PADIC: extension and new experiments

K. Meftouh, S Harrat, Kamel Smaïli

PADIC: extension and new experiments

K. Meftouh[1], S. Harrat[2], K. Smaïli[3]
[1]Badji Mokhtar University, BP 12 23000 Annaba, Algeria
karima.meftouh@univ-annaba.org
[2]ENSB* , ESI** , Algiers, Algeria
slmhrrt@gmail.com
[3]Campus Scientifique, LORIA, Nancy, France
smaili@loria.fr

***Abstract*— PADIC is a multidialectal parallel Arabic corpus. It was composed initially by five Arabic dialects, three from the Maghreb and two from the Middle East, in addition to standard Arabic. In this paper, we present an augmented version of PADIC with a Moroccan dialect. We give also an evaluation, using the σ–index, of the computerization level of the Arabic dialects present in PADIC which reveals that these languages are really under-resourced. Several experiments in machine translation, in both sides between all the combinations of language pairs, are discussed too. For each language, we interpolated the corresponding Language Model (LM) with a large Arabic corpus based LM. The results show that this interpolation is in some cases without effect on the performances of translation systems and in others is rather penalizing.**

***Keywords*Machine translation system, Standard Arabic, Arabic dialect, PADIC, Under-resourced language, Interpolated language model**

I. Introduction

Spoken Arabic is often referred to as colloquial Arabic, dialects, or vernaculars. It's a mixed form, which has many variations, and often a dominating influence from local languages (from before the introduction of Arabic). Differences between variants of spoken Arabic can be large enough to make them incomprehensible to each other. Hence, regarding the large differences between such spoken languages, we can consider them as disparate languages or more exactly as different dialects depending on the geographical place in which they are practiced.

The varieties of Arabic dialects are distributed over the 22 countries in the Arab World and are classified by Natural Language Processing (NLP) researchers community into five groups namely: Maghrebi (spoken in all of North Africa), Egyptian (spoken in Egypt, but understood universally), Levantine (spoken primarily in the Levant, Syria and Palestine), Iraqi (spoken in Iraq) and Gulf (spoken primarily in Saudi Arabia, UAE, Kuwait and Qatar) [1].

Arabic dialect (AD) is the mother tongue for all native speakers of Arabic. Furthermore, there is no native speakers of Modern Standard Arabic (MSA). However, MSA is the unique written standard using the Arabic script. AD is not a written form of the language but there is a tendency nowadays to use the colloquial spoken dialects in written forms as well, especially in social media channels and forums. People tend to use it as it is easier and it reflects their daily life style. Some AD differ significantly from MSA on all levels of linguistic representation that results in huge inconsistencies in orthography [2].

NLP for Arabic dialects has grown widely these last years. Indeed, several works were proposed dealing with all aspects of Natural Language Processing. However, some AD varieties, notably Egyptian Arabic, have received more attention and have a growing collection of resources [3]. Others varieties, such as Maghrebi, still lag behind in that respect. In fact, most of Arabic dialects (if not all) could be considered as under-resourced languages. They are impoverished in terms of available tools and resources compared to MSA for which most of the research effort, in creating tools and resources for Arabic, has focused on.

The research work we carry is focused on the processing of Arabic dialects; and more precisely on Statistical Machine Translation (SMT) of these dialects to standard Arabic and vice versa. PADIC (Parallel Arabic DIalect Corpus)[4] is one of the important contributions we made in this field. It is a multi-dialectal parallel corpus. Initially, it was composed of 5 Arabic dialects namely: ALG (the dialect of Algiers capital of Algeria), ANB (the one of Annaba in eastern Algeria), TUN (a dialect of Sfax in western Tunisia), SYR (spoken in Damascus capital of Syria) and PAL (spoken in Gaza in Palestine) .

In this paper, we present an augmented version of PADIC with a Moroccan dialect and discuss Machine Translation (MT) results obtained for all the pairs of dialects contained into PADIC. In addition, we will proceed to an evaluation of these dialects in terms of resources using the σ–index.

II. Related work

Number of researchers have exploited the existing tools of standard Arabic NLP to develop their Machine Translation systems of Arabic dialects. For example, Tachicart et al. in [5] used tools designed for standard Arabic and adapted them to Moroccan dialect in order to built a translation system of MSA to Moroccan dialect by combining a rule-based approach and a statistical approach. Sawaf [6] built a hybrid AD-English MT system that uses an MSA pivoting approach. In this approach, AD is normalized into

*Ecole Normale Supérieure Bouzareah.
**Ecole Supérieure d'Informatique.

MSA using character-based AD normalization rules, an AD morphological analyzer, an AD normalization decoder that relies on language models, and a lexicon. Similarly, Salloum and Habash [7] presented Elissa , a MT system from AD to MSA which employed a rule-based approach that relies on morphological analysis, morphological transfer rules and dictionaries in addition to language models to produce MSA paraphrases of dialectal sentences. Elissa handles Levantine, Egyptian, Iraqi, and to a lesser degree Gulf Arabic. Zbib et al. [8] used crowdsourcing to build Levantine-English and Egyptian-English parallel corpora. They selected dialectal sentences from a large corpus of Arabic web text, and translated using Amazons Mechanical Turk. They used this data to build Dialectal Arabic MT systems, and find that small amounts of dialectal data have a dramatic impact on translation quality.

Several works have aimed multidialectal Arabic corpus construction. For example, Almeman and Lee in [9] built automatic Arabic dialects corpora by exploiting the web as a corpus. A survey has been conducted to categorise distinct words and phrases that are common to a specific dialect only, and not used in other dialects in order to download a specific dialect text corpus. They obtained 48M tokens from different Arabic dialects. These dialects were categorised into four main dialects Gulf, Levantine, Egyptian and North African, resulting in 14.5M, 10.4M, 13M and 10.1M tokens being obtained respectively. In [1], Cotterell and Callison-Burch present a multi-dialect, multi-genre, human annotated corpus of dialectal Arabic with data obtained from both online newspaper commentary and Twitter. This corpus cover five dialects of Arabic: Egyptian, Gulf, Levantine, Maghrebi and Iraqi. The authors also provide results for the Arabic dialect identification task. Another multi-dialect corpus based on the geographical information of tweets was presented in [10]. They mapped information of user locations to one of the Arab countries, and extracted tweets that have dialectal word(s).

In the other hand, multidialectal Arabic parallel corpora does not exist. The first corpus of such kind is presented in [11]. it's a collection of 2000 sentences in Standard Arabic, Egyptian, Tunisian, Jordanian, Palestinian and Syrian Arabic, in addition to English. The sentences were selected from the Egyptian part of the Egyptian-English corpus built by Zbib et al.[8]. The second one is PADIC. The approach we used to built it is almost similar to that of Bouamor et al. [11] except that in our case, we started from scratch. We have created everything. We present the details in section IV.

III. AD UNDER-RESOURCED LANGUAGES

There are more than 7000 languages in the world but only a small portion of these languages has resources for automatic processing and are supported by text processing software, information search engines, automatic translators, processing tools and speech synthesis, etc.

Measuring the availability of these resources or services for a given language allows to define its computerization level. Berment, in his thesis [12], defines such a metric called "σ index" which can be calculated as follows: a list of services is evaluated for a given language by an expert and a mean score is calculated (marks for each service are weighted by the criticity or importance of the service). An under-resourced language (or a language-π) is defined as a language which has a score below $10/20$.

For the evaluation of Arabic dialects (present initially in PADIC in addition to Moroccan dialect), we adapted the list of services and resources considered by Berment as follows:

Text processing Text selection and lexicographical sort services have been removed due to the fact that they are applied to non-segmented languages. So, they are not applicable to Arabic dialects. We also added a new service which is supported for non-Arabic characters used in Arabic dialects. These are the characters corresponding to the phonemes /P/, /V/ and /G/ which are present generally in borrowed words (particularly from French).

Machine translation We divided it in text translation and speech translation for a more precise assessment.

Resources We introduced monolingual and multilingual corpora as evaluation criterion given their importance in the implementation of NLP tools.

The evaluation table is completed according to our knowledge of existing software and resources for each dialect. The criticity related to speech processing, machine translation and the resources are the highest because of the importance of these services in recent NLP applications. Lower values were assigned to text processing service, as the Arabic dialects are supported by text processing tools for standard Arabic. Table I summarize the evaluation process of Arabic dialects and give their respective σ values.

From Table I, we can note that all the evaluated dialects are under-resourced languages. We note also that the other Arabic dialects can be classified in the same category as the lack of resources is observed for most of them except Egyptian one for which a number of resources exist, but these resources can not classify it in other classes.

IV. EXTENDING PADIC

In this paper we present a new version of PADIC where the list of dialects is augmented by Moroccan dialect (MOR)[1]. The new part is obtained in the same way as for the other dialects using MSA as a pivot language. We summarize the creation process of PADIC, as presented in [4], in the following steps:

Step1 *Collection and transcription*

We were first interested in the two Algerian Arabic dialects, the dialect of Annaba (ANB) and the one spoken in Algiers (ALG). We created ANB corpus by recording different conversations from every day life (in medical offices, cafes, markets, ...), whereas, for ALG,

[1]This version of PADIC is available at `https://www.researchgate.net/publication/316463706\_PADIC\_A\_Parallel\_Arabic\_DIalect\_Corpus`

Dialects			ALG		ANB		TUN		MOR		SYR		PAL	
Services / ressources		Criticity C_k	Mark N_k	weighted mark	Mark N_k	weighted mark	Mark N_k	weighted mark	Mark N_k	weighted mark	Mark N_k	weighted mark	Mark N_k	weighted mark
Text processing	Basic input and support of non-Arabic characters	8	5	40	5	40	5	40	5	40	5	40	5	40
	Visualization / printing	6	8	48	8	48	8	48	8	48	8	48	8	48
	search and replace	6	8	48	8	48	8	48	8	48	8	48	8	48
	spelling Correction	6	0	0	0	0	0	0	0	0	3	18	3	18
	grammatical correctness	6	0	0	0	0	0	0	0	0	2	12	2	12
	stylistic Correction	5	0	0	0	0	0	0	0	0	0	0	0	0
speech processing	Vocal synthesis	8	0	0	0	0	0	0	0	0	0	0	0	0
	Speech recognition	8	0	0	0	0	0	0	0	0	0	0	0	0
Machine translation	Text translation	8	0	0	0	0	2	16	0	0	2	16	2	16
	Speech translation	8	0	0	0	0	0	0	0	0	0	0	0	0
OCR	Optical Character Recognition	5	5	25	5	25	5	25	5	25	5	25	5	25
Resources	Bilingual dictionary	8	0	0	0	0	0	0	0	0	0	0	0	0
	Usability dictionary	8	0	0	0	0	0	0	0	0	0	0	0	0
	monolingual corpus	10	0	0	0	0	3	30	0	0	3	30	0	0
	multilingual corpus	8	0	0	0	0	2	16	0	0	3	24	0	0
Mean (/20)			1,49		1,49		2,06		1,49		2,42		1,92	

TABLE I

COMPUTERIZATION LEVEL FOR THE ARABIC DIALECTS PRESENT IN PADIC.

we used the recordings corresponding to movies and TV shows which are often expressed in the dialect of Algiers. Then we transcribed both of them by hand.

Step2 *Translation by hand*

In order to increase the size of ANB and ALG corpora, we translated by hand each of them into the other. So we got a parallel corpus ANB-ALG made of 6400 sentences. We subsequently proceeded to its translation (also by hand) to MSA.

Step3 *Extension to other dialects*

MSA was used as a pivot language to get other dialectal corpora. To do that, we translated the MSA part of the parallel corpus, obtained in the previous step, for a first time in TUN, SYR, and PAL; and for a second time in MOR. The translation in each dialect was handled by one or more native speakers. We give in Table II, for each language, the original town and the number of native speakers who participated in the creation of the concerned part of the corpus. Also, we report in Table III some statistics on the various parts of PADIC [2].

As we mention before, there is no standard of writing for Arabic dialects. So, when transcribing the ANB and ALG dialects (first step), we used a set of conventions that are detailed in the following section. We have taken care to respect these conventions also for writing the other dialects (TUN, SYR, PAL and MOR).

[2]The reader can refer to [13] for a more detailed analytical study of PADIC

TABLE II

ORIGINAL TOWN AND NUMBER OF NATIVE SPEAKERS (N.S)WHO PARTICIPATED IN THE CREATION OF EACH PART OF PADIC.

AD	**ALG**	**ANB**	**TUN**	**MOR**	**SYR**	**PAL**
City	Algiers	Annaba	Sfax	Rabat	Damascus	Gaza
Nb of N.S	1	20	20	1	2	2

TABLE III

PADIC DESCRIPTION.

Corpus	**#Distinct words**	**#Words**
ALG	9151	40750
ANB	9530	41382
TUN	10040	38966
MOR	9703	42587
SYR	10322	40110
PAL	9642	42273
MSA	10314	44270

V. WRITING CONVENTIONS

For the transcription of speech to text as well as translation to dialects, we agreed to use the following spelling rules:

1) Transcribe each dialectal word by adopting the Arabic notation. In other terms, if a dialectal word exists in standard Arabic, we adopt the standard Arabic form, without any changes (قَال *qāl* "he said", اَلأَرض *alʿarḍ* "the earth" or "the floor" according to the context). If not, the word is written as it is pronounced like for

صوَارد *ṣwārd* "Money".

2) Adopt the definite article أَلـ *al* of standard Arabic for dialect words (اَلمسِيد *al-msīd* "the school") and for words of French or other origin also (اَلمَانطُ *al-mānṭo LE MANTEAU* "the coat").
3) Use ة *h* for feminine dialectal words (طفلَة *ṭflah* "a girl")
4) Write the suffixed pronouns as in standard Arabic, whether attached to a verb or a noun (كَتبُه *katbuh* "He writes it", دَارهُم *dārhum* "their house").
5) Write the dialectal prepositions "ف *f*", "ب *b*" and "ل *l*" as in MSA (فِي الدَار *fī 'ldār* "in the house, بالصح *b-alṣḥ* "really", لَلمسِيد *lalmsīd* "for school").
6) Write French or English words with Arabic letters as they are pronounced (كُنكسيُون *kunaksyūn*, كُنكشَن *kunakšan* "connection" respectively for French and English prononciation).
7) Use the letters پ *p* and ڤ *v* when necessary (پُوپِي *pūpī POUPÉE* "Doll", قَاف *vāf GAFFE* "gaffe").
8) For French or English words used in dialect, when we have the following form "article + name", write the article and the name as one word (لَجُوپ *laǧūp LA JUPE* "The skirt").

VI. Machine translation experiments

Arabic dialects, although they are mainly inspired from Arabic, significant differences may exist and make the communication between people of the Arab world uncomfortable. Indeed, we observe in our daily life the difficulty that someone can encounter when it comes to talking to an Arab person from the Middle East, for example. Due to the popularity of middle eastern, especially Egyptian movies and other media, we think that communication happen very smoothly. Many real experiences have shown that this communication is not always obvious and is not as easy as we thought. We often use the standard Arabic or even French or English to convey an idea that we are unable to communicate. In this context, we propose machine translation between Arabic dialects and standard Arabic.

In the following, we present several experiments in machine translation between all the combinations of dialect pairs present into PADIC. We conduct also experiments of machine translation between these dialects and MSA in both sides.

All the MT systems, we used, are phrase-based [14] with the following settings: bidirectional phrase and lexical translation probabilities, distortion model, a word and a phrase penalty and a trigram language model. We have not used a larger language model because PADIC is not suitable for large ngrams. We used GIZA++ [15] for alignment and SRILM toolkit [16] to compute trigram language models using Kneser-Ney smoothing technique. Many automatic measures have been proposed to facilitate the evaluation of MT systems, the most widely used of which is BLEU [17]. In this paper, we present in Table IV the results conducted on a test set of 500 sentences using BLEU.

TABLE IV
BLEU score of Machine Translation on different pairs of languages using Kneser-Ney smoothing technique.

	Target						
Source	**ALG**	**ANB**	**TUN**	**MOR**	**SYR**	**PAL**	**MSA**
ALG	-	61.06	9.67	10.22	7.29	10.61	15.1
ANB	67.31	-	9.08	10.00	7.52	10.12	14.44
TUN	9.89	9.34	-	14.37	13.05	22.55	25.99
MOR	10.13	10.16	14.68	-	9.68	18.91	24.93
SYR	7.57	7.50	13.67	9.93	-	26.60	24.14
PAL	11.28	9.53	17.93	16.08	23.29	-	40.48
MSA	13.55	12.54	20.03	20.02	21.38	42.46	-

A. Cross-translation results comparison

High score of translation has been achieved between ANB and ALG in both sides. This result is natural since these two dialects are spoken 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 (for both sides) showing that this dialect is the closest to MSA. Most surprising results are those relative to SYR, TUN and MOR. It seems that it is easier to translate TUN and MOR to MSA than SYR to MSA. Also, translating from MSA to TUN and MOR 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 and MOR to MSA in comparison to the Algerian dialects.

Even more surprising are the translation results of MOR and TUN to Algerian dialects. Despite the geographical proximity, it seems that it is more difficult to translate TUN to Algerian dialects than MOR. This remains true whatever the direction of translation. The results show also that MOR and TUN are more close compared to Algerian dialects. We think that this is due to the use of MSA as a pivot language. Indeed, the corpora of Algerian dialects are the only ones to have been constructed without resorting to translation. The sentences extracted from spontaneous discussions are therefore expressed in a proper dialectical language. All other corpora, as we have already mentioned, were obtained by translating the MSA side of PADIC. As a result of MSA-pivoting, many dialect words have been substituted by MSA words; which allowed large rates in terms of common words between each other and with MSA, compared to Algerian dialects (see Table V).

In terms of out of vocabulary (OOV) words, we encounter a significant OOV rate between test and training data for

TABLE V

PERCENTAGE OF COMMON WORDS INTER-ARABIC DIALECTS AND MSA.

	ALG	ANB	TUN	MOR	SYR	PAL
MSA	21.18	21.07	37.60	29.34	37.36	51.68
PAL	24.79	24.63	37.20	30.16	49.33	
SYR	21.01	20.73	28.22	26.91		
MOR	27.79	30.37	35.37			
TUN	31.10	30.38				
ANB	72.86					

all the used languages (see Table VI). This is due to the relatively small size of the training corpora.

TABLE VI

OUT OF VOCABULARY RATES

AD	ALG	ANB	TUN	MOR	SYR	PAL	MSA
%	17.65	20.16	20.08	16.19	24.42	18.16	17.03

B. Measuring Human-targeted translation edit rate

HTER, short for Human-targeted Translation Edit Rate, employs human annotation to make TER a more accurate measure of translation quality [18]. It is a metric that requires the creation of targeted references to accurately measuring the number of edits needed to transform a hypothesis into a fluent target language sentence with the same meaning as the references. This is done by human editing of the system hypothesis translation to produce the target reference that has the same meaning as the original references. It is worth noting that the post edition of hypothesis was possible only for MSA, ALG and ANB, the languages that we master. So, we report in Table VII HTER scores computed for SMT systems translating from dialects to MSA , whereas in Table VIII, we give HTER values for MT from MSA to ALG and ANB. We applied the procedure proposed in [18] to create one targeted reference for each system hypothesis translation.

TABLE VII

HTER SCORES (IN%) OF DIALECT-TO-MSA SMT SYSTEMS.

Source	ALG	ANB	TUN	MOR	SYR	PAL
HTER	62.7	49.17	31.71	26.23	28.36	18.98

TABLE VIII

HTER SCORES (IN %) OF MSA-TO-ALG AND MSA-TO-ANB SMT SYSTEMS.

MSA-ALG	MSA-ANB
41.44	29.90

In general, the results show that the HTER reduces considerably the edit rate relative to TER. We can notice that these results confirm the hypothesis that the MSA-pivoting makes the other corpora (not Algerian ones) closer to standard Arabic. That is why they have a much smaller HTER compared to ALG and ANB.

C. Interpolated language model

Arabic dialects are languages derived primarily from standard Arabic; we thought that the use of an interpolated language model can have a positive effect on the translation system performance. So, we trained two language models, one using the target language part of the training corpus and another one computed on the Linguistic Data Consortium (LDC) Arabic Treebank (Part3,V1.0) [19]. Language modeling software such as the SRILM toolkit we used [16] allows the interpolation of these language models. Before interpolating, we compute the optimal interpolation weights for the corresponding models, also using the SRILM toolkit. Table IX shows results of this experiment in terms of differences between the BLEU values of translation systems computed with and without interpolated language model for each pair of languages.

TABLE IX

BLEU SCORES VARIATIONS OF MACHINE TRANSLATION SYSTEMS USING AN INTERPOLATED LANGUAGE MODEL.

	Target						
Source	**ALG**	**ANB**	**TUN**	**MOR**	**SYR**	**PAL**	**MSA**
ALG	-	-0.004	0.004	-3.32	-0.3	-0.6	0.01
ANB	0.55	-	0.07	-3.55	-0.88	-0.68	0.01
TUN	-0.03	-0.02	-	-4.6	-0.21	-1.04	0.27
MOR	-3.14	-2.75	-3.74	-	-2.85	-5.81	-5.04
SYR	0.11	0.00	-0.01	-3.9	-	-2.33	0.1
PAL	0.02	0.01	0.14	-5.62	-2.12	-	-0.38
MSA	0.06	-0.02	-0.15	-6.19	-1.27	-2.99	-

In general, the use of an interpolated model does not provide significant improvements to the translation system performance. In some cases, it is rather penalizing. Indeed, when translating MSA to dialects, we have for example a difference of (-6.19) BLEU points for MSA-MOR system and (-2.99) in the case of MSA-PAL one. In the other direction, interpolation seems to have no significant effect, scores vary by a maximum of (± 0.38) BLEU points except for the MOR-MSA system where the difference is by (-5.04). For inter-dialects translation systems, the largest difference in terms of BLEU scores is noted for the MOR-PAL system (-5.81).

VII. CONCLUSION

In this paper, we first presented an extension of PADIC to a Moroccan dialect. Thus, PADIC covers four dialects of the Maghrebi group and two from Levantine one in addition to MSA. Then, we proceeded to an evaluation of the computerization level of all the six dialects using the σ–index. For all the languages, the σ–index value is below $10/20$ showing that they are π–languages.

Using PADIC, we conducted several machine translation experiments between all the pairs of languages. The best results are achieved with translation systems based on languages that

are closest (ANB-ALG and PAL-SYR). The worst result is achieved between Syrian and Algerian dialects which are, in fact, very different since the Algerian borrowed a lot of French words which do not exist obviously in the Syrian dialect. Concerning the Maghrebi dialects, Moroccan and Tunisian dialects are more close compared to Algerian ones. For MSA, the best results of machine translation have been achieved with Palestinian dialect. This means that the two languages are very close since they share a large number of words.

Due to the small size of the corpus, we analyzed the impact of the language model on the performances of machine translation systems by interpolating it with a larger one trained on well known corpora. Unfortunately the results are not significant even if in some cases we get some improvements.

In the future, we plan to introduce more Arabic dialects to perform more deep experiments and explore ways on how to use the large existing corpora of MSA to rewrite part of them into dialects and by exploiting comparable corpora.

REFERENCES

[1] R. Cotterell and C. Callison-Bursh, "A multidialect, multi-genre corpus of informal written Arabic," in *Proceedings of the Language Resources and Evaluation Conference, LREC-2014*, 2014, pp. 241–245.

[2] P. Dasig and M. Diab, "Codact: Towards identifying orthographic variants in dialectal Arabic," in *Proceedings of the 5th International Joint Conference on Natural Language Processing*. Chiang Mai, Thailand: AFNLP, 2011, pp. 318–326.

[3] K. Salam, N. Habash, and D. Abdulrahim, "A large scale corpus of gulf arabic," in *Proceedings of the Language Resources and Evaluation Conference, LREC-2016*, 2016.

[4] K. Meftouh, S. Harrat, S. Jamoussi, M. Abbas, and K. Smaili, "Machine translation experiments on padic: A parallel Arabic dialect corpus," in *Proceedings of 29th Paclic Asia Conference on Language, Information and Computation*, 2015, pp. 26–34.

[5] R. Tachicart and K. Bouzoubaa, "A hybrid approach to translate moroccan Arabic dialect," in *SITA'14, 9th International Conference on Intelligent Systems*, 2014.

[6] H. Sawaf, "Arabic dialect handling in hybrid machine translation," in *AMTA'2010, 9th Conf. of the Association for Machine Translation in the Americas*, 2010.

[7] W. Salloum and N. Habash, "Elissa: A dialectal to standard Arabic machine translation system," in *Coling'2012, 24th International Conference on Computational Linguistics*, 2012, pp. 385–392.

[8] R. Zbib, E. Malchiodi, D. Jacob, D. Stallard, S. Matsoukas, R. Schwartz, J. Makhoul, O. Zaidan, and C. 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*, ser. NAACL HLT 12, 2012, pp. 49–59.

[9] K. Almeman and M. Lee, "Automatic building of Arabic multi dialect text corpora by bootstrapping dialect words," in *In proceedings of the 1st international conference on Communications, signal processing, and their applications (iccspa)*, 2013, pp. 1–6.

[10] H. Mubarak and K. Darwish, "Using twitter to collect a multi-dialectal corpus of Arabic," in *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP*. Doha, Qatar: Association for Computational Linguistic, 2014, pp. 1–7.

[11] H. Bouamor, N. Habash, and K. Oflazer, "A Multidialectal Parallel Corpus of Arabic," in *Proceedings of the Language Resources and Evaluation Conference, LREC-2014*, 2014, pp. 1240–1245.

[12] V. Berment, "Méthodes pour informatiser les langues et les groupes de langues peu dotées," Ph.D. dissertation, Université Joseph-Fourier-Grenoble I,, 2004.

[13] S. Harrat, K. Meftouh, M. Abbas, S. Jamoussi, and K. Smaili, "Cross-dialectal Arabic processing," in *Computational Linguistics and Intelligent Text Processing, 16th International Conference, CICLing 2015 proceeding, part 1*, April 2015, pp. 620–632.

[14] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open Source Toolkit for Statistical Machine Translation," *Proceedings of the Annual Meeting of the Association for Computational Linguistics, demonstation session*, pp. 177–180, 2007.

[15] F. J. Och and H. Ney, "A Systematic Comparison of Various Statistical Alignment Models," *Computational Linguistics, Volume 29, No 1*, pp. 19–51, 2003.

[16] A. Stolcke, "Srilm – an Extensible Language Modeling Toolkit," in *ICSLP*, Denver, USA, 2002, pp. 901–904.

[17] K. Papineni and al., "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual of the Association for Computational linguistics*, Philadelphia, USA, 2002, pp. 311–318.

[18] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A study of translation edit rate with targeted human annotation," in *Proceedings of of the 7th Conference of the Association for Machine Translation in the Americas (AMTA)*, 2006, pp. 223–231.

[19] M. Maamouri, A. Bies, T. Buckwalter, and H. Jin, "Arabic treebank: Part 3 v 1.0," in *Linguistic Data Consortium*, 2004.