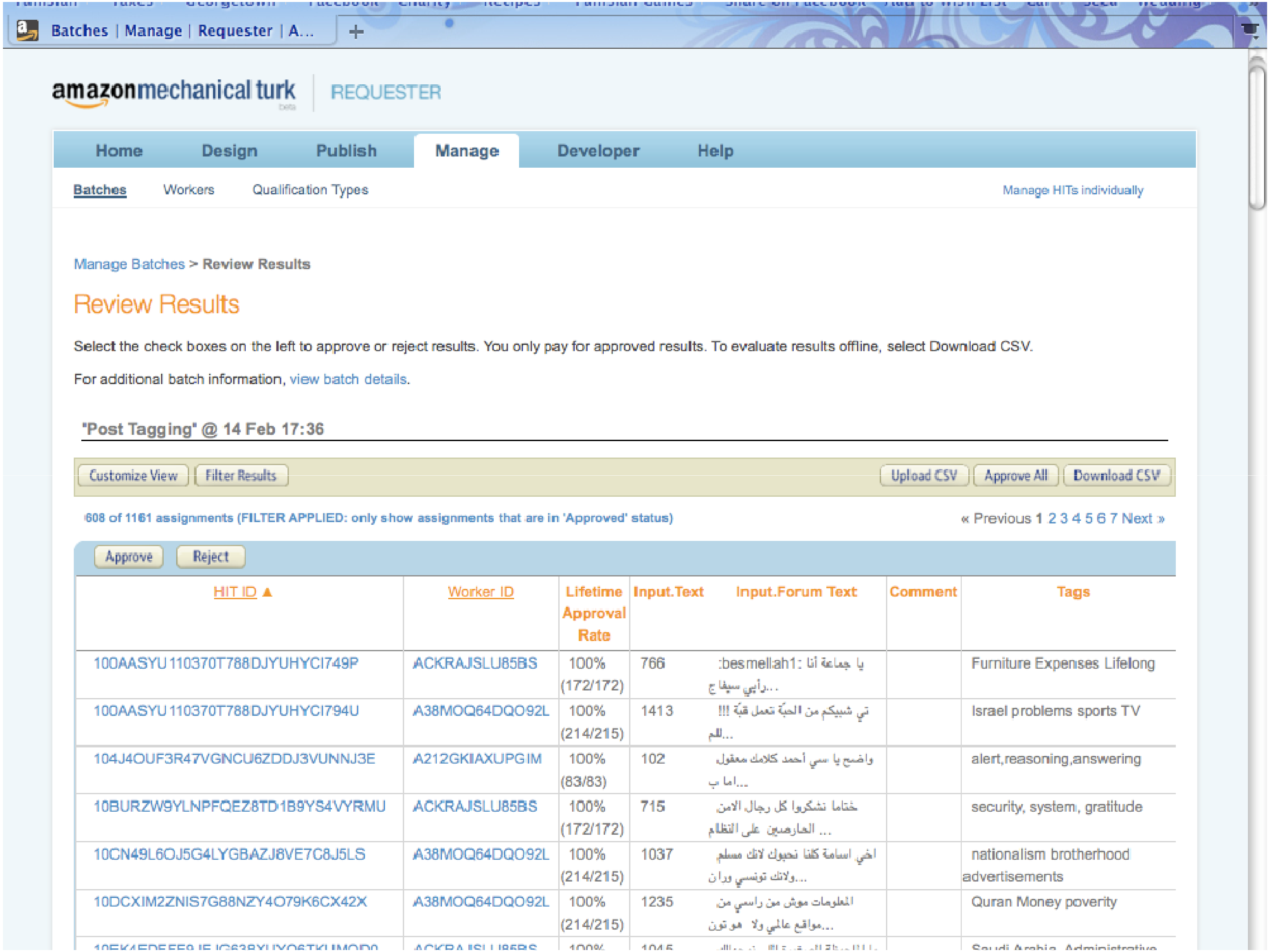
'OCR' @ 04 Apr 20:11

Created:	April 04, 2011	Assignments Completed:	58 / 141
Time Elapsed:	about 20 hours	Estimated Completion Time:	April 06, 2011 9:02 PM PDT (Wednesday)
Average Time per Assignment:	9 minutes 30 seconds	Effective Hourly Rate:	$1.26
Batch Progress:	41% submitted 100% published		

▼ Batches ready for review (3)

'Long Audio Transcription' @ 16 Feb 18:29

Created:	February 16, 2011	Assignments Completed:	1 / 1
Time Elapsed:	7 days	Estimated Completion Time:	COMPLETE
Average Time per Assignment:	6 hours 44 minutes 47 seconds	Effective Hourly Rate:	$7.41
Batch Progress:	100% submitted 100% published		



Manage Batches > Review Results

Review Results

Select the check boxes on the left to approve or reject results. You only pay for approved results. To evaluate results offline, select Download CSV.

For additional batch information, view batch details.

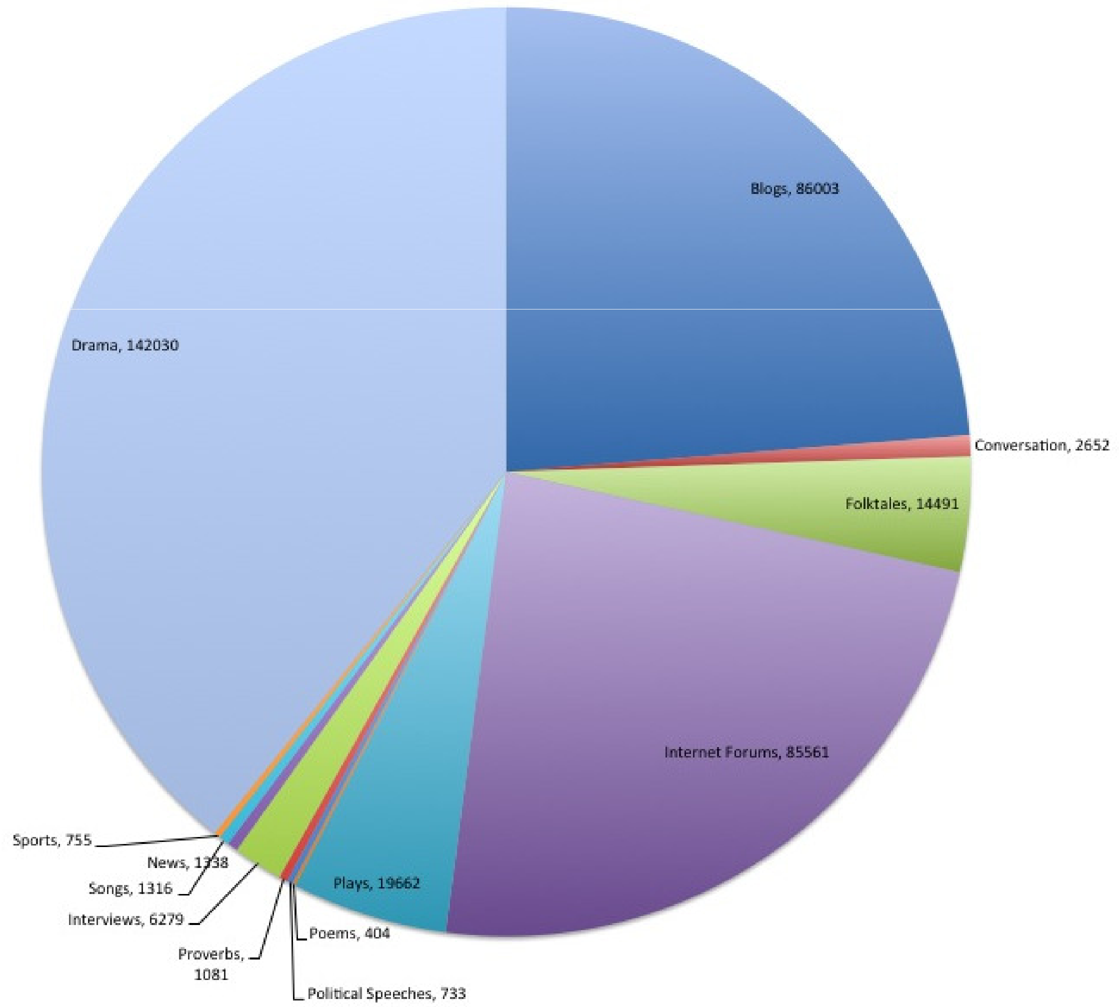
"Post Tagging" @ 14 Feb 17:36

608 of 1161 assignments (FILTER APPLIED: only show assignments that are in 'Approved' status)

« Previous 1 2 3 4 5 6 7 Next »

HIT ID ▲	Worker ID	Lifetime Approval Rate	Input.Text	Input.Forum Text	Comment	Tags
10OAASYU110370T788DJYUHYCI749P	ACKRAJSLU85BS	100% (172/172)	766	:besmellah1: يا جماعة أنا رأبي سيفاج...		Furniture Expenses Lifelong
10OAASYU110370T788DJYUHYCI794U	A38MOQ64DQO92L	100% (214/215)	1413	تي شبيكم من الحبّة تعمل قبّة !!! الم...		Israel problems sports TV
104J4OUF3R47VGNCU6ZDDJ3VUNNJ3E	A212GKIAXUPGIM	100% (83/83)	102	واضح يا سي أحمد كلامك معقول اما ب...		alert,reasoning,answering
10BURZW9YLNPFQEZ8TD1B9YS4VYRMU	ACKRAJSLU85BS	100% (172/172)	715	ختاما تشكروا كل رجال الامن الخارجين على النظام ...		security, system, gratitude
10CN49L6OJ5G4LYGBAZJ8VE7C8J5LS	A38MOQ64DQO92L	100% (214/215)	1037	اخي اسامة كلنا نحبوك لانك مسلم ولانك تونسي وران...		nationalism brotherhood advertisements
10DCXIM2ZNIS7G88NZY4O79K6CX42X	A38MOQ64DQO92L	100% (214/215)	1235	المعلومات موش من راسي من مواقع عالمي ولا هو تون...		Quran Money poverty
10EK4EDFFF9JFJG63BXUYO6TKUMOD0	ACKRAJSLU85BS	100%	1045	[illegible]		Saudi Arabia Administrative

Genres / Balance



Spelling

برشا

برشه

برشة

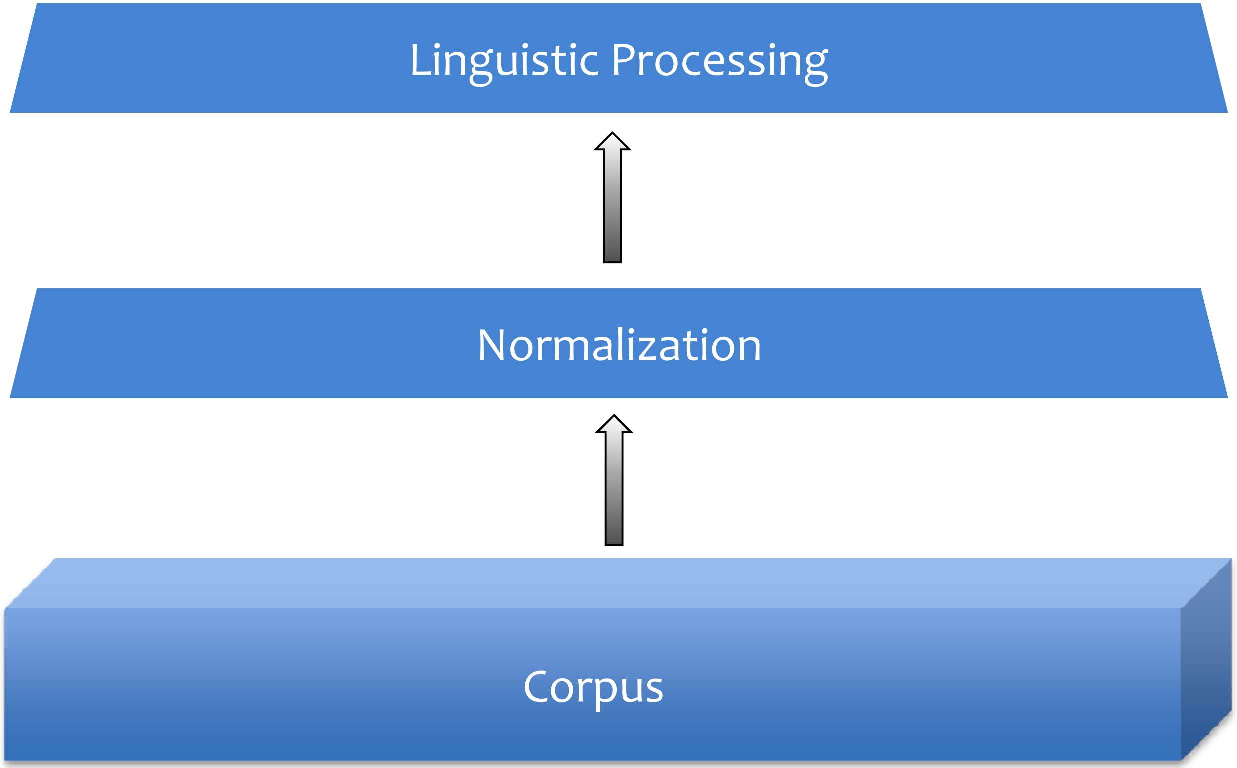
Spelling

699	برشا
33	برشه
563	برشة

Normalization

برشا : برشا، برشه، برشة

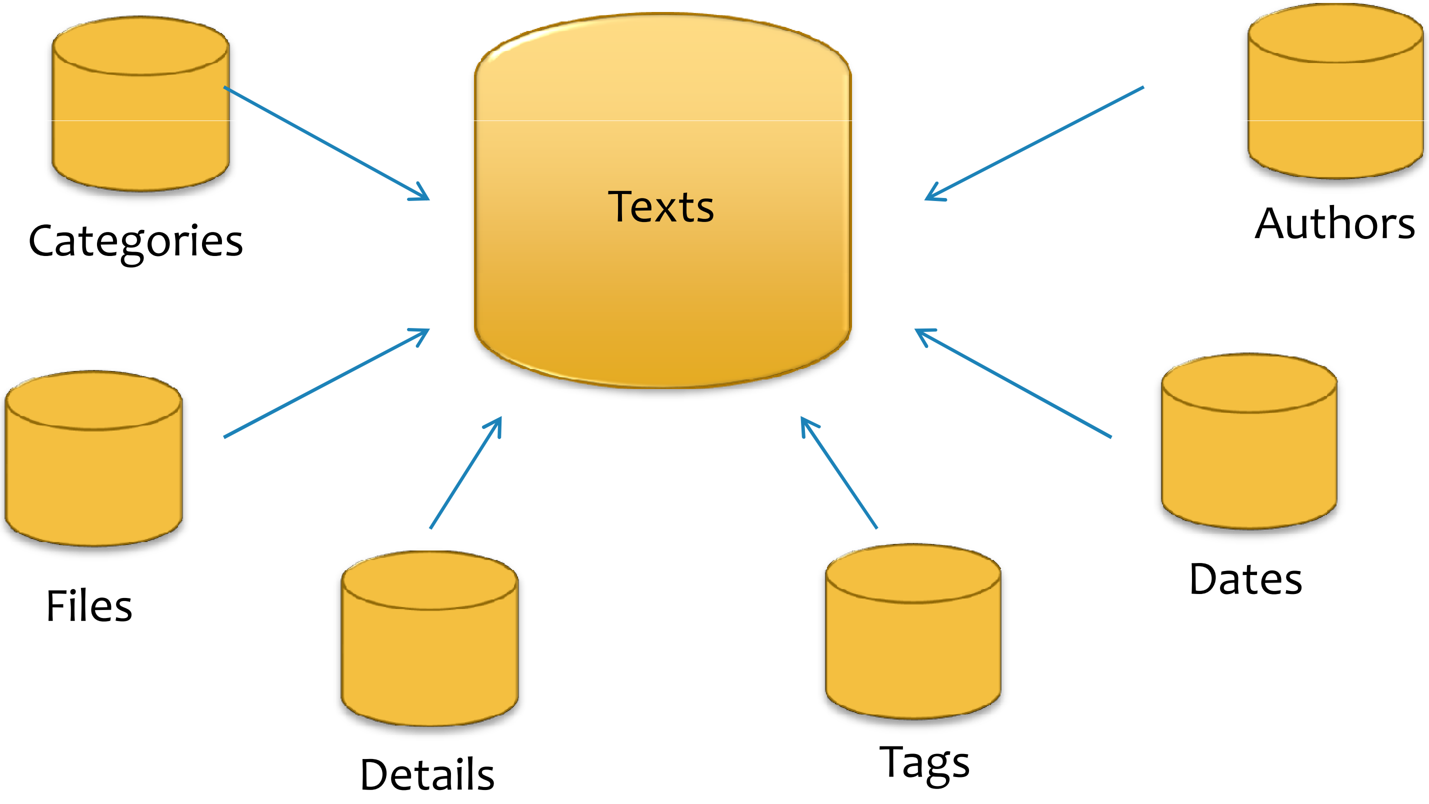
شنوه : شنو، شنوه، آشنو، آش



TAC Management Tool

- Requirements:
  - Allow easy management of Corpus and files
  - Allow remote collaboration
  - Flexible and expandable
  - Attractive and High Usability
- Technical Specs
  - Python
  - Django (Web Framework)
    - Model – View – Controller (MVC) separation

Database Structure



تونسية

- www.tunisiya.org