Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Analyse de sentiments des vidéos en dialecte algérien

Mohamed Amine Menacer[1] Karima Abidi[1] Nouha Othman[1,2] Kamel Smaïli[1]
(1) LORIA, Campus Scientifique, BP 239, 54506 Vandoeuvre-lès-Nancy, France
(2) LARODEC, Institut Supérieur de Gestion de Tunis, 2000 Bardo, Tunisia
{mohamed-amine.menacer,karima.abidi,nouha.othman,kamel.smaili}@loria.fr

RÉSUMÉ

La plupart des travaux existant sur l'analyse de sentiments traitent l'arabe standard moderne et ne prennent pas en considération les spécificités de l'arabe dialectal. Cet article présente un système d'analyse de sentiments de textes extraits de vidéos exprimées en dialecte algérien. Dans ce travail, nous avons deux défis à surmonter, la reconnaissance automatique de la parole pour le dialecte algérien et l'analyse de sentiments du texte reconnu. Le développement du système de reconnaissance automatique de la parole est basé sur un corpus oral restreint. Pour pallier le manque de données, nous proposons d'exploiter des données ayant un impact sur le dialecte algérien, à savoir l'arabe standard et le français. L'analyse de sentiments est fondée sur la détection automatique de la polarité des mots en fonction de leur proximité sémantique avec d'autres mots ayant une polarité prédéterminée.

ABSTRACT

Sentiment analysis of videos in Algerian dialect

Most of the existing works on sentiment analysis deal only with Modern Standard Arabic (MSA), and do not take into account the dialects. This article presents a system for analyzing the sentiments of the utterances extracted from videos, in which the language used is Algerian dialects. We have two challenges to overcome, the automatic speech recognition for the Algerian dialect and the sentiment analysis of the recognized text. A spoken corpus has been recorded in order to develop a baseline system for recognizing the videos. This system is then improved by taking advantage of the acoustic data having an impact on the Algerian dialect, namely standard Arabic and French. The sentiment analysis is based on the automatic detection of the polarity of words according to their semantic proximity to other words with a predetermined polarity.

MOTS-CLÉS : Analyse de sentiments, Dialecte algérien, Vidéos, Reconnaissance automatique de la parole.

KEYWORDS: Sentiment analysis, Algerian dialect, Videos, Automatic speech recognition.

1 Introduction

Plusieurs recherches ont été conduites sur la langue arabe. En revanche, la majorité des travaux destinés au traitement automatique de la langue arabe écrite s'est focalisée, de façon presque exclusive, sur l'arabe moderne standard, en laissant de côté les formes vernaculaires. En effet, l'arabe moderne standard est la langue officielle dans le monde arabe. Elle se trouve principalement dans les livres, les journaux, les magazines, et les médias officiels. Elle représente la forme de l'arabe universel enseignée dans les écoles et utilisée dans les discussions formelles. Cependant, la communication

Volume 2 : Traitement Automatique des Langues Naturelles, pages 296–304.
HAL : hal-02784779.

dans la vie quotidienne se fait à travers le dialecte qui est propre à chaque région du monde arabe. Cette forme parlée est essentiellement basée sur l'arabe moderne standard en relâchant plusieurs contraintes morpho-syntaxiques de la langue d'origine pour laisser place à une langue informelle plus simple d'usage. Le dialecte est parfois combiné avec d'autres langues étrangères comme le français ou l'anglais et il ne s'agit pas de simples emprunts, mais d'utilisation de phrases entières en langues étrangères.

Depuis l'apparition des réseaux sociaux, la communauté TAL s'est lancée dans une activité de recherche accrue sur les dialectes arabes. En effet, les internautes expriment leurs sentiments et opinions à propos de différents sujets dans les réseaux sociaux essentiellement en dialecte. L'analyse de sentiments qu'elle soit parlée ou textuelle est un domaine riche en publications (Kiritchenko *et al.*, 2016; Barhoumi *et al.*, 2018; Brahimi *et al.*, 2019). Néanmoins, très peu de travaux dans ce domaine ont été réalisés sur les dialectes.

Dans ce travail, nous nous intéressons à l'étude de sentiments dans les réseaux sociaux où le dialecte algérien est utilisé comme support de communication. Pour ce faire, nous proposons un système de détection de la polarité (sentiment positif ou négatif) pour une collection de vidéos en dialecte algérien. Ces vidéos sont collectées à partir des chaînes algériennes disponibles sur YouTube. Les vidéos sont transcrites à l'aide d'un système de reconnaissance automatique de la parole (SRAP) pour le dialecte algérien, et ensuite l'étude de sentiments est effectué sur les transcriptions.

Le dialecte algérien est l'un des dialectes les plus difficiles à reconnaître par un SRAP. Cela est dû au fait que cette variante de la langue arabe utilise de nombreuses séquences de mots empruntées (principalement de la langue française). En outre, dans ce dialecte les mots de l'arabe standard sont altérés phonologiquement afin d'en faciliter la prononciation (Harrat *et al.*, 2017, 2018). Par ailleurs, les mots empruntés peuvent être utilisés tels quels, ou ils peuvent être modifiés afin de respecter la structure morphologique de la langue arabe.

Pour construire un SRAP robuste, il faut disposer d'une grande quantité de données orales et écrites de la langue à reconnaître. Malheureusement, ce type de données n'existe pas pour le dialecte algérien puisqu'il est principalement parlé de plus, il n'existe pas de normes ni de règles pour l'écrire ce qui rend le traitement des textes existant plus complexe. Notre approche pour reconnaître le dialecte algérien est d'exploiter des données d'autres langues ayant un impact sur le dialecte, à savoir le MSA et le français. Une autre ressource primordiale dans les SRAP est le dictionnaire de prononciation. L'approche la plus simple pour le générer se base sur la décomposition en caractères de chaque mot pour avoir sa prononciation (Le & Besacier, 2009; Killer *et al.*, 2003; Gizaw, 2008). Une autre approche consiste à utiliser des méthodes statistiques pour convertir les graphèmes en phonèmes (Cucu *et al.*, 2011; Karanasou & Lamel, 2010; Harrat *et al.*, 2014; Masmoudi *et al.*, 2018). C'est cette approche que nous avons adoptée pour notre système.

Une fois la transcription des vidéos est générée par le SRAP, nous procédons ensuite à l'analyse de sentiments qui est basée sur la détection de polarité des mots dialectaux composant cette transcription. Cette polarité est déterminée en fonction des mots proches ayant une orientation prédéterminée.

2 Les corpus

Afin de développer et évaluer un système permettant l'analyse de sentiments de vidéos en dialecte algérien, nous avons utilisé plusieurs sources de données qui sont décrites ci-dessous :

YouTubAlg : nous utilisons ce corpus pour calculer l'orientation sémantique des mots du dialecte algérien et pour apprendre le modèle de langage du SRAP. Il comporte des commentaires collectés à partir de YouTube en utilisant l'API [1] de Google. Pour récupérer un maximum de données correspondant au dialecte algérien, nous avons utilisé une liste de mots-clés spécifiques dressée au préalable. Ces mots-clés correspondent principalement à des évènements ou à des personnalités connues relatives à l'actualité algérienne et ne présentant aucun intérêt au niveau international. En effet, ce principe a été utilisé pour éliminer l'éventualité de collecter des commentaires d'arabophone autres qu'algériens. Le corpus obtenu est composé de 18,3M de mots (Abidi *et al.*, 2017)).

ADIC : l'apprentissage du modèle acoustique dans les SRAP est basé sur une collection de données orales avec leur transcription. ADIC (Algerian DIalect Corpus) a été construit en enregistrant, à l'aide d'un microphone unidirectionnel professionnel, 4,6k phrases par 7 locuteurs natifs algériens. Les phrases ont été sélectionnées à partir de deux corpus : YouTubAlg et PADIC (Meftouh *et al.*, 2015, 2018). Ce dernier est une collection de 6,4K phrases en arabe standard avec leurs traductions dans plusieurs dialectes arabes dont deux dialectes algériens. Le corpus obtenu contient 6 heures de parole réparties comme suit : 240 minutes sont utilisées pour l'apprentissage, 40 minutes pour la validation et 70 minutes pour le test.

SentAlgVid : nous utilisons ce corpus pour l'évaluation finale de notre modèle d'analyse de sentiments de vidéos en dialecte. *SentAlgVid* est une collection de vidéos en dialecte diffusées par des chaînes de télévision algériennes comme *Ennahar TV, Echorouk TV, et El Bilad TV*. Le nombre total de vidéos est égale à 30 vidéos d'une durée moyenne de 2 minutes. Les vidéos de ce corpus ont été annotées manuellement en termes de polarité (positive et négative) par des locuteurs natifs.

3 Modèles proposés

Dans ce travail, nous avons deux défis à surmonter, la RAP du dialecte algérien et l'analyse de sentiments de ce dernier. Le modèle final est basé sur une architecture *pipeline* où la sortie du SRAP est utilisée comme entrée de système de l'analyse de sentiments. Dans ce qui suit, nous présentons chaque composant du modèle final proposé.

3.1 Reconnaissance automatique de la parole pour le dialecte algérien

Le développement d'un SRAP est basé sur trois composants : le modèle acoustique modélisant le système phonologique de la langue, le modèle de langage assurant le respect des règles grammaticales et le modèle de prononciation définissant le vocabulaire et les différentes variantes de prononciation.

3.1.1 La modélisation acoustique

Le modèle acoustique est basé sur les réseaux de neurones de type perceptrons multicouches. Ces modèles sont entraînés pour estimer la probabilité d'associer chaque observation acoustique à un triphone. Les observations acoustiques sont des vecteurs fMLLR (feature-space Maximum Likelihood

1. Disponible sur : https ://developers.google.com/YouTube

Linear Regerssion) (Gales, 1998) de dimension 40. Ces observations sont souvent utilisés pour l'apprentissage adaptatif (*speaker Adaptive Training (SAT)* (Anastasakos *et al.*, 1996)) qui vise à rapprocher les observations acoustiques initiaux et cibles par une transformation linéaire. L'architecture est basée sur 6 couches cachées de 2048 neurones chacune. La couche en entrée est composée de 440 neurones représentant la concaténation de 11 observations acoustiques. L'estimation des paramètres du réseau de neurones nécessite une grande quantité de données acoustiques, en revanche, on ne dispose que de 4 heures du dialecte pour l'apprentissage. Pour cette raison, nous avons décidé de tirer profit des langues influençant le dialecte algérien, à savoir : le MSA et le français. C'est pourquoi, ADIC a été étendu en ajoutant progressivement 4 heures de chaque langues jusqu'à arriver à 44 heures. Les données de l'arabe standard sont extraites de deux corpus NEMLAR (Yaseen *et al.*, 2006) et NetDC (Choukri *et al.*, 2004), tandis que les données de la langue française ont été collectées à partir du corpus ESTER (Galliano *et al.*, 2005). La quantité optimale de données acoustiques de chaque langue a été déterminée en minimisant le WER (Word Error Rate) sur la partie de validation de ADIC. Nous sommes arrivés à la conclusion qu'en ajoutant plus de 12 heures du MSA et plus de 12 heures du français aux données dialectales, les performances du SRAP se dégradent.

3.1.2 La modélisation du langage

L'apprentissage du modèle de langage pour le dialecte algérien n'est pas limitée aux données dialectales (les deux corpus PADIC et YouTubAlg), nous utilisons également des données de l'arabe standard. Comme la quantité des différentes données textuelles est déséquilibrée, le modèle de langage, que nous proposons, est une combinaison linéaire de quatre modèles bi-grammes. Deux d'entre eux ont été entraînés sur des données textuelles de l'arabe standard : la version arabe de Gigaword (1 milliard de mots) et la transcription des données acoustiques utilisées pour enrichir ADIC (315 000 mots), les deux autres ont été entraînés sur des données dialectales : PADIC et YouTubAlg. Les poids de l'interpolation linéaire ont été estimés pour maximiser la probabilité d'un corpus de développement composé d'un mélange de données du MSA et du dialecte. Les poids de pondération, calculés sur le corpus de développement, pour chaque corpus sont les suivants : 0,48 pour YouTubAlg, 0,22 pour Gigaword, 0,19 pour la transcription des données orales du MSA et 0,11 pour PADIC.

3.1.3 La modélisation de la prononciation

Le lexique de prononciation est composé de l'union des mots les plus fréquents extraits à partir de chaque ensemble de données utilisé pour l'apprentissage du modèle de langage. Pour chaque mot du lexique, il faut disposer de toutes ses variantes de prononciation. La question est de savoir comment produire toutes les variantes de prononciation possibles pour les mots arabes, et plus particulièrement les mots dialectaux, sachant que les textes arabes sont écrits sans aucune diacritique. Nous avons utilisé un lexique externe (Ali *et al.*, 2014) comme une table de recherche à partir de laquelle les prononciations des mots arabes sont extraites. Malheureusement, nous ne disposons pas de l'équivalent de cette ressource pour le dialecte algérien. Pour remédier à ce problème, nous avons adopté une approche de type G2P (*grapheme-to-phoneme*) afin de produire les variantes de prononciation pour les mots dialectaux. Pour ce faire, nous avons adapté l'approche proposée dans (Harrat *et al.*, 2014). Le processus de conversion G2P commence par la restitution des diacritiques avec un processus automatique basé sur une approche statistique. Ce problème est considéré comme un problème de traduction automatique où la langue source est un ensemble de phrases non voyellées et la langue cible est un ensemble de phrases avec voyelles. Une fois que les diacritiques sont

restituées, un ensemble de règles est utilisé pour produire la prononciation de mots dialectaux (Harrat *et al.*, 2014). Le lexique final contient 125k mots et 538k variantes de prononciation.

3.2 L'analyse de sentiments

Pour déterminer la polarité des mots dialectaux, nous nous sommes basés sur la proximité de leur orientation avec celle de mots de base que nous appellerons des mots-germes. Pour ce faire, nous proposons une méthode qui s'inspire des travaux de (Turney & Littman, 2003) et (Htait *et al.*, 2017) tous deux appliqués à l'anglais. La méthode que nous proposons est composée de deux sous-tâches.

3.2.1 L'identification des mots germes

L'idée consiste à estimer la polarité d'un mot en se basant sur celles des mots d'une liste établie en amont (mots-germes) (Turney & Littman, 2003). Dans ce travail, les mots germes sont ceux dont la polarité est évidente. Dans (Htait *et al.*, 2017), en plus des mots identifiés par Turney, les auteurs ont ajouté une liste d'une quarantaine de mots-germes identifiés à partir des mots les plus fréquents. Ces mots ont été étiquetés, en termes de polarité, manuellement par les auteurs.

Pour ce qui nous concerne, nous avons choisis 80 mots germes à partir d'une liste des mots les plus fréquents de *YouTubAlg*. Dans le tableau 3.2.1 nous donnons quelques exemples de ces mots-germes retenus.

Mots-germes positifs	chaba (*jolie*), bravo, هَايل (*super*), الصحة (*la santé*), belle, شكرًا (*merci*)
Mots-germes négatifs	شيّات (*lèche botte*), harki (traître), جَاهل (*ignorant*), زعفَان (*énervé*), mafia

TABLE 1 – Quelques exemples de mots germes positifs et négatifs.

3.2.2 L'estimation de la polarité des mots du lexique

Dans (Turney & Littman, 2003) les auteurs estiment qu'un mot positif est plus proche des mots germes positifs que des mots germes négatifs s'il est proche des mots-germes positifs et inversement. L'orientation d'un mot est calculée sur la base de la différence entre entre sa similitude avec les mots-germes positifs et les mots-germes négatifs, comme le montre l'équation 1 :

$$SO(w) = \sum_{w_p \in MGP} sim(w, w_p) - \sum_{w_n \in MGN} sim(w, w_n) \quad (1)$$

Où MGP et MGN correspondent respectivement à la liste des mots-germes positifs et négatifs. Le calcul de la similarité est effectué par les auteurs de (Turney & Littman, 2003) en utilisant l'information mutuelle. Pour ce qui nous concerne, nous avons utilisé une représentation distribuée (*word embedding*) apprise sur le corpus YouTubAlg. Ensuite, nous avons calculé la similarité cosinus entre les vecteurs représentatifs de ces mots. La méthode proposée nous a permis de construire un lexique de polarité pour le dialecte algérien comportant 11,2k entrées.

4 Expérimentations

La démarche expérimentale que nous mettons en place consiste à évaluer chaque composant séparément, à savoir : le SRAP et le système d'analyse de sentiments, pour évaluer enfin la sortie finale.

4.1 La reconnaissance automatique de la parole

L'évaluation de notre système de reconnaissance de la parole est basée sur la partie test de ADIC (70 minutes de parole). Nous n'avons pas utilisé le corpus *SentAlgVid* qui est destiné pour l'évaluation de la sortie finale car on ne dispose pas de la transcription de vidéos de ce corpus. Les résultats en terme du WER sont présentés dans le tableau 2.

Système	**Données acoustiques**	**WER(%)**	**OOV (%)**
S_{base}	ADIC	40.0	6.8
S_1	ADIC+44hMSA+40hFr	38.8	
S_2	ADIC+12hMSA+12hFr	**37.7**	

TABLE 2 – Les résultats de reconnaissance de la parole sur la partie test de ADIC.

Le WER du système de base S_{base} entraîné avec seulement les 4 heures de la partie d'apprentissage de ADIC est de 40%. En intégrant toutes les données acoustiques (système S_1) du MSA (44 heures) et du français (40 heures), les performances de S1 sont meilleures de 1,2% par rapport au système de base. Mieux encore, en optimisant la taille des données acoustiques provenant du MSA et du français, nous avons obtenu une amélioration absolue de 2,3% (S_2). Il est à noter que l'intervalle de confiance pour le système de base est de $\pm$1,2%, ce qui signifie que S_2 atteint une amélioration significative par rapport au système de base. Cela montre également que la taille des données utilisées pour apprendre le modèle acoustique pour le dialecte algérien affecte les performances du système de reconnaissance.

Il est à noter que les travaux de recherches sur la RAP pour le dialecte algérien sont relativement moins nombreux pour pouvoir comparer nos résultats. Cependant, dans la dernière édition de la compétition MGB, MGB5 (Ali *et al.*, 2019), il y avait une tâche de RAP pour le dialecte marocain. Ce dernier est relativement proche du dialecte algérien, ils partagent plusieurs aspects linguistiques et acoustiques. Le meilleur système a obtenu un WER de 37,6%, sachant que 13 heures de la parole dialectale ont été utilisées avec 1200 heures de l'arabe standard pour apprendre le modèle acoustique. Cela montre la difficulté de reconnaître les dialectes maghrébins en particulier le dialecte algérien et que les résultats de notre système sont acceptables.

4.2 L'analyse de sentiments

Afin d'évaluer le lexique que nous avons construit d'une manière automatique, il faut disposer d'un corpus de commentaires en dialecte algérien où chacun d'entre eux est associé à une polarité. Ensuite, il faut utiliser le lexique de polarité que nous avons développé pour calculer la polarité sur ce corpus. Malheureusement, ce type de corpus n'existe pas pour le dialecte algérien. Par conséquent, nous avons dû en construire un. Pour ce faire, nous avons annoté manuellement 750 commentaires extraits de YouTube. Cela a donné lieu à 390 commentaires positifs et à 360 commentaires négatifs, avec une

moyenne de 9 mots par commentaire. Nous avons ensuite estimé la qualité du lexique construit en utilisant ce corpus que nous avons nommé *SentAlg*. Pour calculer la polarité d'un commentaire nous sommons la polarité de chacun de ses termes. Dans le tableau 4 nous donnons les résultats obtenus en terme de rappel et de précision. De ces résultats on peut constater que la méthode proposée est

Corpus	Rappel	Précision
SentAlg	88.11%	88.64%

TABLE 3 – Résultats expérimentaux sur le corpus *SentAlg*.

pertinente et a conduit à la construction d'un lexique de polarité pertinent.

Nous avons testé ce lexique également sur les transcriptions automatiques des vidéos de *SentAlgVid*. Les résultats du tableau 4 sont intéressants, même s'ils ne sont pas de la même qualité que ceux obtenus sur des corpus de textes simples. En effet, rappelons que ces résultats sont calculés sur des transcriptions automatiques obtenus à l'aide d'un SRAP dont le WER du système de reconnaissance de la parole est de 37,7%. Nous considérons ce résultat comme très encourageant étant donné le taux d'erreur élevé.

Corpus	Rappel	Précision
SentAlgVid	60%	64.28%

TABLE 4 – Résultats expérimentaux sur le corpus de vidéos de dialecte algérien *SentAlgVid*.

5 Conclusion

Dans cet article, nous avons proposé un système d'analyse de sentiments de vidéos en dialecte algérien. Nous y avons abordé deux problèmes critiques, à savoir la reconnaissance automatique de la parole pour le dialecte algérien et l'analyse de sentiments du texte reconnu. Pour surmonter le problème de manque de données dialectales nécessaires aux différents modèles du SRAP, nous avons exploité des données de langues ayant un impact sur le dialecte, à savoir l'arabe standard et le français. Nous avons montré qu'il est important de doser la quantité de données à utiliser de chaque langue étrangère afin d'améliorer le SRAP. En ce qui concerne l'analyse de sentiments, une méthode a été proposée pour construire automatiquement un lexique de polarité qui a permis d'analyser le contenu de vidéos en dialecte algérien.

Références

ABIDI K., MENACER M. A. & SMAILI K. (2017). CALYOU : A Comparable Spoken Algerian Corpus Harvested from YouTube. In *18th Annual Conference of the International Communication Association (Interspeech)*, Conference of the International Communication Association (Interspeech), Stockholm, Sweden.

ALI A., SHON S., SAMIH Y., MUBARAK H., ABDELALI A., GLASS J., RENALS S. & CHOUKRI K. (2019). The mgb-5 challenge : Recognition and dialect identification of dialectal arabic speech.

ALI A., ZHANG Y., CARDINAL P., DAHAK N., VOGEL S. & GLASS J. (2014). A complete KALDI recipe for building Arabic speech recognition systems. In *Spoken Language Technology Workshop (SLT), 2014 IEEE*, p. 525–529. DOI : 10.1109/SLT.2014.7078629.

ANASTASAKOS T., MCDONOUGH J., SCHWARTZ R. & MAKHOUL J. (1996). A compact model for speaker-adaptive training. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, volume 2, p. 1137–1140 vol.2. DOI : 10.1109/ICSLP.1996.607807.

BARHOUMI A., ALOULOU C., CAMELIN N., ESTÈVE Y. & BELGUITH L. (2018). Arabic Sentiment analysis : an empirical study of machine translation's impact. In *Language Processing and Knowledge Management international conference (LPKM2018)*, Sfax, Tunisia.

BRAHIMI B., TOUAHRIA M. & TARI A. (2019). Improving sentiment analysis in arabic : A combined approach. *Journal of King Saud University - Computer and Information Sciences*.

CHOUKRI K., NIKKHOU M. & PAULSSON N. (2004). Network of data centres (NetDC) : BNSC - an Arabic broadcast news speech corpus. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal : European Language Resources Association (ELRA).

CUCU H., BESACIER L., BURILEANU C. & BUZO A. (2011). Investigating the role of machine translated text in ASR domain adaptation : Unsupervised and semi-supervised methods. In *2011 IEEE Workshop on Automatic Speech Recognition Understanding*, p. 260–265. DOI : 10.1109/ASRU.2011.6163941.

GALES M. J. (1998). Maximum likelihood linear transformations for HMM-based speech recognition. *Computer speech & language*, **12**(2), 75–98.

GALLIANO S., GEOFFROIS E., MOSTEFA D., CHOUKRI K., BONASTRE J.-F. & GRAVIER G. (2005). The ester phase ii evaluation campaign for the rich transcription of french broadcast news. In *Ninth European Conference on Speech Communication and Technology*.

GIZAW S. (2008). Multiple pronunciation model for Amharic speech recognition system. In *Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU)*.

HARRAT S., MEFTOUH K., ABBAS M. & SMAÏLI K. (2014). Grapheme to phoneme conversion - an Arabic dialect case. In *Spoken Language Technologies for Under-resourced Languages*.

HARRAT S., MEFTOUH K. & SMAÏLI K. (2017). Creating Parallel Arabic Dialect Corpus : Pitfalls to Avoid. In *18th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING)*, Budapest, Hungary.

HARRAT S., MEFTOUH K. & SMAÏLI K. (2018). Maghrebi Arabic dialect processing : an overview. *Journal of International Science and General Applications*, **1**.

HTAIT A., FOURNIER S. & BELLOT P. (2017). Identification semi-automatique de mots-germes pour l'analyse de sentiments et son intensité. In *COnférence en Recherche d'Informations et Applications - CORIA 2017, 14th French Information Retrieval Conference, Marseille, France, March 29-31, 2017. Proceedings.*, p. 415–424.

KARANASOU P. & LAMEL L. (2010). Comparing SMT methods for automatic generation of pronunciation variants. In *International Conference on Natural Language Processing*, p. 167–178 : Springer.

KILLER M., STUKER S. & SCHULTZ T. (2003). Grapheme based speech recognition. In *Eighth European Conference on Speech Communication and Technology*.

KIRITCHENKO S., MOHAMMAD S. & SALAMEH M. (2016). Semeval-2016 task 7 : Determining sentiment intensity of english and arabic phrases. In S. BETHARD, D. M. CER, M. CARPUAT, D. JURGENS, P. NAKOV & T. ZESCH, Éds., *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*, p. 42–51 : The Association for Computer Linguistics.

LE V.-B. & BESACIER L. (2009). Automatic speech recognition for under-resourced languages : application to Vietnamese language. *IEEE Transactions on Audio, Speech, and Language Processing*, **17**(8), 1471–1482.

MASMOUDI A., BOUGARES F., ELLOUZE M., ESTÈVE Y. & BELGUITH L. (2018). Automatic speech recognition system for Tunisian dialect. *Language Resources and Evaluation*, **52**(1), 249–267. DOI : 10.1007/s10579-017-9402-y.

MEFTOUH K., HARRAT S., JAMOUSSI S., ABBAS M. & SMAILI K. (2015). Machine translation experiments on PADIC : A parallel arabic dialect corpus. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, p. 26–34.

MEFTOUH K., HARRAT S. & SMAÏLI K. (2018). PADIC : extension and new experiments. In *7th International Conference on Advanced Technologies ICAT*, Antalya, Turkey. HAL : hal-01718858.

TURNEY P. D. & LITTMAN M. L. (2003). Measuring praise and criticism : Inference of semantic orientation from association. *ACM Transactions on Information Systems*, **21**(4). DOI : 10.1145/944012.944013.

YASEEN M., ATTIA M., MAEGAARD B., CHOUKRI K., PAULSSON N., HAAMID S., KRAUWER S., BENDAHMAN C., FERSØE H., RASHWAN M., HADDAD B., MUKBEL C., MOURADI A., AL-KUFAISHI A., SHAHIN M., CHENFOUR N. & RAGHEB A. (2006). Building annotated written and spoken Arabic LRs in NEMLAR project. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy : European Language Resources Association (ELRA).