# A Lexical Distance Study of Arabic Dialects

Kathrein Abu Kwaik (Chatrine Qwaider) [1] Motaz Saad [2]
Stergios Chatzikyriakidis [1] Simon Dobnik[1]

[1] CLASP, Department of Philosophy, Linguistics and Theory of Science (FLoV), University of Gothenburg, Sweden

[2]The Islamic University of Gaza, Gaza, Palestine

13th Feb 2019

CLASP centre for linguistic theory and studies in probability

# Objective

- Building a Linguistic resource for the Levantaine dialects (Palestinian, Jordanian, Syrian, Lebanese)

# Objective

- Building a Linguistic resource for the Levantaine dialects (Palestinian, Jordanian, Syrian, Lebanese)
- Conduct a computational cross dialectal lexical distance study to measure the similarities and differences between dialects and the MSA

CLASP centre for linguistic theory and studies in probability

# Objective

- Building a Linguistic resource for the Levantaine dialects (Palestinian, Jordanian, Syrian, Lebanese)
- Conduct a computational cross dialectal lexical distance study to measure the similarities and differences between dialects and the MSA
- Highlight the lexical relation between the MSA and Dialectal Arabic (DA) in more than one Arabic region

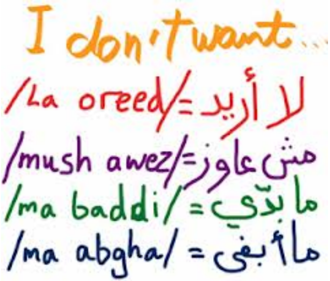
CLASP centre for linguistic theory and studies in probability

# Objective

- Building a Linguistic resource for the Levantaine dialects (Palestinian, Jordanian, Syrian, Lebanese)
- Conduct a computational cross dialectal lexical distance study to measure the similarities and differences between dialects and the MSA
- Highlight the lexical relation between the MSA and Dialectal Arabic (DA) in more than one Arabic region
- A basis for building NLP tools for dialectal processing by adapting MSA tools and focusing on areas of similarity and degrees of difference

CLASP centre for linguistic theory and studies in probability

# Diglossia




https://middleeasttransparent.com/en/diglossia/arabic-dialects/

- Diglossia is a very common phenomenon in Arabic-speaking communities, where the spoken language is different from both Classical Arabic (CA) and Modern Standard Arabic (MSA)
- The spoken language is characterised as a number of dialects used in everyday communication as well as informal writing

# Classification of Arabic Language

- The Classical Arabic: the language of the Holy Quran
- The Modern Standard Arabic: the formal spoken and written language
- The Dialectal Arabic: the informal spoken variety and nowadays an informal written language (as: social media)

CLASP centre for linguistic theory and studies in probability

# Building Levantine Corpus SHAMI

- the first Levantine dialect corpus that contains the largest volume of data separated as individual Levantine dialects
- it is not a crafted and also not a parallel corpus; it contains real conversations as written in social media and blogs;
- it includes several topics from regular conversations such as politics, education, society and others;
- SDC has been created from scratch by collecting Levantine data through automatic and manual approaches.

CLASP centre for linguistic theory and studies in probability

# Data Collection

- Automatic Collection
  - collect IDS for activist and some public Figures
  - we use *tweepy* to collect tweets and replies from these IDs.
  - extract data according to geographical location.
- Manual Collection
  - We harvest the web and choose online dialectal blogs and forums in Levantine countries.
- Overall, this gives us sentences of various lengths.

# Data Pre-processing

- Remove diacritics: أّ Tashdid, أَ *a* Fatha, أً *an* Tanwin Fath.
- Remove non Arabic words, Latin characters, numbers and dates, emoticons, and symbols.

CLASP centre for linguistic theory and studies in probability

# Data Pre-processing

- Normalization: there is no standard orthography for Arabic dialects. we implement finer rules that work more reliably and preserve the semantic meaning of the text, for example:
  1. Aleph: we only convert Aleph with an accent أ *ʾa* to Aleph without an accent ا *ā* if it appears at the beginning of the word. This is because we want to mark the accent in other contexts in order to preserve the meaning of dialectal words. For example, ('هلأ *hlʾa*' / now) from ('هلا *hlā*'/ Hello).
  2. Alef Maqsora (ى *ā*) at the end of the word: in most processing steps the letter (ى *ā*) is converted to a (ي *y*), but we did not do so because a lot of words would change the meaning. For example: (على *ʿlā* / on preposition) and (علي *ʿly* Ali / a personal name).

CLASP centre for linguistic theory and studies in probability

# Data Pre-processing

- Remove repeated characters: (for example Waaaaaaw).
  1. We extract all words containing repeated characters in MSA texts and keep them in a list.
  2. All words containing duplicate characters from the previous list are abbreviated to two characters.
  3. The rest of the characters are reduced to only one character, for example the repeating character و *w* in (مبرووووك *mbrwwwwwk* / congratulation) is converted to (مبروك *mbrwk*/ congratulation).
  4. The conjunction letter (و *w* / and), We postulated that if the given word begins with more than one (و *w*), the first (و *w*) and the rest of the word are separated and the original word is processed according to the previous algorithm.

CLASP centre for linguistic theory and studies in probability

# Corpus size

| Shami Corpus | | | |
|---|---|---|---|
| | sentences | tokens | types |
| Jordanian | 32 K | 0.47 M | 69 K |
| Palestinian | 21 K | 0.35 M | 56 K |
| Syrian | 48 K | 0.7 M | 63 K |
| Lebanese | 16 K | 0.2 M | 34 K |
| **Total** | **117 K** | **1.72 M** | **222 K** |

Table: Statistics for SDC

CLASP centre for linguistic theory and studies in probability

# Difference

Gulf (Aish)

Egypt (Aish)

https://www.aljamila.com/node/118751/

https://www.youtube.com/watch?v=ubC9j1xrNTU

CLASP centre for linguistic theory and studies in probability

# Difference

على كيفك
ALA KAIFAK

Palestine, Egypt → As you like

Syria→ Perfect

Iraq → Take your time

https://twitter.com/alakaifakco

CLASP centre for linguistic theory and studies in probability

# Qualitative differences between MSA and DA

There are qualitative differences at all levels of linguistic representations:

1. Orthographical and Phonological Differences
2. Morphological Differences
3. Syntactic Differences
4. Lexical and Semantic differences

CLASP centre for linguistic theory and studies in probability

# Orthographical and Phonological Differences

- Dialectal Arabic (DA) does not have an established standard orthography like MSA.
- Arabic script and the Latin alphabet is used for writing short messages or posting on social media, For example, كيفك *kyfk* / "how are you" is represented as Keifk.

CLASP centre for linguistic theory and studies in probability

# Orthographical and Phonological Differences

- Dialectal Arabic (DA) does not have an established standard orthography like MSA.
- Arabic script and the Latin alphabet is used for writing short messages or posting on social media, For example, كيفك *kyfk* / "how are you" is represented as Keifk.
- Pronunciation of dialectal words containing the letter ق *q* which depends on the dialect and regions. For instance, the Palestinian speakers from rural and urban regions pronounce it like /'/ glottal stop or /k/ while Bedouin pronounce it as /g/ .
  - The word قال *qāl* /say is pronounced and sometimes written as قال *qāl* , كال *kāl* , ئال *ʾyāl* or جال *ǧāl*

CLASP centre for linguistic theory and studies in probability

# Morphological Differences

- There are some important differences between MSA and dialectal Arabic in terms of morphology because of the way of using these clitics, particles and affixes

| Example | Dialect word | Dialect | MSA | English |
|---|---|---|---|---|
| Using multiple words together | كيفك *kyfk*<br>معلش *mʕlš* | Levantine<br>Egyptian | كيف حالك *kyf ḥālk*<br>لا يهم *lā yhm* | How are you?<br>Does not matter |
| Sharing the stem with different affixes | مبدرسش *mbdrsš*<br>ما بيدرس *mā bydrs*<br>مبيدرسش *mbydrsš* | Palestinian<br>Syrian<br>Egyptian | لايدرس *lāydrs* | He does not study |
| The future marker | ح ، راح *ḥ, rāḥ*<br>حيلعب *ḥylʕb*<br>راح يلعب *rāḥ ylʕb* | Palestinian | سوف *swf*<br>سوف يلعب *swf ylʕb* | will<br>He will play |
| Clitics | ب *b* for present<br>بياكل *byākl*<br>عم بطبخ *ʕm bṭbẖ* | <br>Egyptian<br>Syrain | <br>يأكل *yʔakl*<br>أنا أطبخ *ʔanā ʔaṭbẖ* | <br>He is eating<br>I am cooking |

centre for linguistic theory and studies in probability

# Syntactic Differences

In regarding to S(Subject) , V(verb), O(Object) order in the sentence.

CLASP centre for linguistic theory and studies in probability

# Lexical and Semantic differences

| MSA | Englihs | Negation | English |
|---|---|---|---|
| أعرف *ʔaʕrf* | know | لا أعرف *lā ʔaʕrf* | Don't know |
| Palestinian | Jordanian | Syrian | Lebanese |
| مش عارف *mš ʕārf* | مش عارف *mš ʕārf* | ما بعرِف *mā bʕrif* | ما بعرِف *mā bʕrif* |
| Egyptian | Algerian | | Tunisian |
| معرفش *mʕrfš* | مش نعرف *mš nʕrf* | ملبعاليش *mlbʕālyš* | منيش عارف *mnyš ʕārf* |
| Gulf | Iraqi | | |
| مدري *mdry* | ما أدري *mā ʔadry* | | |

Figure: Differences in negation between the dialects

CLASP centre for linguistic theory and studies in probability

# Lexical and Semantic differences

| MSA أعرف *ʔaʕrf* | Englihs know | Negation لا أعرف *lā ʔaʕrf* | English Don't know |
|---|---|---|---|
| Palestinian مش عارف *mš ʕārf* | Jordanian مش عارف *mš ʕārf* | Syrian ما بعرِف *mā bʕrif* | Lebanese ما بعرِف *mā bʕrif* |
| Egyptian معرفش *mʕrfš* | Algerian مش نعرف *mš nʕrf* | ملبعاليش *mlbʕālyš* | Tunisian منيش عارف *mnyš ʕārf* |
| Gulf مدري *mdry* | Iraqi ما أدري *mā ʔadry* | | |

Figure: Differences in negation between the dialects

| MSA الآن *ālʔān* | English Now | | |
|---|---|---|---|
| Levantine هلأ، هلقيت *hlʔa, hlqyt* | Bedouin هلحين *hlḥyn* | Saudi Arabia دحين *dḥyn* | Iraqi هالوقت *hālwqt* |
| Libyan توا *twā* | Tunisian توة *twh* | Algerian توا *twā* | Egyptian دلوقتي، دلوقت *dlwqty, dlwqt* |

Figure: Examples for new lexicon in dialects

CLASP centre for linguistic theory and studies in probability

# Lexical and Semantic differences

| Word | Original | MSA | English | Word | Original | MSA | English |
|---|---|---|---|---|---|---|---|
| طربيزة *ṭrbyzh* | Turkish | طاولة *ṭāwlh* | Table | بندورة *bndwrh* | Italian | طماطم *ṭmāṭm* | Tomatoes |
| أستاذ *ʔastāḏ* | Persion | مدرس *mdrs* | Teacher | توف *twf* | Hebrew | جيد *ǧyd* | Good |
| أفوكادو *ʔafwkādw* | French | محامي *mḥāmy* | lawyer | تليفون *tlyfwn* | English | هاتف *hātf* | Telephone |

Figure: Examples of borrow words from other languages

CLASP centre for linguistic theory and studies in probability

# Quantitative differences between MSA and DA

- Arabic Corpora

| Corpus Name | Type | Dialects | Description |
| --- | --- | --- | --- |
| PADIC (Parallel Arabic Dialect Corpus) | Parallel | MSA, Algerian, Tunisian, Palestinian, Syrian | The corpus is collected from Algerian chats and conversations which are translated to MSA and then to other dialects. |
| Multi-dialectal Arabic parallel corpus | Parallel | MSA, Egyptian, Syrian, Palestinian, Tunisian, Jordanian | This corpus is originally build on Egyptian dialects extracted from Egyptian-English corpus. It has been translated to the remaining dialects by four translators |
| SDC (Shami Dialect Corpus) | Non-parallel | Palestinian, Syrian, Jordanian, Lebanese | The corpus is collected from different sources of social media, blogs, stories and public figures on the Internet. |
| WikiDocs Corpus | Comparable | MSA, Egyptian | It contains a comparable documents from Wikipedia. |

Figure: List of Arabic corpora used to investigate the differences between dialects

CLASP centre for linguistic theory and studies in probability

# Measuring quantitative differences between MSA and DA

- For parallel corpora: the comparison is at the **document (sentence) level**
- For comparable and non-parallel corpora: the comparison is at the **corpus level**
- **Programming language**: Python (Gensim Library)

CLASP centre for linguistic theory and studies in probability

# Methods

- Exploit several methods from Natural Language Processing (NLP) and Information Retrieval (IR)
  - Vector Space Model (VSM)
  - Latent Semantic Indexing (LSI)
  - Hellinger Distance (HD)

CLASP centre for linguistic theory and studies in probability

# Methods

- Exploit several methods from Natural Language Processing (NLP) and Information Retrieval (IR)
  - Vector Space Model (VSM)
  - Latent Semantic Indexing (LSI)
  - Hellinger Distance (HD)
- Apply different Arabic dialectal corpora

CLASP centre for linguistic theory and studies in probability

# Methods

- Exploit several methods from Natural Language Processing (NLP) and Information Retrieval (IR)
  - Vector Space Model (VSM)
  - Latent Semantic Indexing (LSI)
  - Hellinger Distance (HD)
- Apply different Arabic dialectal corpora
- Measure the overlap among all the dialects and compute the frequencies of the most frequent words in every dialect

CLASP centre for linguistic theory and studies in probability

# Lexical Sharing and Overlapping **(Jaccard Index)**

- We compute the percentage of vocabularies that overlap between these dialects

$$JaccardIndex(A, B) = \frac{|A \bigcap B|}{|A \bigcup B|} \tag{1}$$

**Note**

- The Multi-dialect corpus is biased towards the EGY dialect, as EGY was the pivot language when the corpus was built. This is reflected in all the measures used in this study
- the bias of the pivot language is not reflected between ALG and MSA in the PADIC corpus as these are the least similar varieties

CLASP centre for linguistic theory and studies in probability

# Lexical Sharing and Overlapping (Jaccard Index)

| | PADIC | | | | Multi-dialect corpus | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | ALG | TN | SY | PA | | EG | JO | TN | SY | PA |
| MSA | 0.1 | 0.14 | 0.14 | 0.19 | MSA | 0.21 | 0.14 | 0.13 | 0.15 | 0.16 |
| PA | 0.13 | 0.14 | 0.25 | | PA | 0.23 | 0.25 | 0.18 | 0.24 | |
| SY | 0.12 | 0.16 | | | SY | 0.23 | 0.26 | 0.18 | | |
| TN | 0.17 | | | | TN | 0.18 | 0.18 | | | |
| | | | | | JO | 0.21 | | | | |

| | SDC | | | | WikiDocs corpus | |
|---|---|---|---|---|---|---|
| | LB | JO | SY | | | EG |
| PA | 0.15 | 0.21 | 0.19 | | MSA | 0.1 |
| SY | 0.16 | 0.2 | | | | |
| JO | 0.16 | | | | | |

- PAL is the most similar to MSA, that coming after the EGY
- The measurement on the SDC shows a reasonable overlapping across the Levantine dialects
- In the comparable corpus the overlapping between the MSA and the Egyptian does not exceed the 0.1

CLASP centre for linguistic theory and studies in probability

# Vector Space Model (VSM)

VSM is broken down into three steps

1. **Document indexing** where each document is represented by the content bearing words (document-terms vector)
2. **Term weighting** : employ the frequency of occurrence expressed as a ration between frequency and inverse document frequency (tf-idf)
3. **Similarity coefficient** : cosine similarity is computed between each pair of vectors to indicate a ranking of documents

CLASP centre for linguistic theory and studies in probability

# Vector Space Model (VSM)

| | PADIC | | | | | Multi-dialect corpus | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | ALG | TN | SY | PA | | EG | JO | TN | SY | PA |
| MSA | 0.27 | 0.38 | 0.37 | 0.5 | MSA | 0.5 | 0.38 | 0.37 | 0.4 | 0.4 |
| PA | 0.38 | 0.47 | 0.63 | | PA | 0.59 | 0.66 | 0.48 | 0.62 | |
| SY | 0.34 | 0.41 | | | SY | 0.63 | 0.7 | 0.5 | | |
| TN | 0.44 | | | | TN | 0.49 | 0.47 | | | |
| | | | | | JO | 0.56 | | | | |

| | SDC | | | | WikiDocs corpus |
|---|---|---|---|---|---|
| | LB | JO | SY | | EG |
| PA | 0.84 | 0.86 | 0.77 | MSA | 0.4 |
| SY | 0.81 | 0.9 | | | |
| JO | 0.84 | | | | |

- PAL in both the PADIC and the Multi dialect corpus are closer to MSA, with 0.5 and 0.4 similarity
- TN and ALG are furthest from MSA.
- on SDC we can demonstrate a high similarity between individual LEV .

CLASP centre for linguistic theory and studies in probability

# Latent Semantic Indexing (LSI)

- Analyzes the documents in order to represent the concepts they contain

CLASP centre for linguistic theory and studies in probability

# Latent Semantic Indexing (LSI)

- Analyzes the documents in order to represent the concepts they contain
- Map the vector space into a new compressed space by reducing the dimensions of the terms matrix using Singular Value Decomposition (SVD)

CLASP centre for linguistic theory and studies in probability

# Latent Semantic Indexing (LSI)

- Analyzes the documents in order to represent the concepts they contain
- Map the vector space into a new compressed space by reducing the dimensions of the terms matrix using Singular Value Decomposition (SVD)
- **Method**: We build the model with all the dialects and test it on one dialect in each run. The model outputs the similarity between the test dialect and every dialect used to build the model

CLASP centre for linguistic theory and studies in probability

# Latent Semantic Indexing (LSI)

| PADIC | ALG | TN | SY | PA |
|---|---|---|---|---|
| MSA | 0.68 | 0.75 | 0.69 | 0.75 |
| PA | 0.78 | 0.82 | 0.85 | |
| SY | 0.74 | 0.74 | | |
| TN | 0.82 | | | |

| Multi-dialect corpus | EG | JO | TN | SY | PA |
|---|---|---|---|---|---|
| MSA | 0.72 | 0.37 | 0.75 | 0.4 | 0.41 |
| PA | 0.82 | 0.88 | 0.63 | 0.9 | |
| SY | 0.7 | 0.94 | 0.59 | | |
| TN | 0.74 | 0.55 | | | |
| JO | 0.73 | | | | |

| SDC | LB | JO | SY |
|---|---|---|---|
| PA | 0.84 | 0.86 | 0.77 |
| SY | 0.81 | 0.9 | |
| JO | 0.84 | | |

| WikiDocs corpus | EG |
|---|---|
| MSA | 0.8 |

- PAL appears to be close to MSA only in PADIC
- TN shows a close relation to MSA in both corpora
- The relation between the dialects in the (SDC) is very strong as well as the relation between the ALG and TN

CLASP centre for linguistic theory and studies in probability

# Hellinger Distance

- Hellinger Distance (HD) measures the difference between two probability distributions

# Hellinger Distance

- Hellinger Distance (HD) measures the difference between two probability distributions
- Method:
  1. **A Bag Of Words BOW** model is used to represent the data from our corpora

CLASP centre for linguistic theory and studies in probability

# Hellinger Distance

- Hellinger Distance (HD) measures the difference between two probability distributions
- Method:
  1. **A Bag Of Words BOW** model is used to represent the data from our corpora
  2. **Latent Dirichlet Allocation LDA** gives us a probability distribution over a specified number of unknown topics

CLASP centre for linguistic theory and studies in probability

# Hellinger Distance

- Hellinger Distance (HD) measures the difference between two probability distributions
- Method:
  1. **A Bag Of Words BOW** model is used to represent the data from our corpora
  2. **Latent Dirichlet Allocation LDA** gives us a probability distribution over a specified number of unknown topics
  3. **Hellinger Distance HD** is then used to measure the distance between these topics and new documents (dialect)

CLASP centre for linguistic theory and studies in probability

# Hellinger Distance

- Hellinger Distance (HD) measures the difference between two probability distributions
- Method:
  1. **A Bag Of Words BOW** model is used to represent the data from our corpora
  2. **Latent Dirichlet Allocation LDA** gives us a probability distribution over a specified number of unknown topics
  3. **Hellinger Distance HD** is then used to measure the distance between these topics and new documents (dialect)
- **The greater the distance the less the similarity between the dialects and vice versa**

CLASP centre for linguistic theory and studies in probability

# Hellinger Distance

| | PADIC ALG | TN | SY | PA |
|---|---|---|---|---|
| MSA | 0.91 | 0.83 | 0.77 | 0.77 |
| PA | 0.73 | 0.64 | 0.58 | |
| SY | 0.87 | 0.81 | | |
| TN | 0.72 | | | |

| | Multi-dialect corpus EG | JO | TN | SY | PA |
|---|---|---|---|---|---|
| MSA | 0.01 | 0.77 | 0.76 | 0.78 | 0.78 |
| PA | 0.52 | 0.34 | 0.77 | 0.55 | |
| SY | 0.53 | 0.54 | 0.72 | | |
| TN | 0.35 | 0.69 | | | |
| JO | 0.51 | | | | |

| | SDC LB | JO | SY |
|---|---|---|---|
| PA | 0.26 | 0.18 | 0.23 |
| SY | 0.25 | 0.1 | |
| JO | 0.2 | | |

| | WikiDocs corpus EG |
|---|---|
| MSA | 0.73 |

- PAL and SY are both less dissimilar from MSA compared to the rest of the dialects in PADIC
- In Multi-dialect corpus, the TN seems to be the closest to MSA
- In SDC, the JO and the SY dialects are the closest to each other, while the PAL and the LEB dialects are most dissimilar

# Frequent words and Correlation Coefficient

- Extract the 30 most frequent words in each dialect
- Collect those words that appear in all dialects (10 words)
- Calculate the **Pearson correlation coefficient** among them in respect to their frequency

NOTE we have **NOT** eliminated stop words from the corpora as these keywords are discriminative and representative for each dialect and hence can be used to build a dialectal lexicon

CLASP centre for linguistic theory and studies in probability

# Frequent words and Correlation Coefficient

| | PADIC | | | | SDC | | | |
|---|---|---|---|---|---|---|---|---|
| | ALG | TN | SY | PA | | LB | JO | SY |
| MSA | 0.76 | 0.92 | 0.67 | 0.85 | PA | 0.31 | 0.42 | -0.05 |
| PA | 0.97 | 0.95 | 0.86 | | SY | 0.13 | 0.74 | |
| SY | 0.83 | 0.71 | | | JO | 0.47 | | |
| TN | 0.92 | | | | | | | |

- The result shows high correlation for the frequent words between the MSA and TN, followed by the PAL dialects in PADIC
- This sheds the light on the different usage of frequent words cross dialects. **For example** PAL speakers say عشان *ʿšān* / "because" while the SY speakers say منشان *mnšān*

CLASP centre for linguistic theory and studies in probability

# Conclusion

- Most of the measurements used indicate that the LEV are in general the closet to MSA, while the North African dialects the farthest
- Although the results show some differences due to the nature of the corpora, in general, the results are homogeneous
- We have shown the degree of convergence between the dialects of the Levant and the linguistic overlap
- **New Variety:** i.e. an informal writing dialect, which differs from the spoken dialects

# Current/future works

- appling Machine learning methods /Deep learning networks for Fine-Grained Arabic Dialect Identification.
- depending on the previous study, investigate the usage of Arabic sentiment analyzer on levantine dialects then use SDC to build a sentiment analysis corpus for levantine dialects .
- try to learn the mapping between MSA vectore embedding and dialects space.

CLASP centre for linguistic theory and studies in probability