Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 142 (2018) 2–13

Procedia Computer Science

www.elsevier.com/locate/procedia

The 4th International Conference on Arabic Computational Linguistics (ACLing 2018), November 17-19 2018, Dubai, United Arab Emirates

A Lexical Distance Study of Arabic Dialects

Kathrein Abu Kwaik[a,*], Motaz Saad[b], Stergios Chatzikyriakidis[a], Simon Dobnik[a]

[a] *CLASP, Department of Philosophy, Linguistics and Theory of Science, Ölof Wijksgatan 6, Gothenburg, 412 55, Sweden*
[b] *The Islamic University of Gaza, Gaza, Palestine*

Abstract

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. In this paper, we highlight the lexical relation between the MSA and Dialectal Arabic (DA) in more than one Arabic region. We conduct a computational cross dialectal lexical distance study to measure the similarities and differences between dialects and the MSA. We exploit several methods from Natural Language Processing (NLP) and Information Retrieval (IR) like Vector Space Model (VSM), Latent Semantic Indexing (LSI) and Hellinger Distance (HD), and apply them on different Arabic dialectal corpora. We measure the overlap among all the dialects and compute the frequencies of the most frequent words in every dialect. The results are informative and indicate that Levantine dialects are very similar to each other and furthermore, that Palestinian appears to be the closest to MSA.

Keywords: Diglossia; Lexical Distance; Vector Space Model; Latent Semantic Indexing; Hellinger Distance

1. Introduction

The number of the native Arabic speakers in the world varies from 290 million according to UNESCO[1] to 313 million, according to the Ethnologue[2]. There are three varieties in Arabic language: Classical Arabic, Modern Standard Arabic (MSA), and Arabic dialects (Colloquialism). Classical Arabic (CA) is the form of the Arabic language used in Umayyad and Abbasid literary texts from the 7th century AD to the 9th century AD. The orthography of the Quran was not developed for the standardized form of Classical Arabic [1]. MSA is the official language used for education, news, politics, religion and, in general, in any type of formal setting. Colloquialisms (dialects) are used in everyday communication as well as informal writing, e.g. in social media [2].

* Corresponding author.
E-mail address: kathrein.abu.kwaik@gu.se

[1] https://en.unesco.org/news/world-arabic-language-day-2017-looking-digital-world

[2] Simons, Gary F. and Charles D. Fennig (eds.). 2018. Ethnologue: Languages of the World, Twenty-first edition. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com.

1877-0509
10.1016/j.procs.2018.10.456

As a result of this situation, diglossia, a case where two distinct varieties of a language are spoken within the same speech community [3], is a very common phenomenon in Arabic-speaking communities. In some parts of the Arab speaking world, more than two varieties are spoken within the same community. For example, this is the case in North African communities like Morocco where Arabic, Berber, French, English and Spanish are spoken within the same speech community [4]. In a diglossic situation, the standard formal language assumes the role of the High variety (H), while the other languages or dialects act as the Low variety (L) [5]. MSA is so different from the colloquial dialects that they are in some cases not mutually intelligible. The differences are clearly evident in all linguistic aspects: pronunciation, phonology, morphology, lexicon, syntax and semantics. However, the degree in which the individual dialects differ with respect to these aspects has not been yet quantitatively measured.

In this paper, we focus on measuring the lexical distance between MSA and Arabic dialects using natural language processing techniques, tools and text corpora. We use various distance metrics such as the Vector space model (VSM) based on word distribution over documents as common in Information Retrieval (IR) [6], Latent semantic indexing (LSI) [7] and the Divergence Distance algorithm as Hellinger Distance (HD) [8]. We hope that this study will shed light on similarities and differences between the varieties and therefore inform our future work on building NLP tools and applications for these domains, in particular how these can be ported.

To the best of our knowledge, our work is the most extensive effort to measure the distance or similarity across Arabic dialects using natural language processing tools and text corpora.

2. Related Work

Several approaches have been used to measure the distance between European languages [9, 10], Indian dialects [11], and similar languages [12, 13]. These approaches can be classified according to the type of linguistic representations they investigate: characters, terms and documents. Lexical similarity measures operate on string sequences at a character level and on corpora of texts at a term level. Table 1 shows the most popular approaches in the literature. There is not much research on measuring the lexical closeness and divergence between Arabic and its dialects.

Table 1. Summary of the most commonly used approaches for measuring similarity between texts

Approach Level	Approach Name	Description
Character Level	Longest Common SubString	Measures the length of the longest contiguous sequence of characters existing in the string under comparison [14, 15].
	Levenshtein distance	Measure the minimum number of insertions, deletions and substitutions needed to transform one string into another [16, 17].
	N-gram models	Can be used in different ways to estimate similarity or dissimilarity. One of the most effective approaches is to build n-gram models for Language Identification and measure the perplexity of n-grams [18].
	Dynamic programming	Used for biological sequence comparison; e.g. the Needleman-Wunsch and Smith-Waterman algorithms [19].
Term Level	Vector space models	Represent documents as vectors of word frequencies and then apply vector comparison measures to compare vectors of different documents [20].
	Cosine Similarity	Measures the cosine angle as a similarity indicator between two vector spaces [21].
	Divergence Distance	For example, the Kullback-Leibler distance, Hellinger, Manhattan distance, etc. These are used to measure the divergence between probability distributions [22, 23]
	Jacquard similarity	Measures the number of overlapping strings over the number of unique strings between texts to indicate the similarity [24].
	Latent Semantic Indexing	Words that are close in meaning will occur frequently in similar positions in the text [25].

Abunasser in [16] compares five Arabic dialects (MSA, Gulf, Levantine, Egyptian and Morocco) in terms of lexical and pronunciation variation. He depends on the Swadesh list [26] and the concept of non-cognate words to measure the amount of linguistic variations between the dialects. As the Swadesh list is a phonological list rather than a lexicon, the author collected the data from two male speakers for each dialect. The Swadesh list has been adapted to the MSA list using two modern Arabic dictionaries (المورد *ālmwrd* [27] and قاموس ابن اياس *qāmws ābn āyās* [28]). To rule out the chance of lexical ambiguity, a context sentence per each lexical item has been provided. Thus, the

distance between dialects is measured based on the percentage of non-cognates in the MSA Swadesh list. Moreover, he employs Levenshtein distance to compute the distance between lexical items at the phonemic level based on the IPA transcription of the words in the Swadesh list. He concludes that Gulf and Levantine are the closet dialects to MSA followed by Egyptian, while Morocco is the farthest. The most significant limitation of this experiment is how the data were collected where speakers, gender and the geographical location were limited to two male speakers per dialect only. Also, the two modern dictionaries that are used to translate the Swadesh list to the corresponding MSA list have been authored by Levantine authors which might bias MSA to Levantine to some degree. Finally, with the intention to measure lexical variation, the study uses phonemic representation of words which may also reveal other more subtle non-lexical differences.

Meftouh et al. [22] present PADIC (Parallel Arabic Dialect Corpus). It includes five dialects: two Algerian (from the cities of Algiers and Annaba), one Tunisian and two Levantine dialects (Palestinian and Syrian). The authors present a linguistic analytical study of PADIC where they employ experiments on every pair of dialect and MSA, including:

- identifying the most frequent words in each dialect;
- computing the percentage of common lexical units both at the document and the sentence level to emphasize the relation between the dialects and the MSA; and
- measuring the cross language divergence in terms of the Hellinger distance to measure which language is closer to which one.

The experiments have shown that the Palestinian dialect is the closest dialect to MSA followed by Tunisian and Syrian, whereas Algerian dialects are the most different. The results are expected, as they demonstrate that Tunisian is closer to Algerian than to Palestinian and Syrian. In addition, the closest dialects according to the distance measurements are Algerian dialects on one hand and Palestinian and Syrian on the other hand. Even though the results are reasonable, the corpus has a shortcoming that it has been manually translated from Algerian conversations to MSA and further to other dialects by one native speaker of each dialect which introduces several biases.

Rama et al. [12] present a computational classification of the Gondi dialects which are spoken in central India by applying tools from dialectometry and phylogenetics. They use multilingual word lists for 210 concepts at 46 sites where Gondi is the dominate dialect. They depend on the Glottolog classification as a gold standard to evaluate their results. To be able to compute the aggregate distances, they employ the IPA to convert the word lists to pronunciation data. Levenshtein distance and Long-Short Term Memory neural networks are used as dialectometry methods to measure the distance between every word pair of words on the list. Moreover, they also apply Bayesian analysis on cognate analysis as a phylogentic method. They find that phylogentic methods perform best when compared to the gold standard classification.

Ruette et al. [20] measure the distance between Belgian and Netherlandic Dutch using two similarity measures in the Vector Space Model (VSM). They apply the two methods on a Dutch corpus collected from two registers (quality newspapers and Usenet) and topics related to politics and economy. They also exploit the profile-based approach (where the frequency of pre-selected words is compared from speakers' data) in addition to the text categorization method. For the profile based approach they implement the City-Block distance as a straightforward descriptive distance measure. On the other hand, text categorisation is using TFxIDF on documents and cosine similarity to measure distance as the complement of cosine similarity.

3. Qualitative differences between MSA and DA

Arabic is characterized by its rich morphology and vocabulary. For instance the Arabic word وسيعطيك *wsyʕṭyk* means "and he will give you" in English so one word in Arabic may correspond to 5 words in English [29] and that would make the comparison between languages/dialects challenging. This is true for both MSA as well as dialectal Arabic. However, MSA and DA have a number of differences that make it difficult for one to apply state of the art MSA natural language processing tools to DA. Previous attempts to do so have resulted in very low performance due to the significant difference between the varieties. [30] report that over one third of Levantine verbs cannot be analysed using an MSA morphological analyser. The degree of variation between MSA and dialectal Arabic depends

on the specific dialect of Arabic. MSA and dialectal Arabic differ to a different degree phonologically, orthographically, morphologically, syntactically, lexically and semantically [31, 32]. In this section we describe some qualitative differences between MSA and the dialects based on our observation of examples.

3.1. Orthographical and Phonological Differences

Dialectal Arabic (DA) does not have an established standard orthography like MSA. Mostly, Arabic script is used to write DA but in some cases, e.g. in Lebanese, 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*. Another example is the 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* [33].

3.2. Morphological Differences

Dialects, like MSA and other Semitic languages, make extensive use of particular morphological patterns in addition to a large set of affixes (prefixes, suffixes, or infixes) and clitics, and therefore 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 [34]. Some examples are illustrated in Table 2 and 3.

Table 2. Examples for Morphological differences

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

3.3. Syntactic Differences

Syntactically, MSA and DA are very similar with some differences regarding word order. For example, the OVS and OSV word orders are most commonly used in MSA while in dialects other word order patterns can be found. For example, in Levantine SVO is most commonly used, while in Maghrebi VSO is used to a great extent [35]. Furthermore, in dialectal Arabic it is common to use masculine plural or singular forms instead of dual and feminine plural forms [36].

3.4. Lexical and Semantic differences

Many DA words are borrowed from a variety of other languages like Turkish, French, English, Hebrew, Persian and others depending on the speaker contact with these languages. Table 4 shows some of the borrowed words. New

Table 3. Differences in negation between the dialects

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*		

lexical items appear mostly in dialects and not MSA as shown by the example in in Table 5. Another thing to note is dialects and MSA share words but with different meanings. For example, the word دول *dwl* means 'these' in Egyptian but "countries" in MSA.

Table 4. Examples of borrow words from other languages

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

Table 5. Examples for new lexicon in dialects

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

4. Quantitative differences between MSA and DA

4.1. Arabic Corpora

Ferguson [3] was the first to define the term diglossia. He stated and defined the most important features in order to understand the difference between the official languages (H) and the informal varieties (L). One of these features is the lexicon. In his own words: *"A striking feature of Diglossia is the existence of many paired items, one H and one L, referring to fairly common concepts frequently used in both H and L, where the range of meaning of the two items is roughly the same, and the use of one or the other immediately stamps the utterance or written sequence as H or L"*.

In this work, we examine several existing Arabic corpora, so that we can include as many dialectal data as we can. Table 6 shows the corpora we use and the dialectal data they contain. Table 7 shows the statistics about each corpus where |d| is the number of documents (sentences) in the corpus, |w| is the number of words in the corpus, and |v| is the vocabulary size (number of unique words).

Table 6. List of Arabic corpora used to investigate the differences between dialects

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.

The two Algerian dialects are the basis of the Parallel Arabic Dialect Corpus (PADIC), that was collected from daily conversations, movies and TV shows were presented in Annaba and Algeria dialects. The two corpora were transcribed by hand and then translated to MSA. Hence, the MSA is considered the pivot language to construct the Syrian, Palestinian and Tunisian dialects. They adopt Arabic notation to write dialectal words. If the dialectal word does exist in MSA, it is written as MSA without any change, otherwise, it is written as it is uttered. Some consider these rules as drawbacks of the corpus which bias the dialect to the MSA and the translated sentence is subjected to the annotators [22]. The corpus is not considered fully representative for every dialect due to the lack of translators particularly for the Levantine dialects where only 2 translators are involved while for Tunisian they had 20 speakers all of them from the South of Tunisia where their dialect is close to the Standard Arabic.

The Multi-dialectal Arabic parallel corpus is built on the English-Egyptian corpus [37], where the Egyptian sentences have been selected as the starting point for the new parallel corpus. Five translators, one for every dialect, were asked to translate the Egyptian corpus to Palestinian, Jordanian, Syrian and Tunisian dialects, while the Egyptian speakers translated the corpus to corresponding MSA [38]. Using the Egyptian sentences as the pivot dialect makes the corpus heavily influenced and biased by the Egyptian dialects, which is clearly shown in our results in the following sections.

The WikiDocs corpus is extracted from Arabic Wikipedia articles and their corresponding Egyptian Wikipedia articles [39]. It should be noted that a lot of the Egyptian articles are not detailed, as most of these only contain one or two sentences. This is in contrast to the MSA articles, which contain full details on each subject. The Shami Dialect Corpus (SDC) corpus is collected from different domains like social life, sports, house work, cooking, etc. and from resources such as personal blogs, social media public figures posts and stories written in DA. It focuses on public figures from Levantine countries. It is not a parallel corpus, thus the measures are done over the whole corpus and not on every document [32].

Table 7. Statistics about the used corpora

	PADIC					**SDC**			
	MSA	PA	AlG	SY	Tn	PA	JO	SY	LB
$\|d\|$	6.4K	6.4K	6.4K	6.4K	6.4K	21K	32K	48K	16K
$\|w\|$	51K	51K	48K	49K	48K	0.35M	0.47M	0.7M	0.2M
$\|v\|$	9.4K	9.6K	9.4K	10K	10.6K	56K	69K	63K	34K
	Multi-dialect corpus						**WikiDocs corpus**		
	MSA	PA	JO	SY	TN	EG	MSA		EG
$\|d\|$	1 K	1 K	1 K	1 K	1 K	1K	459K		16K
$\|w\|$	11.9K	10.5K	9.7K	11.5K	10.6K	10.9K	83.5M		2.18M
$\|v\|$	4.4K	4K	3.6K	4K	3.8K	4.5K	4.7M		293.5K

In what follows, we exploit various approaches to the lexicon to precisely clarify the difference between MSA and other Arabic dialects in term of lexical distance. The type of corpora affects the way we implement each measure as follows:

- for parallel and comparable corpora: the comparison is at the document (sentence) level, then the average is taken at a corpus level;
- for non-parallel and non-comparable corpora: the comparison is at the corpus level, given that the data belong to the same domain.

In all experiments we have used Python as a programming language to implement the algorithms and used the Gensim library for some methods. As the corpora are already preprocessed, we did not do any further pre-processing.[3] In the next subsections, we present the measures what we use in our experiments.

4.2. Lexical Sharing and Overlapping

Jaccard Index is a measure of how similar two data sets are. Given that dialects share many words, we compute the percentage of vocabularies that overlap between these dialects according to Equation 1. Table 8 presents the similarity overlap across dialects. Palestinian is the most similar to MSA, that coming after the Egyptian dialect, with the highest percentage of vocabulary overlap in both parallel corpora. The measurement on the SDC shows a reasonable overlapping across the Levantine dialects, while in the comparable corpus the overlapping between the MSA and the Egyptian does not exceed the 0.1.

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

Table 8. Percentage of vocabulary overlapping between dialects

	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									

4.3. Vector Space Model (VSM)

VSM is broken down into three steps. First, document indexing where each document is represented by the content bearing words which, in turn, are represented as a document-terms vector. VSM represents all documents as vectors in a high dimensional space in which each dimension of the space corresponds to a term in the document collection [7]. Secondly, term weighting where a weighting schema is used to compute the term weightings for each term in the represented vector (document). The most common weighting schema is to employ the frequency of occurrence

[3] It is possible that the preprocessing techniques that have been used on different corpora might affect their comparison, which is an unfortunate limitation of our approach in terms of the implications for language use in general.

expressed as a ration between frequency and inverse document frequency (tf-idf). A similarity coefficient is then computed between each pair of vectors to indicate a ranking of documents [40].

We utilize the VSM to measure the similarity across dialects and MSA by comparing the similarity between the terms in their documents or sentences. Clearly, not all words in a dialect or a document are equally important. Most current approaches remove all the stop words during the preprocessing phase. However, we have decided to index all words as many of the stop words act like function words and therefore are distinguishing of certain dialects. In order to overcome the out-of-dictionary problem we build a vector for each pair of dialects. Therefore, for the first dialect (MSA), we draw a vector model and employ the tf-idf weighting schema. The second dialect is considered as the query vector compared to the first dialect. Spatial closeness corresponds to conceptual similarity (words that are used in the same documents are similar) so we measure the cosine similarity between the main vector model (first-dialect) and the query vector (second-dialect) (what a vector represents in each case depends on the kind of corpora we are comparing as explained above) which is a symmetric measurement. Table 9 present the similarity across dialects for all corpora.

Table 9. Similarity across dialects for all corpora based on 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									

The results show that the Palestinian dialects in both the PADIC and the Multi dialect corpus are closer to MSA, with 0.5 and 0.4 similarity respectively, while the Tunisian and Algerian dialects are furthest from MSA. Moreover, on SDC we can demonstrate a high similarity between individual Levantine dialects. For example Jordanian is the closest to Palestinian, which seems to coincide with informal observations by native speakers of both dialects.

It is worth mentioning that the Egyptian dialect records the highest relation with MSA in Multi-dialect corpus, as we previously expected. The corpus is biased towards the Egyptian dialect, as Egyptian was the pivot language when the corpus was built. This is reflected in all the measures used here. However, the bias of the pivot language is not reflected between Algerian and MSA in the PADIC corpus as these are the least similar varieties.

4.4. Latent Semantic Indexing LSI

Unlike VSM and other retrieval methods, LSI can address the problem of synonymy and polysemy among words. It analyzes the documents in order to represent the concepts they contain. LSI tries to map the vector space into a new compressed space by reducing the dimensions of the terms matrix using Singular Value Decomposition (SVD). By using SVD, the main associative patterns and trends are extracted from the document space and the noise is ignored. In other words, it makes the best possible reconstruction of the document matrix with the most valuable information [7]. We exploit the LSI model to measure the similarity between the dialects. We build the model with all the dialects (full corpus) 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. Table 10 shows the similarities among the Arabic dialects for all corpora.

Palestinian appears to be close to MSA only in PADIC, whereas the Tunisian dialect shows a close relation to MSA in both corpora. In addition to this, it is obvious that the relation between the dialects in the Levantine corpus (SDC) is very strong as well as the relation between the Algerian and Tunisian. These results show the artefacts of the LSI model which connects the data according to topics and clusters.

Table 10. Similarity across dialects for all corpora based on LSI

	PADIC				**Multi-dialect corpus**					
	ALG	TN	SY	PA		EG	JO	TN	SY	PA
MSA	0.68	0.75	0.69	0.75	MSA	0.72	0.37	0.75	0.4	0.41
PA	0.78	0.82	0.85		PA	0.82	0.88	0.63	0.9	
SY	0.74	0.74			SY	0.7	0.94	0.59		
TN	0.82				TN	0.74	0.55			
					JO	0.73				
	SDC				**WikiDocs corpus**					
	LB	JO	SY			EG				
PA	0.84	0.86	0.77		MSA	0.8				
SY	0.81	0.9								
JO	0.84									

4.5. Hellinger Distance

We are interested to measure the divergence between the dialects. Here we will use the Hellinger Distance (HD) that measures the difference between two probability distributions [22]. In this work we use Latent Dirichlet Allocation (LDA) to model a vector of discrete probability distributions of topics to measure the distance between dialects in comparison. LDA is a very common technique used to uncover topics in the data [41]. For simplicity, a Bag Of Words (BOW) model is used to represent the data from our corpora. LDA gives us a probability distribution over a specified number of unknown topics. LDA therefore works like a way of soft clustering the documents made up of words. Later HD is then used to measure the distance between these topics and new documents. The greater the distance the less the similarity between the dialects and vice versa.

Table 11 shows the distance between the dialects cross all corpora. Palestinian is less dissimilar from MSA compared to the rest of the dialects in PADIC. Even though in the Multi-dialect corpus the results for the distance of all dialects, except of the Egyptian, to MSA is quite close, the Tunisian seems to be the closest to MSA. Considering that the Levantine dialects in SDC are very similar to each other, the Jordanian and the Syrian dialects are the closest to each other, while the Palestinian and the Lebanese dialects are most dissimilar.

Table 11. Distance between dialects for all corpora based on Hellinger Distance

	PADIC				**Multi-dialect corpus**					
	ALG	TN	SY	PA		EG	JO	TN	SY	PA
MSA	0.91	0.83	0.77	0.77	MSA	0.01	0.77	0.76	0.78	0.78
PA	0.73	0.64	0.58		PA	0.52	0.34	0.77	0.55	
SY	0.87	0.81			SY	0.53	0.54	0.72		
TN	0.72				TN	0.35	0.69			
					JO	0.51				
	SDC				**WikiDocs corpus**					
	LB	JO	SY			EG				
PA	0.26	0.18	0.23		MSA	0.73				
SY	0.25	0.1								
JO	0.2									

4.6. Frequent words and Correlation Coefficient

This step consists of two parts. At first, we extract the 30 most frequent words in each dialect and then we collect those words that appear in all dialects to calculate the Pearson correlation coefficient among them in respect to their frequency as shown in Table 12.

Table 12. The Person correlation coefficient between dialects in PADIC and SDC

	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 Tunisian, followed by the Palestinian dialects in PADIC. This sheds the light on the different usage of frequent words cross dialects. For example Palestinian speakers say في المدرسة *fy ālmdrsh* / "at the school" while the Syrian speakers say بالمدرسة *bālmdrsh*.

For the words that are not shared and have not been included in the correlation experiment, we have calculated the Term Frequency (TF) as in Equation 2.

$$TF(t) = \frac{(Number\ of\ times\ term\ t\ appears\ in\ a\ dialect)}{(Total\ number\ of\ terms\ in\ the\ dialect)}. \quad (2)$$

As we have already mentioned, 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. Table 13 shows the 20 most frequent words in PADIC[4].

5. Conclusion

In this paper, we estimate the degree of similarity and dissimilarity between MSA and DA on one hand, and across dialects of Arabic on the other. Different measures have been exploited, such as as VSM, LSI, HD as well as simple measures like vocabulary overlap, coefficient correlation and Jaccard similarity. More than one corpus has been used. In particular, PADIC, the Multi-Dialect corpus, SDC and Wiki-Docs were used, that include MSA, Levantine dialects, Egyptian and Dialects from North Africa. This was done in order to minimise the bias of any of the individual corpora and to address the question of the degree of the text representativeness. Most of the measurements used indicate that the Levantine dialects 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. For example, it is expected that the Egyptian dialects appear very close to MSA in the Multi-Dialect corpus. This is, as mentioned earlier, due to a strong bias of the specific corpus towards the Egyptian dialect, given that it was built from an Egyptian corpus and then translated into other dialects and MSA.

We have shown the degree of convergence between the dialects of the Levant and the linguistic overlap to such an extent that in some cases it seems impossible to distinguish between them in writing without the presence of phonological information or without adding accent diacritic marks.

It is very clear that we have a new variety, i.e. an informal writing dialect, which differs from the spoken dialects. Even if some dialects appear close to each other based on the speakers' intuitions, there may be differences in the writing form due to the lack of accent diacritics. The reverse is also true. Some dialects appear closer lexically in their writing form given that a big part of their vocabulary overlaps, but in their spoken form, they are not that close.

This study can be seen as a basis for building Natural Language Processing tools for dialectal processing by adapting what already exists for MSA and focusing on areas of similarity and degrees of difference. The study is the

[4] The full tables for all corpora can be found in https://github.com/GU-CLASP/DAdistance

Table 13. The percentage of the most frequent words in PADIC

MSA Word	TF%	Palestinian Word	TF %	Syrian Word	TF %	Tunisian Word	TF%	Algerian Word	TF%
لا *lā*	1.96	اللي *ālly*	0.84	بس *bs*	0.98	باش *bāš*	0.85	لي *ly*	1.14
أن *ʾan*	1.44	انه *ānh*	0.83	اي *āy*	0.92	الي *āly*	0.78	واش *wāš*	1
لم *lm*	0.81	بس *bs*	0.81	عم *ʿm*	0.89	ايه *āyh*	0.73	ايه *āyh*	0.82
لي *ly*	0.7	ايش *āyš*	0.8	شو *šw*	0.88	لا *lā*	0.72	تاع *tāʿ*	0.79
نعم *nʿm*	0.7	مش *mš*	0.79	رح *rḥ*	0.85	اما *āmā*	0.59	كي *ky*	0.67
هذا *hḏā*	0.65	اه *āh*	0.77	شي *šy*	0.73	كان *kān*	0.46	لالا *lālā*	0.59
ماذا *māḏā*	0.47	لا *lā*	0.65	لا *lā*	0.7	هذا *hḏā*	0.44	واحد *wāḥd*	0.48
إلى *ʾilā*	0.45	هذا *hḏā*	0.64	انو *ānw*	0.51	تو *tw*	0.37	ولا *wlā*	0.4
هل *hl*	0.45	اشي *āšy*	0.55	مو *mw*	0.48	علاش *ʿlāš*	0.35	راني *rāny*	0.38
ذلك *ḏlk*	0.42	لما *lmā*	0.53	كنير *knyr*	0.47	حتى *ḥtā*	0.34	باش *bāš*	0.37
لكن *lkn*	0.42	هو *hw*	0.5	لما *lmā*	0.45	باهي *bāhy*	0.33	حتى *ḥtā*	0.36
لك *lk*	0.39	عشان *ʿšān*	0.45	اللي *ālly*	0.44	هو *hw*	0.31	والله *wāllh*	0.34
عندما *ʿndmā*	0.39	هيك *hyk*	0.44	هيك *hyk*	0.37	وقت *wqt*	0.31	هو *hw*	0.32
قلت *qlt*	0.83	اذا *āḏā*	0.44	هاد *hād*	0.36	موش *mwš*	0.29	راح *rāḥ*	0.32
إذا *ʾiḏā*	0.35	كتير *ktyr*	0.4	الله *āllh*	0.35	واحد *wāḥd*	0.25	بالصح *bālṣḥ*	0.32
لها *lhā*	0.34	الله *āllh*	0.36	ليش *lyš*	0.34		0.25	دوك *dwk*	0.31
هناك *hnāk*	0.41	هذه *hḏh*	0.35	اذا *āḏā*	0.31	برشه *bršh*	0.25	كيما *kymā*	0.31
الله *āllh*	0.32	زي *zy*	0.34	متل *mtl*	0.31	اللي *ālly*	0.24	برك *brk*	0.3
له *lh*	0.32	ليش *lyš*	0.33	عن *ʿn*	0.29	شي *šy*	0.23	راهي *rāhy*	0.3
شيء *šyʾ*	0.32	اني *āny*	0.3	كان *kān*	0.28	ولا *wlā*	0.23	صح *ṣḥ*	0.29

most extensive of its kind concerned with measuring similarities and differences in Arabic and dialectal Arabic, and represents a basis for new similar investigations, focusing on other criteria such as phonological distance, morphological distance and semantic distance. In the future, we plan to employ other methods of measuring similarity and distance based on the semantics of the words, e.g. word embedding techniques with Word2Vec. In this way, one can extract different words in terms of their lexical relatedness, and use them in automatic machine translation tools for the languages and dialects investigated.

References

[1] Shah, Mustafa., *The Arabic language*, Routledge, 2008.

[2] Versteegh, Kees, *The Arabic language*, Edinburgh University Press, 2014.

[3] Ferguson, Charles A., Diglossia, word 15 (2) (1959) 325–340.

[4] Zouhir, Abderrahman., Language situation and conflict in Morocco, in: *Selected Proceedings of the 43rd Annual Conference on African Linguistics*, ed. Olanike Ola Orie and Karen W. Sanders, 2013, pp. 271–277.

[5] Jabbari, MJ., Diglossia in Arabic – a comparative study of the Modern Standard Arabic and colloquial Egyptian Arabic, Global *Journal of Human Social Sciences* 12 (8) (2012) 23–46.

[6] Clark, Stephen, Vector space models of lexical meaning, in: Lappin, Shalom and FoxS, Chris (Eds.). *Handbook of Contemporary Semantics – second edition*, Wiley – Blackwell, 2015, Ch. 16, pp. 493–522.

[7] Kumar, Ch Aswani, M Radvansky, and J Annapurna, Analysis of a Vector Space Model, Latent Semantic Indexing and formal concept analysis for Information Retrieval, *Cybernetics and Information Technologies* 12 (1) (2012) 34–48.

[8] González-Castro, Víctor, Rocío Alaiz-Rodríguez, and Enrique Alegre. Class distribution estimation based on the Hellinger distance, *Information Sciences 218* (2013) 146–164.

[9] Chiswick, Barry R and Paul W Miller. Linguistic distance: A quantitative measure of the distance between English and other languages, *Journal of Multilingual and Multicultural Development* 26 (1) (2005) 1–11.

[10] Heeringa, Wilbert, Jelena Golubovic, Charlotte Gooskens, Anja Schüppert, Femke Swarte, and Stefanie Voigt. Lexical and orthographic distances between Germanic, Romance and Slavic languages and their relationship to geographic distance, *Phonetics in Europe: Perception and Production* (2013) 99–137.

[11] Sengupta, Debapriya and Goutam Saha. Study on similarity among Indian languages using language verification framework, *Advances in Artificial Intelligence* 2015 (2015) 2.

[12] Rama, Taraka, Çağrı Çöltekin, and Pavel Sofroniev, Computational analysis of Gondi dialects, in: *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects* (VarDial), 2017, pp. 26–35.

[13] Houtzagers, Peter, John Nerbonne, and Jelena Prokić, Quantitative and traditional classifications of Bulgarian dialects compared, Scando-Slavica 56 (2) (2010) 163–188.

[14] Aminul Islam and Diana Inkpen, Semantic text similarity using corpus-based word similarity and string similarity, *ACM Transactions on Knowledge Discovery from Data* (TKDD) 2 (2) (2008) 10.

[15] Robert W Irving and Campbell B Fraser. Two algorithms for the longest common subsequence of three (or more) strings, in: Annual Symposium on Combinatorial Pattern Matching, Springer, 1992, pp. 214–229.

[16] Abunasser, Mahmoud Abedel Kader. Computational measures of linguistic variation: A study of Arabic varieties, Ph.D. thesis, University of Illinois at Urbana-Champaign (2015).

[17] Navarro, Gonzalo. A guided tour to approximate string matching, ACM computing surveys (CSUR) 33 (1) (2001) 31–88.

[18] Kondrak, Grzegorz. N-gram similarity and distance, in: International symposium on string processing and information retrieval, Springer, 2005, pp. 115–126.

[19] Needleman, Saul B and Christian D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins, *Journal of molecular biology* 48 (3) (1970) 443–453.

[20] Ruette, Tom, Dirk Speelman, and Dirk Geeraerts. Measuring the lexical distance between registers in national variaties of Dutch, Aletheia, Publicações da Faculdade de Filosofia da Universidade Católica Portuguesa, 2011.

[21] Anna Huang. Similarity measures for text document clustering, in: *Proceedings of the sixth New Zealand computer science research student conference* (NZCSRSC2008), Christchurch, New Zealand, 2008, pp. 49–56.

[22] HarratSalima, Karima Meftouh, Mourad Abbas, Salma Jamoussi, Motaz Saad, and Kamel Smaili. Cross-dialectal Arabic processing, in: International Conference on Intelligent Text Processing and Computational Linguistics, Springer, 2015, pp. 620–632.

[23] Bigi, Brigitte. Using Kullback-Leibler distance for text categorization, in: European Conference on Information Retrieval, Springer, 2003, pp. 305–319.

[24] Niwattanakul, Suphakit, Jatsada Singthongchai, Ekkachai Naenudorn, and Supachanun Wanapu. Using of Jaccard coefficient for keywords similarity, in: *Proceedings of the International MultiConference of Engineers and Computer Scientists*, Vol. 1, 2013.

[25] Sebastiani, Fabrizio. Machine Learning in automated text categorization, ACM computing surveys (CSUR) 34 (1) (2002) 1–47.

[26] Swadesh, Morris. Salish internal relationships, *International Journal of American Linguistics* 16 (4) (1950) 157–167.

[27] Baalbaki, Munir. المورد: قاموس انجليزي ـ عربي *ālmwrd: qāmws ānğlyzy ʿrby*. بثروت: دار العلم للملايين *dār ālʿlm llmlāyyn: bṯrwt* , 1982.

[28] Elias, Elias Antoon and Ed E Elias. Elias' *modern dictionary, Arabic-English*, (1983).

[29] Saad, Motaz. Fouille de documents et d'opinions multilingue, Ph.D. thesis, Université de Lorraine (2015).

[30] Habash, Nizar and Owen Rambow. a morphological analyzer and generator for the Arabic dialects, in: *Proceedings of the 21st International Conference on Computational Linguistics* and the 44th annual meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2006, pp. 681–688.

[31] Dasigi, Pradeep and Mona T Diab. Towards identifying orthographic variants in dialectal Arabic., in: IJCNLP, 2011, pp. 318–326.

[32] Qwaider, Chatrine, Motaz Saad, Stergios Chatzikyriakidis, and Simon Dobnik. Shami: A Corpus of Levantine Arabic Dialects, in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, European Language Resources Association (ELRA), Miyazaki, Japan, 2018.

[33] Jarrar, Mustafa, Nizar Habash, Diyam Akra, and Nasser Zalmout. Building a corpus for Palestinian Arabic: A preliminary study, in: Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP), 2014, pp. 18–27.

[34] Habash, Nizar, Mona T Diab, and Owen Rambow. Conventional orthography for dialectal Arabic, in: LREC, 2012, pp. 711–718.

[35] Meftouh, Karima, Salima Harrat, Salma Jamoussi, Mourad Abbas, and Kamel Smaili. Machine translation experiments on PADIC: A parallel Arabic Dialect Corpus, in: The 29th Pacific Asia conference on language, information and computation, 2015.

[36] Darwish, Kareem, Hassan Sajjad, and Hamdy Mubarak. Verifiably effective Arabic dialect identification., in: EMNLP, 2014, pp. 1465–1468.

[37] Zbib, Rabih, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F Zaidan, and Chris Callison-Burch. Machine Translation of Arabic Dialects, in: *Proceedings of the 2012 conference of the North American Chapter of the Association for Computational Linguistics*: Human language technologies, Association for Computational Linguistics, 2012, pp. 49–59.

[38] Bouamor, Houda, Nizar Habash, and Kemal Oflazer. A Multidialectal Parallel Corpus of Arabic, in: LREC, 2014, pp. 1240–1245.

[39] Saad, Motaz and Basem O Alijla. Wikidocsaligner: An off-the-shelf Wikipedia documents alignment tool, in: *Information and Communication Technology (PICICT)*, 2017 Palestinian International Conference on, IEEE, 2017, pp. 34–39.

[40] Larson, Ray R. Introduction to Information Retrieval, *Journal of the American Society for Information Science and Technology* 61 (4) (2010) 852–853.

[41] Blei, David M, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet Allocation, *Journal of Machine Learning research* 3 (Jan) (2003) 993–1022.