

Cross-Dialectal Arabic Processing

Salima Harrat, Karima Meftouh, Mourad Abbas, Salma Jamoussi, Motaz Saad, Kamel Smaïli

▶ **To cite this version:**

Salima Harrat, Karima Meftouh, Mourad Abbas, Salma Jamoussi, Motaz Saad, et al.. Cross-Dialectal Arabic Processing. International Conference on Intelligent Text Processing and Computational Linguistics, Apr 2015, cairo, Egypt. 10.1007/978-3-319-18111-0_47 . hal-01261598

HAL Id: hal-01261598
https://hal.science/hal-01261598

Submitted on 25 Jan 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Cross-dialectal Arabic Processing

Salima Harrat[1], Karima Meftouh[2], Mourad Abbas[3], Salma Jamoussi[4], Motaz Saad[5], and Kamel Smaili[5]

[1] ENSB ⋆, Ecole Supérieure d'Informatique (ESI), Algiers, Algeria,
[2] Badji Mokhtar University, Annaba, Algeria
[3] CRSTDLA⋆⋆, Algiers, Algeria
[4] MIRACL⋆⋆⋆, Pole Technologique de Sfax, Tunisia
[5] Campus Scientifique LORIA , Nancy, France

Abstract. We present, in this paper an Arabic multi-dialect study including dialects from both the Maghreb and the Middle-east that we compare to the Modern Standard Arabic (MSA). Three dialects from Maghreb are concerned by this study: two from Algeria and one from Tunisia and two dialects from Middle-east (Syria and Palestine). The resources which have been built from scratch have lead to a collection of a multi-dialect parallel resource. Furthermore, this collection has been aligned by hand with a MSA corpus. We conducted several analytical studies in order to understand the relationship between these vernacular languages. For this, we studied the closeness between all the pairs of dialects and MSA in terms of Hellinger distance. We also performed an experiment of dialect identification. This experiment showed that neighbouring dialects as expected tend to be confused, making difficult their identification. Because the Arabic dialects are different from one region to another which make the communication between people difficult, we conducted cross-lingual machine translation between all the pairs of dialects and also with MSA. Several interesting conclusions have been carried out from this experiment.

1 Introduction

In Arab countries, the majority of people speaks dialects. Modern Standard Arabic (MSA) is the official language used only in formal speeches, media and education. What may be surprising is that even educated people, in their daily life prefer speaking the dialect which is their mother-tongue. Consequently, studying the dialects becomes a priority which could take benefit from natural language processing tools.

During the last decade, researchers have been interested to Arabic dialects processing, like building lexicon, morphological analysis, POS tagging, etc, [1–4].

⋆ Ecole Normale Supérieure Bouzareah.
⋆⋆ Centre de Recherche Scientifique et Technique pour le Développement de la langue Arabe.
⋆⋆⋆ Multimedia, InfoRmation systems and Advanced Computing Laboratory.

Recent works have been dedicated to other tasks, such as Machine Translation [5, 6] and dialect identification [7, 8]. In [9], a work of building a small multilingual dialectal corpus is presented further including the MSA.

In this paper, we will focus on a set of Arabic dialects and more particulary on three from Maghreb (two from Algeria and one from Tunisia). On the other side we will conduct a study and experiment on Palestinian and Syrian dialects. To do that, we build a parallel corpus, study the relationship between dialects and MSA, distinguish one dialect from another and present few experiments of machine translation between MSA and the different dialects. This paper is structured as follows: in section 2 we give an overview of the considered dialects. We discuss in section 3 how resources are built with some related works, then we detailed how we created our parallel corpus. Section 4 is dedicated to an analytical comparison of all dialects and MSA. Section 5 presents dialect identification experiments whereas the last one gives results of machine translation between all dialects and MSA. Finally, we conclude in section 7 by summarizing all the work.

2 Overview of the Used Dialects

Arabs use in their daily conversations dialects which could be considered such as variants of MSA. Tunisian and Algerian dialects share many features with the other Maghrebi dialects because of their similar history. It is worth mentioning that they contain many words borrowed from other languages, mainly Berber, French, Turkish, Italian and Spanish. Syrian and Palestinian dialects share an important number of features since they are included in the Levantine Arabic dialect continuum. In the following, we give a short overview about each dialect we study in this article.

2.1 Algerian Dialect

Algerian dialect is an informal spoken language, not used in official speech. Its vocabulary is roughly similar through all Algeria. However, in the east of the country, the dialect is closer to Tunisian whereas in the west it is closer to Moroccan. Most of the words of Arabic dialect come from MSA [10], but there is significant variation in the vocalization in most cases, and omission of some letters in other cases. Contrary to MSA, few letters are not used in Algerian as ظ and ذ, where most of the time they are respectively pronounced as ض and د.

Moreover Algerian dialect uses some non-Arabic letters like ڤ and پ.

2.2 Tunisian Dialect

Like other Maghrebi dialects, the vocabulary of the Tunisian dialect is mostly Arabic, with significant Berber substrates. However, its morphology, syntax, phonology and vocabulary differ from standard Arabic. The Tunisian dialect is

very agglutinative: people tend to use very few words for conversation where one word may express a whole sentence. It differs from MSA especially in its negation form where the markers are always agglutinated to other words as affixes or suffixes. Moreover, in Tunisian dialect, several Arabic words are used with substantial changes in their stem formation.

2.3 Syrian and Palestinian Dialects

Syrian and Palestinian dialects are part of Levantine Spoken Arabic which covers also dialects spoken in Lebanon and Jordan. Levantine Arabic shares most phonological, structural, and lexical features with other varieties of Arabic. At the same time, there are differences among Levantine dialects based on geography and urban/rural division. Arabic Syrian dialect is influenced by the Syriac language, a Semitic language of the Middle East, belonging to the Aramaean language group. It contains a large proportion of Arabic words and also words borrowed from Turkish and French. Palestinian dialect has slight phonetic differences from north Levantine dialects. It can be classified into two main categories: urban and countryside. It can be classified also according to geographical area (north and south). Palestinian dialect built in this work is mainly the dialect of people who live in Gaza strip.

3 Building a Parallel Corpus

It is well known that parallel corpora are the foundation stone of several natural language processing tasks, particularly cross-language applications such as machine translation, bilingual lexicon extraction and multilingual information retrieval. Building this kind of resources is a challenging task especially when it deals with under-resourced languages [11]. Arabic is one of these languages for which parallel corpora are scarce. The problem is much deeper with the Arabic dialects which are used by a huge number of people but, unfortunately they are often not written. To overcome the need of text corpora covering these languages, researchers can choose one of two main possible solutions: building the corpus from scratch or crawl the web to build a parallel set of sentences.

The solution adopted for our work is the first one: from scratch, since the overall goal of this work is Speech-to-Speech translation we need real everyday conversations. In the following, we focus on Annaba's dialect (ANB), the language spoken in the east of Algeria, on Algiers's dialect (ALG), the language used in the capital of Algeria, on Sfax's dialect (TUN) spoken in the south of Tunisia, Syrian (SYR) and Palestinian (PAL) dialects.

ANB corpus was created by recording different conversations from every day life whereas, for ALG, we used the recordings corresponding to movies and shows which are often expressed in the dialect of Algiers. Then we transcribed both of them by hand. To increase the size of the two corpora, we translated each of them into the other. Afterwards, these two corpora have been translated into MSA.

In order to introduce both Tunisian, Syrian and Palestinian dialects, we used MSA as a pivot language. We translated the MSA corpus to TUN, SYR and PAL. The Tunisian corpus was produced by 20 native speakers. Each one was responsible of translating almost 320 sentences from MSA to TUN. Speakers have very slight differences in their spoken languages. All of them are from the south of Tunisia where people tend to use Arabic words rather than French words as it is the case in the north of the country. In fact, the dialect used in the south is closer to the Standard Arabic than that used in the north of Tunisia. Syrian and Palestinian corpora were created in the same way as Tunisian one except that each of them has been obtained by two translators. Finally, we get a parallel corpus including ANB, ALG, TUN, SYR, PAL and MSA.

It should be noted that each dialect word is written by adopting the Arabic notation, that means if a dialectal word does exist in MSA, it is written in a standard form without any change, otherwise the word is written as it is uttered. We give in Table 1 an example of a same sentence from the built corpus. We can remark even if we do not read Arabic that some words are the same from one dialect to another, while others are completely different.

Table 1. An example of a sentence from the parallel corpus

Dialect/Language	Sentence
ALG	الوقت ڤاع لي درتو باش نجي ليك حبيتي تخسريه في دقيقة
ANB	الوقت أدا كل لي عڤبتو باه نجي عندك حايبة تفزديه في دقيقة
TUN	الوقت الكل الي قعدتو باش نجيك اتحب اتضيعو في دقيقة
SYR	كل الوقت اللّي قضيتو و انا جاي عندك بدك تروحيه بدقيقة
PAL	كل الوقت اللّي اخدتو عشان آجي عندك بدك تضيعيه في دقيقة
MSA	كل الوقت الذي استغرقته لكي آتي عندك تريدين أن تهدريه في دقيقة
Meaning	All the time that I put to visit you, you want to spoil it in one minute

4 Analytical Comparison

In the following, we will compare dialects between them and with MSA. The idea is to understand what is close to what? What is different? etc. We hope that this will help us in a future work to take advantage of MSA in order to develop linguistic tools for processing dialects.

4.1 Multi-dialect Corpus Statistics

The obtained parallel corpus is made up of 6400 parallel sentences. The MSA part contains 40906 words including 9131 different words. The five dialects, ALG,

ANB, TUN, SYR and PAL include an average of 37500 words with a vocabulary which does not exceed 10250 words (see Table 2). The average number of words in a dialect sentence is of 6 while it is of 7 for MSA. The shortest sentence in the corpus is composed of 4 words and the longest one contains 25 words.

Table 2. Parallel corpus description

Corpus	#Distinct words	#Words
ALG	8966	38707
ANB	9060	38428
TUN	10215	36648
SYR	9825	37259
PAL	9196	39286
MSA	9131	40906

4.2 Common Lexical Units between MSA and Dialects

MSA language is the same throughout the Arab world, while the dialects vary according to the geographical location. In this Section, we are interested in measuring how much the dialectal vocabulary is close to MSA by using the aforementioned parallel corpus. The experiments we achieved, show that the dialects employ many MSA words, even if the utterance of these words depends strongly on each dialect. Particularly, PAL is closest to MSA than other dialects are (Table 3).

Table 3. Percentage of common words between Arabic dialects and MSA

Dialect	ALG	ANB	TUN	SYR	PAL
%	21.18	21.07	37.60	37.36	51.68

These results are not surprising. Indeed, Middle-East Arabic dialects tend to be closer to MSA than those of the Maghreb. Also, it would be noticed that Arabic dialects spoken in south of Maghreb countries include more Arabic words than those spoken in the north. This explains the different rates in terms of common words between the two Algerian dialects on one side and the Tunisian dialect on another side. Indeed TUN is spoken in the south of Tunisia while ALG and ANB are dialects of northern Algeria. In Table 4, we give few examples of the most frequent words between the studied Arabic dialects and MSA.

In the same way, we computed the percentage of common words between all pairs of dialects. The Table 5 represents the percentage of common words between different dialects. These values show that ALG and ANB share the largest

Table 4. The most frequent words of each dialect relatively to our corpus

Dialect	Most Frequent words			
ALG	واحد *one*	صح *right*	راح *he went*	كامل *full*
ANB	واحد *one*	صح *right*	راح *he went*	عندك *you have*
TUN	كان *it was*	وقت *time*	واحد *one*	الكل *all*
SYR	اليوم *today*	مرة *one time*	واحد *one*	عندي *i have*
PAL	اليوم *today*	واحد *one*	طيب *good*	راح *he went*

number of words, followed by PAL and SYR. These results were excepted because ALG dialects and ANB are close since they are used in two cities separated by 372 miles, as PAL and SYR which are used in the same geographic location separated by only 175 miles. Also, TUN shares more words with PAL than the two other Maghrebi dialects do. Only 23% in average of words are common to Syrian and Maghrebi dialects. This result reinforces the fact that we made at the beginning of the article about the difficulty of conversing between Arabic people, from Maghreb and middle-east.

Table 5. Cross dialect percentage of common words.

	Percentage of common words				
Ref.	ALG	ANB	TUN	SYR	PAL
ALG	-	73.62	35.43	24.16	25.43
ANB	72.86	-	34.25	23.59	25.00
TUN	31.10	30.38	-	29.79	33.49
SYR	21.01	20.73	29.52	-	44.00
PAL	24.79	24.63	37.20	49.33	-

We estimated also the percentage of common words at sentence level between each pair of languages. For each pair of the k^{th} aligned sentences $S^k_{L_i}$ and $S^k_{L_j}$ from the bitext (L_i, L_j). The common words is calculated as in formula 1 , it corresponds to the percentage of common words in the two sentences to the total number of words in both sentences. Then we estimate the average common number of words over all the sentences.

$$Ovp(S^k_{L_i}, S^k_{L_j}) = \frac{|S^k_{L_i} \cap S^k_{L_j}|}{|S^k_{L_i} \cup S^k_{L_j}|} \tag{1}$$

Table 6 presents the overlap between the Arabic dialects and MSA at sentence level. The achieved results confirm those of the two last experiments. PAL is the closest dialect to MSA followed by TUN then SYR, while ALG and ANB are the farthest. This experiment also highlights the closeness between Algerian dialects (ALG and ANB) and Levantine dialects (PAL and SYR). It shows also That TUN is closer to PAL and to SYR than ALG and ANB.

Table 6. Overlapping of vocabularies between Dialects and MSA

	ALG	ANB	TUN	SYR	PAL
MSA	0.12	0.10	0.16	0.14	0.21
PAL	0.13	0.11	0.17	0.21	
SYR	0.09	0.09	0.13		
TUN	0.16	0.13			
ANB	0.32				

4.3 Measuring the Cross-language Divergence

In this section, we are interested by measuring the divergence between dialects and MSA throw unigram language models. For this purpose we choose to use the Hellinger Distance (HD) [12][13], a measure of distributional divergence. It quantifies the similarity between two probability distributions. It has been used to detect failures in classification performance [14] and in machine learning it is used to estimate the class distribution [15]. In [16], this distance was used to measure information loss in data protection. Hellinger distance is symmetric and non-negative, and obeys to triangle rule.

In order to measure the divergence between two languages with HD, let consider a bi-text(L_i, L_j) with the vocabularies V_i and V_j respectively. We constitute V, a vocabulary including 10K words including the common words between V_i and V_j and from the remaining words of the two vocabularies we include the most frequent ones to complete V. For each side of the bi-text, a unigram probability distribution $P(w|L_i)$ is computed over V. To avoid zero probabilities due to the words not belonging to the considered language, we decided to smooth the probabilities. The comparison of the two distributions is then calculated as follows:

$$HD(L_i, L_j) = \sqrt{\frac{1}{2} \sum_{w \in V} (\sqrt{P(w|L_i)} - \sqrt{P(w|L_j)})^2} \quad (2)$$

Table 7 draws HD values computed between all dialects and MSA. These values show that PAL is the closest dialect to MSA followed by TUN then SYR, whereas ALG and ANB are the most divergent. The closest dialects according to HD are ALG and ANB and also PAL and SYR. The closest dialect to MSA

is PAL and the farthest are ALG and ANB. Another interesting and expected result is the one related to the distance between TUN and the other dialects, TUN is closer to ALG and ANB than to PAL and SYR.

Table 7. Hellinger distance values for the different pairs of languages

	ALG	ANB	TUN	SYR	PAL
MSA	0.72	0.72	0.60	0.62	0.55
PAL	0.85	0.86	0.81	0.76	
SYR	0.84	0.86	0.81		
TUN	0.79	0.80			
ANB	0.73				

5 Dialect Identification

In this section, we deal with the issue of using several languages in the same sentence. This is very common in Arabic world and especially in Maghreb. This phenomenon is commonly named code switching. Arabic people often switch between several languages. For instance, in Algeria, people could switch from dialect, to MSA to French. In the following, French will not be taken into account. To identify the different languages in order to apply the appropriate tools, we consider the identification of language such as a classification issue which will be treated in the following by a Naive Bayes classifier (NBC). NBC is probabilistic learning algorithm, it is used for many issues in NLP [17–19]. A naive Bayes classifier assumes that all features representing a given problem are conditionally independent given the value of classification variables.

For our purpose, NBC is based on 3-grams features. Given n classes corresponding to n languages, the purpose is to assign the most suitable class C_i in accordance to a set of features $F = \{f_1, ..., f_n\}$ which maximizes the conditional probability:

$$p(C_i \mid f_1, ..., f_n) = p(C_i) \prod_{j=1}^{n} p(f_j|C_i) \tag{3}$$

where $p(C_i)$ is the probability of the class C_i and $p(f_j|C_i)$ is the conditional probability of the feature f_j observed with the class C_i.

For the experiment, we created a corpus by merging MSA, ALG, ANB, TUN, SYR and PAL for which each sentence is annotated by its corresponding language class. By selecting randomly 80% of the data, we create the training corpus and the remaining has been dedicated for test. Classification results in Table 8 show that the recall for MSA is the highest one (75%); this could be explained by the fact that MSA writing obeys to strict rules contrary to dialects for which no

Table 8. Dialect identification results using the parallel corpus

Language	Precision	Recall	F
ALG	0.48	0.50	0.49
ANB	0.49	0.49	0.49
TUN	0.68	0.52	0.59
SYR	0.62	0.55	0.58
PAL	0.53	0.57	0.55
MSA	0.64	0.75	0.69

formal writing rules exist: a dialect word could be written in different forms which are all acceptable. Consequently, this phenomenon generates a larger distribution probability for dialects than MSA ones.

Table 9 draws the confusion matrix of the classifier. For dialect side, it is clearly shown that the highest confusion rates are those between ALG and ANB and between PAL and SYR, this confusion is justified by the closeness between these pairs of dialects; ALG and ANB for example share an important vocabulary in spite of their difference. For MSA side, it is shown that the highest confusion

Table 9. Confusion matrix rates for dialect identification using the parallel corpus.

	Estimated language classes					
True language classes	ALG	ANB	TUN	SYR	PAL	MSA
ALG	**50**	35	4	6	2	4
ANB	38	**49**	5	2	3	3
TUN	12	8	**52**	6	9	14
SYR	3	3	4	**55**	24	11
PAL	4	3	4	16	**57**	17
MSA	2	2	4	5	12	**75**

rates related to MSA are those with PAL, whereas for ALG and ANB dialects, confusion rates related to MSA do not exceed 4% for both dialects.

6 Machine Translation

Arabic language translation has been widely studied. The rich morphology of Arabic is seen as a rocky barrier in building efficient translation systems. Indeed, Arabic is characterized by complex a morphology and a rich vocabulary. It is a derivational, flexional and agglutinative language. We recall that, in order to compare it to English, an Arabic word (or more rigorously a lexical entry) can sometimes correspond to a whole English sentence [20].

Moreover, Arabic words are often ambiguous because a single word could have multiple morphological analyses. This is due to the richness of the Arabic

affixation and the omission of short vowels. In addition, articles, prepositions, pronouns, etc. can be affixed to adjectives, nouns, verbs and particles to which they are related. All these phenomena increase the ambiguity and make the traditional issues of NLP more challenging such as machine translation from and to Arabic.

As shown in the previous experiments, dialects even if they are inspired strongly from Arabic, the significant differences may prevent communication between people of Arabic world. That is why, it is very important to propose machine translation between different dialects and MSA. In the following, we present several experiments in order to develop machine translation between Arabic dialects and MSA. For each pair of languages we used a parallel corpus of 6400 sentences (5900 have been dedicated to training and the remaining for tests).

Table 10. BLEU score of Machine Translation on different pairs of languages using two smoothing techniques

	Target											
	ALG		ANB		TUN		SYR		PAL		MSA	
Source	KN	WB	KN	WB	KN	WB	KN	WB	KN	WB	KN	WB
ALG	-	-	61.06	**60.81**	**9.67**	9.36	7.29	**7.95**	**10.61**	10.14	**15.1**	14.64
ANB	**67.31**	65.55	-	-	**9.08**	8.64	7.52	**7.95**	**10.12**	9.84	**14.44**	13.95
TUN	**9.89**	9.48	**9.34**	9.01	-	-	**13.05**	12.93	**22.55**	22.21	**25.99**	25.21
SYR	**7.57**	7.50	7.50	**7.64**	**13.67**	13.23	-	-	**26.60**	25.74	**24.14**	22.96
PAL	**11.28**	10.67	**9.53**	9.15	**17.93**	16.64	**23.29**	23.07	-	-	**40.48**	39.76
MSA	**13.55**	13.05	**12.54**	11.72	20.03	**20.44**	**21.38**	20.32	**42.46**	41.37	-	-

All the MT systems we used are phrase-based [21] with default settings: bidirectional phrase and lexical translation probabilities, distortion model, a word and a phrase penalty and a trigram language model. We used GIZA++ [22] for alignment and SRILM toolkit [23] to compute trigram language models. Since the parallel corpus is small, we experimented the Kneser-Ney and Witten-Bell smoothing techniques hoping to identify the one which best fits. The results conducted on the test set are presented in terms of BLEU in Table 10. This experiment leads to very interesting conclusions. First of all, for small parallel corpus, it seems that the smoothing technique has an impact on translation results. A difference of almost 2 points in terms of BLEU scores has been observed for translating from ANB to ALG. But, we can not generalize by affirming that one smoothing technique is definitely better than another. High score of translation has been achieved between ANB and ALG in both sides. This result is natural since these two dialects are used in the same country and share up to 60% of words. Almost the same observation is made for the pair SYR and PAL since these two dialects belong to the same language family (Levantine).

Another interesting and expected result is BLEU score between MSA and dialects. In fact, the highest one is related to PAL in both sides since this dialect is the closest one to MSA as shown in other experiments of sections 4.2 and 4.3. Most surprising results are those relative to SYR and TUN. It seems that it is easier to translate TUN to MSA than SYR to MSA. Also, translating from MSA to TUN gives better results than from MSA to the Algerian dialects. In the symmetric side of translation we get the same scale of results. This definitely shows the closeness of TUN dialect to MSA in comparison to the Algerian dialects.

7 Conclusion

In this paper we present an analytical study of Arabic dialects from Middle-east and Maghreb. Maghreb's dialects use several words from French, Turkish, ... and adapted them phonetically so they become full words of these dialects. In the opposite, Middle-east dialects are more close to MSA because they share an important vocabulary with it.

To make this research and study possible, we started from scratch because for these vernacular languages, there is no available written resources. We build a parallel corpus including 5 dialects (two from Algeria, one from Tunisia, and the two others from Middle-east: Palestine and Syria) and MSA. We perform different experimentations in order to study the relationship between MSA and dialects on one hand and cross-dialects on the other hand. For this, we calculated the overlapping between each pair of vocabularies. We then calculated the distance between the distributions of each pair of languages in order to measure which language is closer to which one. The carried out results are consistent with the fact that Middle-East Arabic dialects are closer to MSA than those of the Maghreb. This has been confirmed by the other experiments of identification handled by machine learning techniques. We showed that it is easier to identify MSA than dialects because it is a natural language with the whole standard linguistic constraints. Concerning the experience on identification, the results could be separated into two classes. The first one concerns ALG and ANB and the other one the three other dialects. In fact for this last class, the F-measure results are close and the difference between them are not statistically significant. This means that it is easier to identify PAL, SYR and TUN than Algerian dialects.

We conducted also several experiments of machine translation between all the pairs of languages. We took advantage from this experiment to try to understand whether the smoothing techniques could have an impact on BLEU score when we are faced to small corpora. We remarked that in some cases, the used method could improve BLEU significantly. High score of translation has been achieved between ANB and ALG in both sides. This result is natural since these two dialects are used in the same country and share up to 70% of words. In the near future, we will extend this work to other dialects and will propose a speech to speech system which is the main objective of this work.

References

1. Kilany, H., Gadalla, H., Arram, H., Yacoub, A., El-Habashi, A., McLemore, C.: Egyptian Colloquial Arabic Lexicon. In: LDC catalog number LDC99L22. (2002)
2. Kirchhoff, K., Bilmes, J., Das, S., Duta, N., Egan, M., Ji, G., He, F., Hopkins, J., Liu, D., Noamany, M., Schone, P., Schwartz, R., Vergyri, D.: Novel Approaches to Arabic Speech Recognition: Report from the 2002 Johns-Hopkins Summer Workshop. In: Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing. (2003) 344–347
3. Habash, N., Rambow, O.: Magead: A Morphological Analyzer and Generator for the Arabic Dialects. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics. (2006) 681–688
4. Chiang, D., Diab, M., Habash, N., Rambow, O., Shareef, S.: Parsing Arabic Dialects. In: Proceedings of the European Chapter of ACL (EACL). (2006)
5. Zbib, R., Malchiodi, E., Jacob, D., Stallard, D., Matsoukas, S., Schwartz, R., Makhoul, J., Zaidan, O., Callison-Burch, C.: 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. NAACL HLT 12 (2012) 49–59
6. Salloum, W., Habash, N.: Dialectal Arabic to English Machine Translation: Pivoting through Modern Standard Arabic. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. NAACL HLT 13 (2013) 348–358
7. Zaidan, O., Callison-Burch, C.: Arabic Dialect Identification. Computational Linguistics, Volume 40 (2014) 171–202
8. Elfardy, H., Diab, M.: Sentence Level Dialect Identification in Arabic. In: ACL (2). (2013) 456–461
9. Bouamor, H., Habash, N., Oflazer, K.: A Multidialectal Parallel Corpus of Arabic. In: Proceedings of the Language Resources and Evaluation Conference, LREC-2014. (2014) 1240–1245
10. Meftouh, K., Bouchemal, N., Smaili, K.: A Study of a Non-resourced Language: an Algerian Dialect. In: Third International Workshop on Spoken Languages Technologies for Under-resourced Languages. (2012) 125–132
11. Skadiņa, I., Aker, A., Giouli, V., Tufis, D., Gaizauskas, R., Mieriņa, M., Mastropavlos, N.: A Collection of Comparable Corpora for Under-resourced Languages. In: Proceedings of the 2010 Conference on Human Language Technologies – The Baltic Perspective: Proceedings of the Fourth International Conference Baltic HLT 2010. (2010) 161–168
12. Kailath, T.: The Divergence and Bhattacharyya Distance Measures in Signal Selection. In: Communication Technology, IEEE Transactions. Volume 15. (1967) 52–60
13. Rao, C.R.: A Review of Canonical Coordinates and an Alternative to Correspondence Analysis Using Hellinger Distance. Quaderns dEstadistica i Investig Ope, Questiio, Volume 19 (1995) 23–63
14. Cieslak, D.A., Chawla, N.V.: A Framework for Monitoring Classifiers Performance: When and Why Failure Occurs? Knowledge and Information Systems (2009) 83–109
15. González-Castro, V., Alaiz-Rodrguez, R., Alegre, E.: Class Distribution Estimation Based on the Hellinger Distance. Information Sciences (2013) 146–164

16. Torra, V., Carlson, M.: On the Hellinger Distance for Measuring Information Loss in Microdata. In: Joint UNECE/Eurostat work session on statistical data confidentiality. (2013)
17. Pop, I.: An Approach of the Naive Bayes Classifier for the Dcument Classification. General Mathematics, Volume 14, No 4 (2006) 135–138
18. Pedersen, T.: A Simple Approach to Building Ensembles of Naive Bayesian Classifiers for Word Sense Disambiguation. In: Proceedings of 1st Annual Meeting of the North American Chapter of the Association for Computational Linguistics. (2000) 63–69
19. Ahmed, F., Nurnberger, A.: Arabic/English Word Translation Disambiguation Using Parallel Corpora and Matching Schemes. In: 12th EAMT conference. (2008) 6–11
20. Badr, I., Zbib, R., Glass, J.: Segmentation for English-to-Arabic Statistical Machine Translation. In: Proceedings of the ACL 2008 Conference Short Papers. (2008) 153–156
21. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: Open Source Toolkit for Statistical Machine Translation. Proceedings of the Annual Meeting of the Association for Computational Linguistics, demonstation session (2007) 177–180
22. Och, F.J., Ney, H.: A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics, Volume 29, No 1 (2003) 19–51
23. Stolcke, A.: Srilm – an Extensible Language Modeling Toolkit. In: ICSLP, Denver, USA (2002) 901–904