Shamela: A Large-Scale Historical Arabic Corpus

Yonatan Belinkov, Alexander Magidow, Maxim Romanov,
Avi Shmidman, Moshe Koppel

MIT, URI, Leipzig University, Bar Ilan University and Dicta

LT4DH Workshop at Coling, December, 2016

Outline

- Introduction
- Related work
- Initial corpus preparation
  - Metadata/data wrangling
  - Lemmatization
  - Statistics and characteristics
- Corpus enhancements
  - Text reuse
  - Text dating
- Applications
  - DH: Brown workshop, Maxim’s work
  - Linguistics: word life span, first attestations
- Conclusion

Introduction

- Arabic as a written language: >14 centuries

Introduction

- Arabic as a written language: >14 centuries
- Lack of large-scale historical corpora

Introduction

- Arabic as a written language: >14 centuries
- Lack of large-scale historical corpora
- Why this matters:
  - DH studies
  - NLP tools

Introduction

- Arabic as a written language: >14 centuries
- Lack of large-scale historical corpora
- Why this matters:
  - DH studies
  - NLP tools
- Shamela: 1 billion words from Al-Maktaba Al-Shamela (http://shamela.ws)
  - Basic preparation
  - Enhancements
  - Applications

Related Arabic Corpora

- Most focus on modern written texts

Related Arabic Corpora

- Most focus on modern written texts
- Few historically oriented corpora
    - KSUCCA (Alrabiah+13); a few others
    - All small, many lack temporal data, use texts from the Shamela website

Related Arabic Corpora

- Most focus on modern written texts
- Few historically oriented corpora
- Online corpora
  - KACST, Leeds, ICA, ArabiCorpus, CLAUDia
  - Large, but not downloadable, lack temporal information

Related Arabic Corpora

- Most focus on modern written texts
- Few historically oriented corpora
- Online corpora
- Shamela
    - Fine-grained time information
    - Covers most of the history
    - Available for download (inside the RAWrabica collection)

Initial Corpus Preparation

- Metadata/data wrangling
  - Website not designed as a corpus: original texts manually digitized
  - Semi-automatic process for metadata organization
  - Basic de-duplication

Initial Corpus Preparation

- Metadata/data wrangling
- Lemmatization
  - Importance for Arabic
  - MADAMIRA (Pasha+14)
  - Reduction in vocabulary size

Words	Lemmas
16.8M	95K

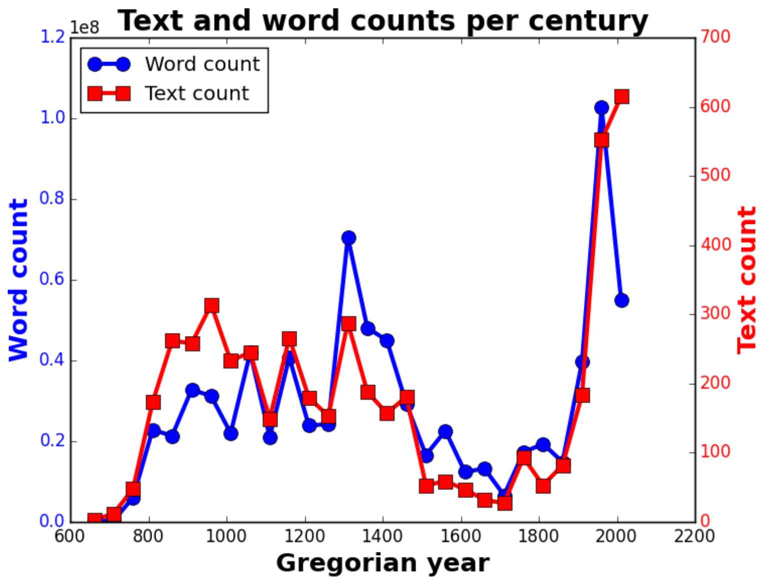
Initial Corpus Preparation

- Metadata/data wrangling
- Lemmatization
- Statistics and Characteristics

	Texts	Words
Dated	4,900	800M
Undated	1,200	200M
Total	6,100	1B

Initial Corpus Preparation

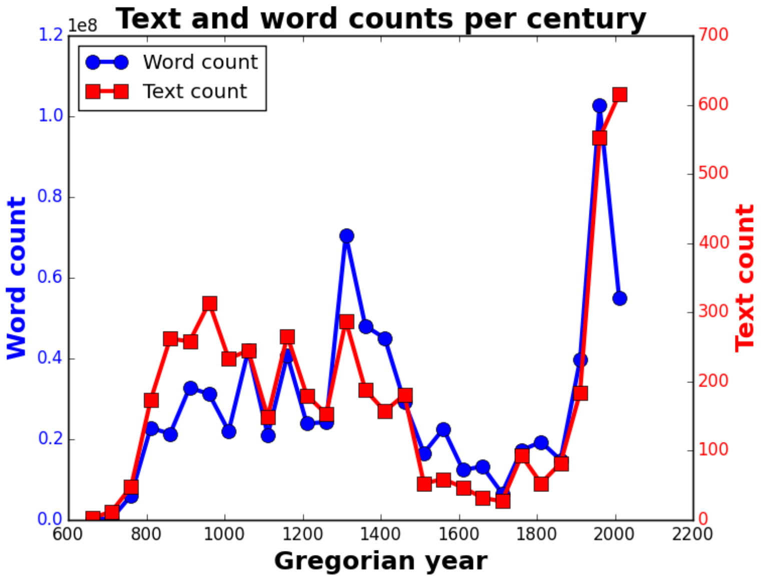
- Metadata/data wrangling
- Lemmatization
- Statistics and Characteristics



	Texts	Words
Dated	4,900	800M
Undated	1,200	200M
Total	6,100	1B

Initial Corpus Preparation

- Metadata/data wrangling
- Lemmatization
- Statistics and Characteristics



	Texts	Words
Dated	4,900	800M
Undated	1,200	200M
Total	6,100	1B

Genre	Average Date	Texts
Hadith Collections	946	179
Biographies	1334	377
Jurisprudence (*Fiqh*)	1486	157
Popular religious writing	1998	298

Challenges

- Text reuse and duplication
  - Writing style in religious texts
  - Quotations, paraphrases, copying

- Undated texts
  - Large portion of undated texts
  - Contemporary introductions to classical texts

Text Reuse

- Important for DH and computational linguistics

Text Reuse

- Important for DH and computational linguistics
- Previous work
  - Law bills, newspaper texts (Smith+14, Wilkerson+15)
  - Text alignment based on n-grams (Smith+14, Li 16)

Text Reuse

- Important for DH and computational linguistics
- Previous work
- Our approach
  - First step: exclude “boiler-plate” text chunks (blessings, formulae, etc.)
  - Second step: skip-gram matching over two-letter hashes (Shmidman et al 2016)

Text Reuse

- Important for DH and computational linguistics
- Previous work
- Our approach
  - First step: exclude “boiler-plate” text chunks (blessings, formulae, etc.)
  - Second step: skip-gram matching over two-letter hashes (Shmidman et al 2016)
- Results
  - 18M words of very frequent passages
  - >5 million pairwise approximate matches, average length of 40 words

Text Reuse

- Boiler-plate: Almost always Quranic verses, occasionally general prayers
  - أَنِ الْحَمْدُ لِلَّهِ نَحْمَدُهُ ، وَنَسْتَعِينُهُ ، وَنَسْتَغْفِرُهُ ، وَنَعُوذُ بِاللَّهِ مِنْ شُرُورِ أَنْفُسِنَا ، وَمِنْ سَيِّئَاتِ أَعْمَالِنَا
  - "All praise to god who we worship, and seek comfort in, and seek forgiveness from, and we seek shelter from the evil of our selves and of our deeds…" (part of a much longer prayer attested as a *hadith)*
- Non-trivial matches: Longer *hadith* texts, longer quotations
  - ابراهيم بن يزيد النخعي ذكره الحاكم وغيره من المدلسين وحكى خلف بن سلام عن عدة من مشايخه أن تدليسه من أعمض شيء وكانوا يتعجبون منه
  - "Ibrahim bin Yaziid al-Nakha`i was mentioned by Al-Hakim and others as a forger (of *hadith),* and Khalaf bin Salaam says on behalf of many of his teachers that his fabrications are quite obscure(?) and they were amazed by him" (original from 1359, then 1437, 1505 and in a modern text)

Text Dating

- Large portion of undated texts

Text Dating

- Large portion of undated texts
- Previous work (de Jong+05, Dalli&Wilks 06, Chambers 12, Niculae+14, Popescu&Strapparava 15)
  - Variety of features, methods, and granularity levels

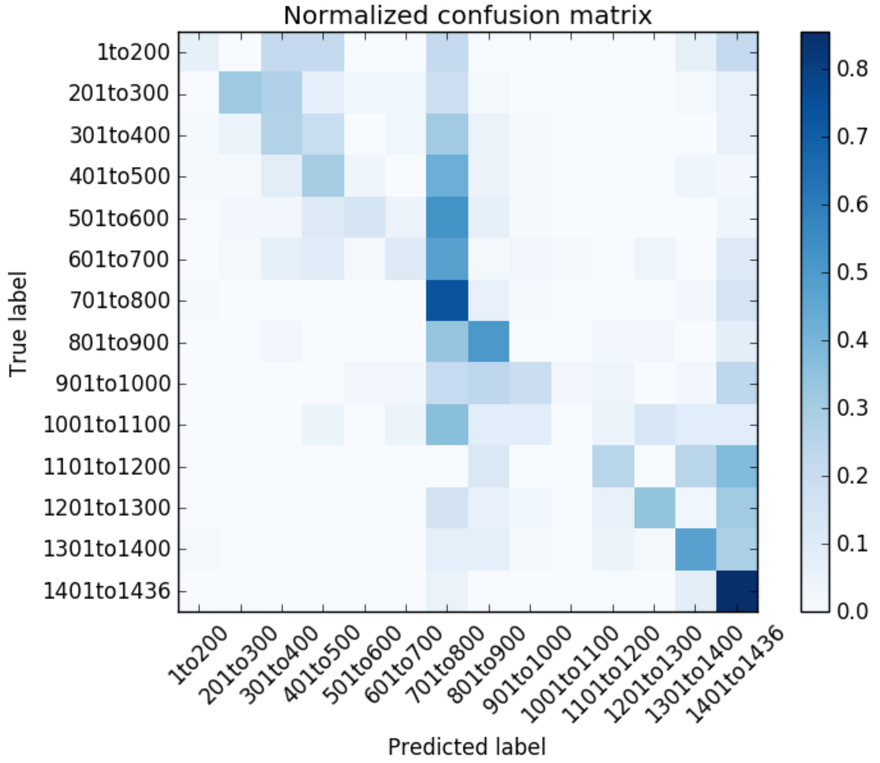
Text Dating

- Large portion of undated texts
- Previous work
- We take a simple language modeling approach
  - Train language models on dated texts (5-gram LM with Knesser-Nay smoothing)
  - Rank undated texts by perplexity
  - Validate on held-out dated texts
  - Bucket at a 100 years granularity

Text Dating

- Results

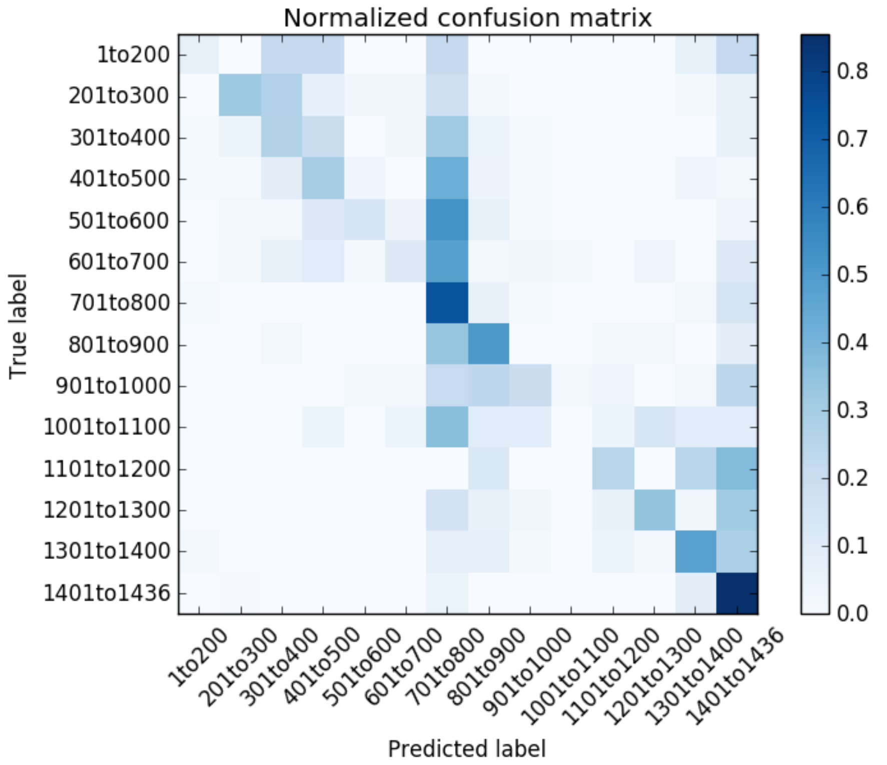
System	Accuracy
Random	7.14
Majority	20.29
LM	42.95
LM (top 3)	71.14



Text Dating

- Results

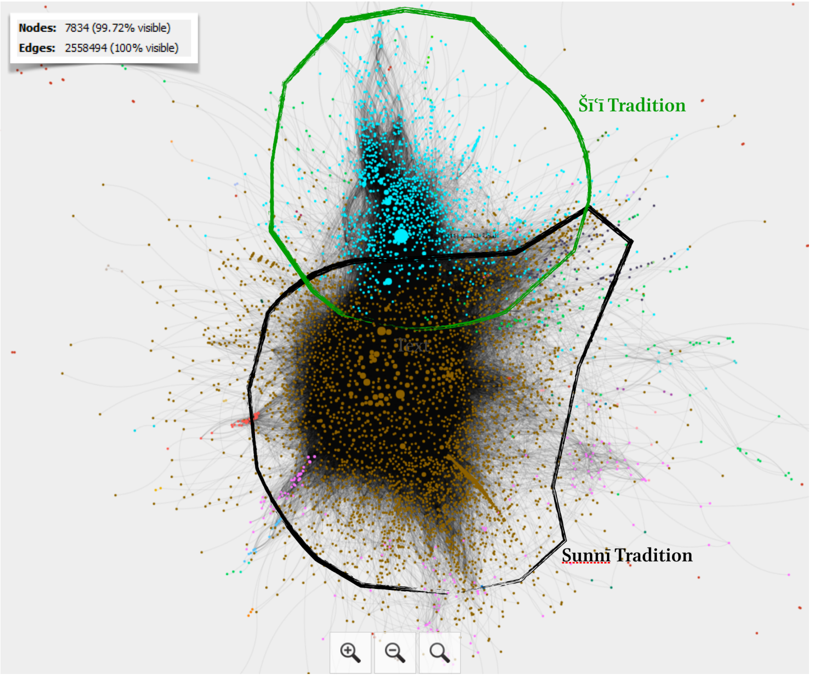
System	Accuracy
Random	7.14
Majority	20.29
LM	42.95
LM (top 3)	71.14



- Manual inspection
  - Confusion between subsequent periods
  - Identification of mixed texts
  - Prioritization of manual tagging

Applications

- DH
  - Brown University's workshops on Islamic Digital Humanities
  - Intellectual networks, transmissions, cultural geography (Romanov 13)



Applications

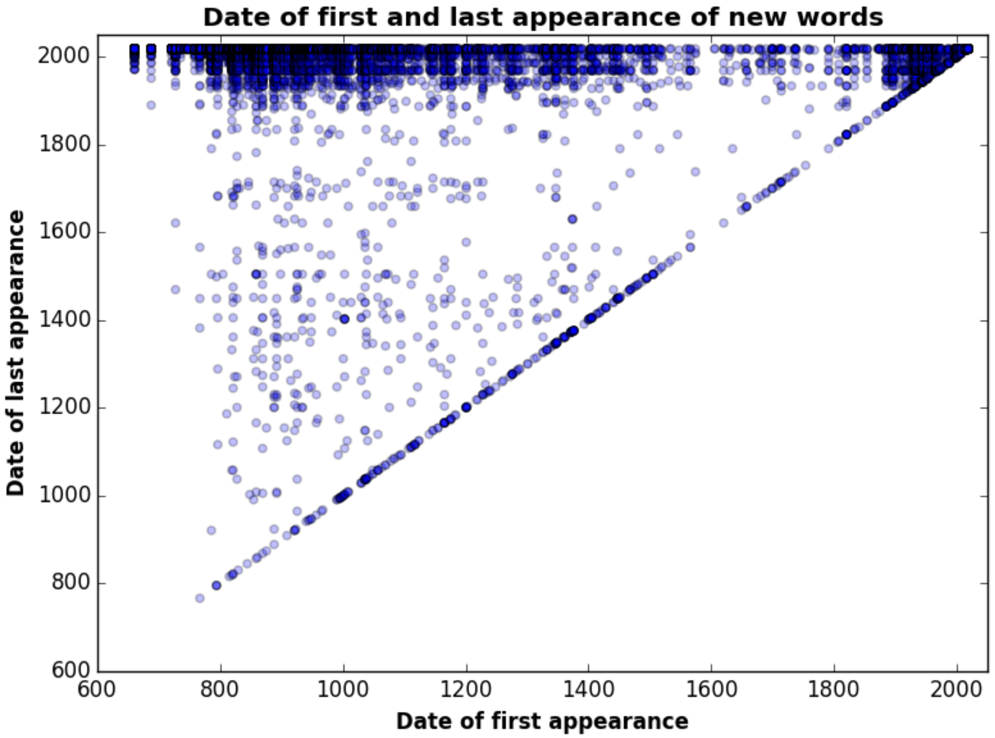
- DH
- Linguistics
  - Lifespan of Arabic words
  - First attestations of words

Arabic Word Lifespan

- Impression of little variation between modern and classical Arabic

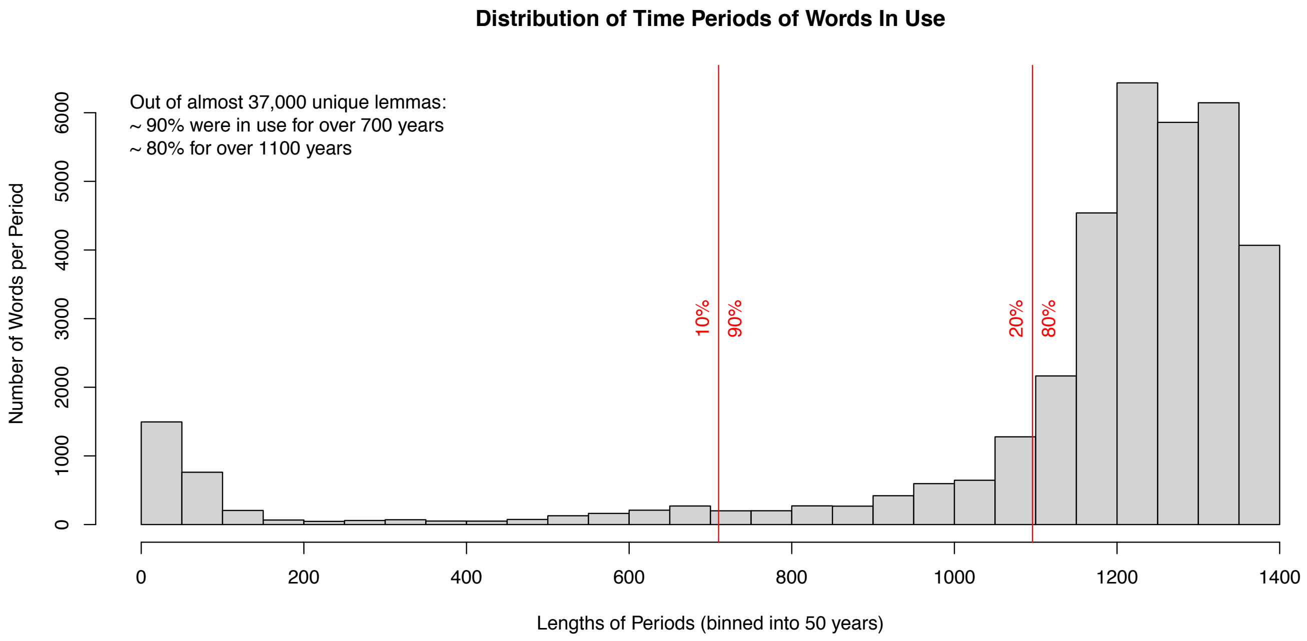
Arabic Word Lifespan

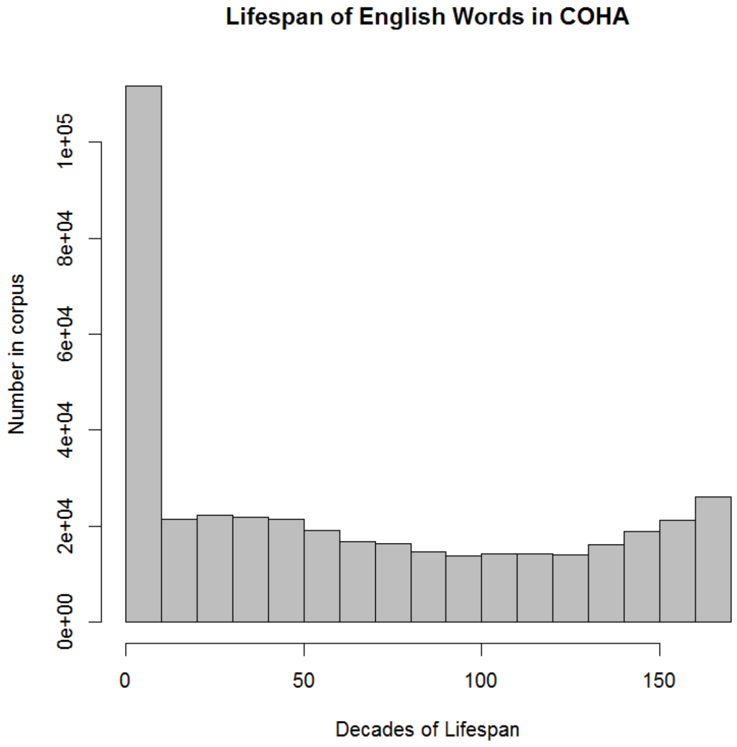
- Contrasting first and last usages



Arabic Word Lifespan

- Contrasting first and last usages





First Attestations

- ”Say/don’t say" statements
  - Ḥawālay (حوالي) “around, approximately”
  - “Should not be used for approximation of number”

First Attestations

- ”Say/don’t say" statements
  - Ḥawālay (حوالي) “around, approximately”
  - “Should not be used for approximation of number”
- In Shamela
  - Indeed, very early usages for physical approximation
  - But, fairly early usage of numerical approximation (1201 CE)

Conclusion

- Contributions
  - Making available a 1 billion word historical corpus of Arabic
  - Improving corpus quality
  - Demonstrating its utility

Conclusion

- Contributions
  - Making available a 1 billion word historical corpus of Arabic
  - Improving corpus quality
  - Demonstrating its utility
- Future work
  - Periodization of Arabic
  - You?