ADIDA: Automatic Dialect Identification for Arabic

Ossama Obeid, Mohammad Salameh,[†] Houda Bouamor,[†] Nizar Habash
New York University Abu Dhabi, UAE
[†]Carnegie Mellon University in Qatar, Qatar
{oobeid,nizar.habash}@nyu.edu
{msalameh,hbouamor}@cmu.edu

Abstract

This demo paper describes ADIDA, a web-based system for automatic dialect identification for Arabic text. The system distinguishes among the dialects of 25 Arab cities (from Rabat to Muscat) in addition to Modern Standard Arabic. The results are presented with either a point map or a heat map visualizing the automatic identification probabilities over a geographical map of the Arab World.

1 Introduction

The last few years have witnessed an increased interest within the natural language processing (NLP) community in the computational modeling of dialectal and non-standard varieties of languages (Malmasi et al., 2016; Zampieri et al., 2017, 2018). The Arabic language, which is a collection of variants or dialects, has received a decent amount of attention in this regard with a number of efforts focusing on dialect identification, translation and other forms of modeling. In this demo paper, we present ADIDA,[1] a public online interface for visualizing fine-grained dialect identification of Arabic text (Salameh et al., 2018). The dialect identification system produces a vector of probabilities indicating the likelihood an input sentence is from 25 cities (Table 1) and Modern Standard Arabic (MSA). ADIDA displays the results with either a point map or a heat map overlaid on top of a geographical map of the Arab World.

2 Arabic and its Dialects

Although MSA is the official language across the Arab World, it is not the native language of any speakers of Arabic. Dialectal Arabic (DA), on the other hand, is the daily informal spoken variety. DA is nowadays emerging as the primary language of communication – not just spoken, but also written, particularly in social media. Arabic dialects are often classified in terms of geographical regions, such as Levantine Arabic, Gulf Arabic and Egyptian Arabic (Habash, 2010). However, within each of these regional groups, there is significant variation down to the village, town, and city levels. The demo we present is based on the work of Salameh et al. (2018), who utilize the MADAR Project parallel corpus of 25 Arab cities plus MSA (Table 1) (Bouamor et al., 2018).[2]

Arabic dialects differ in various ways from MSA and from each other. These include phonological, morphological, lexical, and syntactic differences (Haeri, 1991; Holes, 2004; Watson, 2007; Bassiouney, 2009). Despite these differences, distinguishing between Arabic dialects in written form is an arduous task because: (i) dialects use the same writing script and share part of the vocabulary; and (ii) Arabic speakers usually resort to repeated code-switching between their dialect and MSA (Abu-Melhim, 1991; Bassiouney, 2009), creating sentences with different levels of dialectness (Habash et al., 2008).

3 Related Work

3.1 Arabic Dialect Processing

While automatic processing of DA is relatively recent compared to MSA, it has attracted a considerable amount of research in NLP (Shoufan and Al-Ameri, 2015). Most of it focuses on (i) collecting datasets from various sources and at different levels (Zaidan and Callison-Burch, 2011; Khalifa et al., 2016; Abdul-Mageed et al., 2018; Bouamor et al., 2018), (ii) creating processing tools (Habash et al., 2013; Al-Shargi and Rambow, 2015; Obeid et al., 2018) (iii) developing DA to English ma-

[1] https://adida.abudhabi.nyu.edu/
The Arabic word عديدة /ςadida/ means 'numerous'.

[2] https://camel.abudhabi.nyu.edu/madar/

Proceedings of NAACL-HLT 2019: Demonstrations, pages 6–11
Minneapolis, Minnesota, June 2 - June 7, 2019.

Region	Maghreb				Nile Basin	Levant		Gulf		Yemen
Sub-region	**Morocco**	**Algeria**	**Tunisia**	**Libya**	**Egypt/Sudan**	**South Levant**	**North Levant**	**Iraq**	**Gulf**	**Yemen**
Cities	Rabat Fes	Algiers	Tunis Sfax	Tripoli Benghazi	Cairo Alexandria Aswan Khartoum	Jerusalem Amman Salt	Beirut Damascus Aleppo	Mosul Baghdad Basra	Doha Muscat Riyadh Jeddah	Sana'a

Table 1: Different city dialects covered in ADIDA and the regions they belong to.

chine translation systems (Zbib et al., 2012; Sajjad et al., 2013), (iv) or performing dialect identification (Zaidan and Callison-Burch, 2014; Huang, 2015; Salameh et al., 2018).

3.2 Dialect Identification

Dialect Identification (DID) is a particularly challenging task compared to Language Identification (Etman and Beex, 2015). Since Arabic dialects use the same script and share part of the vocabulary, it is quite arduous to distinguish between them. Hence, developing an automatic identification system working at different levels of representation and exploring different datasets has attracted increasing attention in recent years. For instance, DID has been the goal of a dedicated shared task (Malmasi et al., 2016; Zampieri et al., 2017, 2018), encouraging researchers to submit systems to recognize the dialect of speech transcripts for dialects of four main regions: Egyptian, Gulf, Levantine and North African, and MSA. Several systems implementing a range of traditional supervised learning (Tillmann et al., 2014) and deep learning methods (Belinkov and Glass, 2016; Michon et al., 2018) were proposed.

In the literature, a number of studies have been exploring DID using several datasets, ranging from user-generated content (i.e., blogs, social media posts) (Sadat et al., 2014), speech transcripts (Biadsy et al., 2009; Bougrine et al., 2017), and other corpora (Elfardy and Diab, 2012, 2013; Zaidan and Callison-Burch, 2014; Salameh et al., 2018; Dinu et al., 2018; Goldman et al., 2018). Shoufan and Al-Ameri (2015) and Al-Ayyoub et al. (2017) present a survey on NLP and deep learning methods for processing Arabic dialectal data with an overview on Arabic DID of text and speech. While most of the proposed approaches targeted regional or country level DID, Salameh et al. (2018) introduced a fine-grained DID system covering the dialects of 25 cities from several countries across the Arab world (from Rabat to Muscat), including some cities in the same country.

3.3 Visualization

Map visualizations are used in multiple fields of study including linguistics, socio-linguistics, and political science to display geographical relations of non-geographic data. Geographical visualizations may include point maps to display individual data points, choropleths and Voronoi tessalation maps that cluster data points by region, and heat maps and surface maps that interpolate data over some geographical area.

In the general context of visualization of language data, one example is the Visualizing Medieval Places project (Wrisley, 2017, 2019), which extracted place names from medieval French texts and overlaid them over their physical locations as a point map with a color ramp to display their frequency. The Linguistic Landscapes of Beirut Project (Wrisley, 2016) visualizes the presence of multilingual written samples within the greater Beirut area using different geographical visualizations to explore different aspects of its data. Specifically in the context of dialectometric visualizations, most relevant to this paper, Scherrer and Stoeckle (2016) provide surface and Voronoi tessalation maps[3] to visualize difference in Swiss German dialects using data extracted from the *Sprachatlas der deutschen Schweiz*. Similarly, data collected from *The Harvard Dialect Survey* (Vaux and Golder, 2003) used point maps to display phrase variation across American English dialects. Katz and Andrews (2013) provide further visualization of *The Harvard Dialect Survey* using heat maps to interpolate data from the survey.

4 Design and Implementation

4.1 Design Considerations

The underlying system we use for dialect identification can work with any number of words (single words, phrases or sentences) and produces probabilities of occurrence in different locales in a one dimensional vector (with 26 values in our case). As such, we want an interface that can visualize

[3] http://dialektkarten.ch/dmviewer

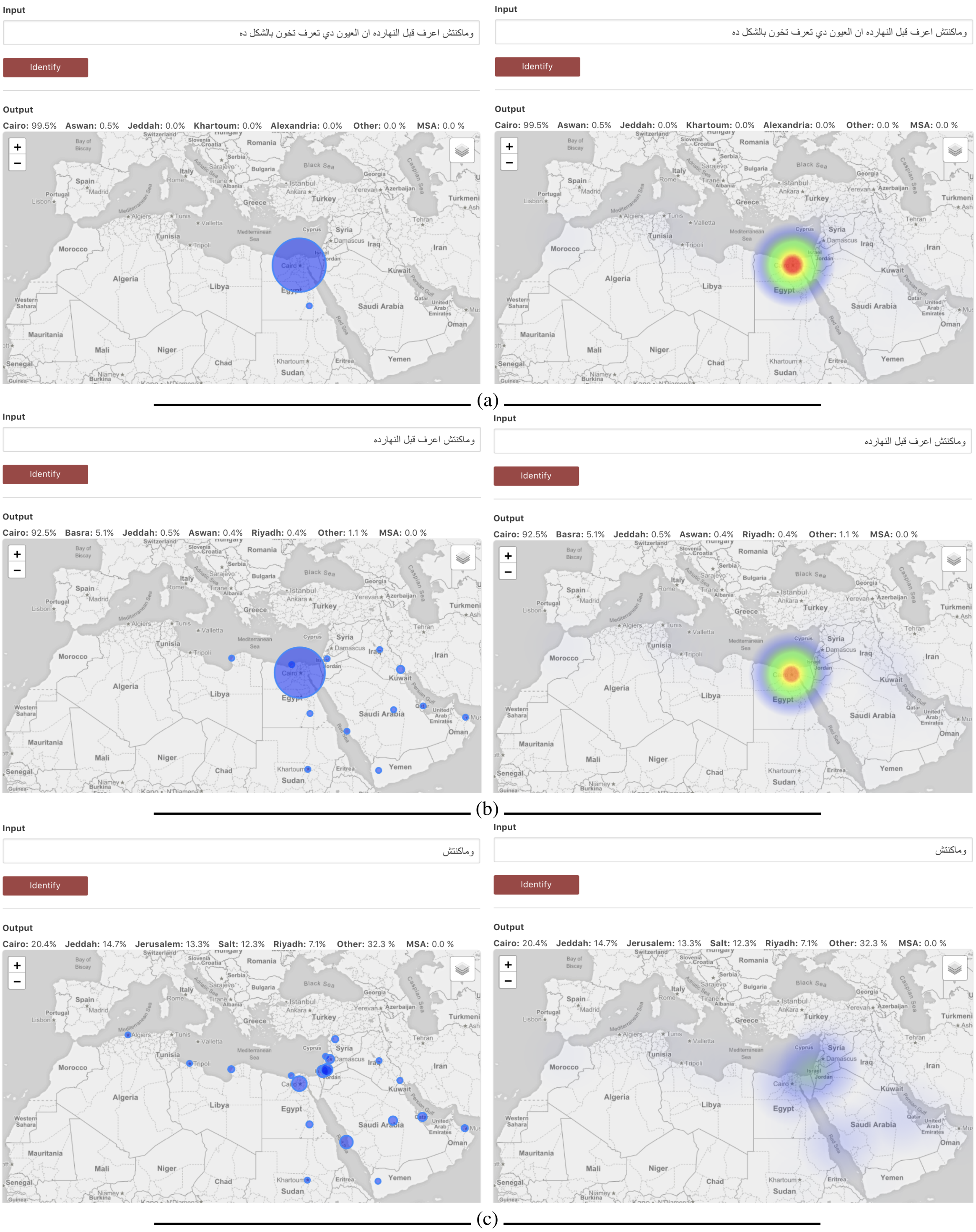

Figure 1: ADIDA Interface showing the output for a verse from an Egyptian Arabic song in the two display modes: point map (right) and heat map (left). The subfigures (a), (b) and (c) correspond to different lengths of the verse: (a) full, (b) first four words, and (c) the first word only.

the probability distribution into a two-dimensional geographical map space allowing us to easily observe and debug connections and patterns relating to dialectal similarities and differences that are harder to catch in the one dimensional output of the system classifier. We also want to visualize aggregations of probabilities of nearby cities that give a sense of regional presence.

Our setup and needs are different from other dialect map visualization efforts discussed in Section 3.3 which mostly focus on specific concepts and their realizations in different forms.

4.2 The ADIDA Interface

The ADIDA interface is publicly available at `https://adida.abudhabi.nyu.edu/`. Figure 1.(a, *left side*) presents the basic structure of the interface. At the top there is a box to input the Arabic text to dialect identify. The web page automatically fills the box with a randomly selected song verse from a set of well known songs from different dialects. This is intended to make it easy for the user to understand the task of the interface. After the user clicks on the *Identify* button, a geographical map of the Arab world is shown with one of two toggleable overlays: (1) a point map displaying one point per city scaled to the probability of attribution to the city (default mode), or (2) a heat map that plots the probabilities as Gaussians centered on each city with proportional intensities that aggregate any nearby points at a given zoom level. The point map only shows cities that have an attribution probability larger than 0.1% while the heat map displays Gaussians for all cities. Both visualization modes exclude MSA as there is no geographical location that can represent it. The heat map should not be interpreted to make claims about the attribution probabilities of regions between the considered cities. The falloff of each Gaussian and their aggregates are used solely as a high-level visualization aid through allowing aggregation of probabilities of nearby cities. Additionally, the interface presents the top five cities with their probabilities, together with that of MSA and of the remaining probability mass assigned to *Other*. We discuss the rest of the screen shots in Figure 1 in Section 4.4.

4.3 Implementation

Back-end The ADIDA back-end was implemented in Python using `Flask`[4] to create a Web API wrapper for the dialect ID code. The core dialect ID application is based on the best performing model distinguishing between 26 classes (25 dialects and MSA), described in Salameh et al. (2018). The application makes use of scikit-learn (Pedregosa et al., 2011) to learn a Multinomial Naive Bayes (MNB) classifier using the MADAR corpus (Bouamor et al., 2018), a large-scale collection of parallel sentences built to cover the dialects of 25 cities from the Arab World (Table 1), in addition to MSA. The model is fed with a suite of features covering word unigrams and character unigrams, bigrams and trigrams weighted by their Term Frequency-Inverse Document Frequency (TF-IDF) scores, combined with language model scores.The output of the MNB model is a set of 26 probability scores referring to the 25 cities and MSA. Results on a test set show that the model can identify the exact city of a speaker at an accuracy of 67.9% for sentences with an average length of 7 words. Salameh et al. (2018) reported on an oracle study showing that accuracy can reach more than 90% with 16-word inputs.

Front-end The front-end was implemented using `Vue.js`[5] for model view control. We use `Leaflet`[6] with `Mapbox`[7] to provide the geographical map display. We also use `heatmap.js`[8] to generate the heat maps.

4.4 Example

Figure 1 demonstrates the output of ADIDA for a verse from an Egyptian Arabic song (Hafez, 1963). The left side of Figure 1 shows the default point-map mode, while the right side shows the heat-map mode. In Figure 1.(a), the full verse of 11 words is returns a correct preference for Cairo at a high degree of confidence (99.5% probability). In Figure 1.(b) and (c), the length is reduced first to the first four words, and then to the very first word only. In all three cases, Cairo is the top choice, but with decreasing confidence correlating with the length of the input: 99.5% $>$ 92.5% $>$ 20.4%. Additionally we see a great diffusion of the probability score, with the case of one word input resulting with more probability mass in the other 20 cities that are not shown than in the first choice.

5 Conclusion and Future Work

We presented ADIDA, a public online interface for visualizing a system for fine-grained dialect identification. This system produces a vector of probabilities indicating the likelihood an input sentence is from 25 cities and MSA. ADIDA displays the results as a point map or a heat map overlaid on top of a geographical map of the Arab World.

In the future, we plan to continue improving our dialect identification back-end. We also plan to extend the interface in a number of ways: (a) provide

[4] `http://flask.pocoo.org/`

[5] `https://vuejs.org/`

[6] `https://leafletjs.com/`

[7] `https://www.mapbox.com/`

[8] `https://www.patrick-wied.at/static/heatmapjs/`

a display mode that better serves color-blind individuals, (b) provide a feedback mode that can be used to collect additional data provided by users with their quality judgments, and (c) gamify the interface to allow the use of it as a tool to identify more cities in the Arab World.

The data we use in building the back-end is made available as part of a shared task on Arabic fine-grained dialect identification (Bouamor et al., 2019).

Acknowledgments The work presented was made possible by grant NPRP 7-290-1-047 from the Qatar National Research Fund (a member of the Qatar Foundation). The statements made herein are solely the responsibility of the authors. We would like to thank David Wrisley and Yves Scherrer for their valuable feedback and insights.

References

Muhammad Abdul-Mageed, Hassan Alhuzali, and Mohamed Elaraby. 2018. You tweet what you speak: A city-level dataset of Arabic dialects. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan.

Abdel-Rahman Abu-Melhim. 1991. Code-switching and linguistic accommodation in Arabic. In *Proceedings of the Annual Symposium on Arabic Linguistics*, volume 80, pages 231–250.

Mahmoud Al-Ayyoub, Aya Nuseir, Kholoud Alsmearat, Yaser Jararweh, and Brij Gupta. 2017. Deep learning for Arabic nlp: A survey. *Journal of Computational Science*.

Faisal Al-Shargi and Owen Rambow. 2015. Diwan: A dialectal word annotation tool for Arabic. In *Proceedings of the Workshop for Arabic Natural Language Processing (WANLP)*, pages 49–58, Beijing, China.

Reem Bassiouney. 2009. *Arabic Sociolinguistics: Topics in Diglossia, Gender, Identity, and Politics*. Georgetown University Press.

Yonatan Belinkov and James Glass. 2016. A Character-level Convolutional Neural Network for Distinguishing Similar Languages and Dialects. In *Proceedings of the Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 145–152, Osaka, Japan.

Fadi Biadsy, Julia Hirschberg, and Nizar Habash. 2009. Spoken Arabic Dialect Identification Using Phonotactic Modeling. In *Proceedings of the Workshop on Computational Approaches to Semitic Languages (CASL)*, pages 53–61, Athens, Greece.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. The MADAR Arabic Dialect Corpus and Lexicon. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan.

Houda Bouamor, Sabit Hassan, Nizar Habash, and Kemal Oflazer. 2019. The MADAR shared task on Arabic fine-grained dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop (WANLP)*, Florence, Italy.

Soumia Bougrine, Hadda Cherroun, and Djelloul Ziadi. 2017. Hierarchical Classification for Spoken Arabic Dialect Identification using Prosody: Case of Algerian Dialects. *CoRR*, abs/1703.10065.

Liviu P. Dinu, Alina Maria Ciobanu, Marcos Zampieri, and Shervin Malmasi. 2018. Classifier ensembles for dialect and language variety identification. *CoRR*, abs/1808.04800.

Heba Elfardy and Mona Diab. 2012. Token level identification of linguistic code switching. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, Mumbai, India.

Heba Elfardy and Mona Diab. 2013. Sentence Level Dialect Identification in Arabic. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, pages 456–461, Sofia, Bulgaria.

Asma Etman and Louis Beex. 2015. Language and Dialect Identification: A Survey. In *Proceedings of the Intelligent Systems Conference (IntelliSys)*, London, UK.

Jean-Philippe Goldman, Yves Scherrer, Julie Glikman, Mathieu Avanzi, Christophe Benzitoun, and Philippe Boula de Mareil. 2018. Crowdsourcing Regional Variation Data and Automatic Geolocalisation of Speakers of European French. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Nizar Habash, Owen Rambow, Mona Diab, and Reem Kanjawi-Faraj. 2008. Guidelines for Annotation of Arabic Dialectness. In *Proceedings of the Workshop on HLT & NLP within the Arabic World*, Marrakech, Morocco.

Nizar Habash, Ryan Roth, Owen Rambow, Ramy Eskander, and Nadi Tomeh. 2013. Morphological Analysis and Disambiguation for Dialectal Arabic. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, Atlanta, Georgia.

Nizar Y Habash. 2010. *Introduction to Arabic natural language processing*, volume 3. Morgan & Claypool Publishers.

Niloofar Haeri. 1991. Sociolinguistic Variation in Cairene Arabic: Palatalization and the qaf in the Speech of Men and Women.

Abdel Halim Hafez. 1963. Gabbar (Arrogant). Lyrics by Hessien El Sayed.

Clive Holes. 2004. *Modern Arabic: Structures, Functions, and Varieties*. Georgetown Classics in Arabic Language and Linguistics. Georgetown University Press.

Fei Huang. 2015. Improved Arabic Dialect Classification with Social Media Data. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2118–2126, Lisbon, Portugal.

Josh Katz and Wilson Andrews. 2013. How Yall, Youse and You Guys Talk. https://www.nytimes.com/interactive/2014/upshot/dialect-quiz-map.html.

Salam Khalifa, Nizar Habash, Dana Abdulrahim, and Sara Hassan. 2016. A Large Scale Corpus of Gulf Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Portorož, Slovenia.

Shervin Malmasi, Marcos Zampieri, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, and Jörg Tiedemann. 2016. Discriminating between Similar Languages and Arabic Dialect Identification: A Report on the Third DSL Shared Task. In *Proceedings of the Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 1–14, Osaka, Japan.

Elise Michon, Minh Quang Pham, Josep Crego, and Jean Senellart. 2018. Neural network architectures for Arabic dialect identification. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 128–136.

Ossama Obeid, Salam Khalifa, Nizar Habash, Houda Bouamor, Wajdi Zaghouani, and Kemal Oflazer. 2018. MADARi: A Web Interface for Joint Arabic Morphological Annotation and Spelling Correction. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

Fatiha Sadat, Farnazeh Kazemi, and Atefeh Farzindar. 2014. Automatic Identification of Arabic Dialects in Social Media. In *Proceedings of the Workshop on Natural Language Processing for Social Media (SocialNLP)*, pages 22–27, Dublin, Ireland.

Hassan Sajjad, Kareem Darwish, and Yonatan Belinkov. 2013. Translating dialectal Arabic to English. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, pages 1–6, Sofia, Bulgaria.

Mohammad Salameh, Houda Bouamor, and Nizar Habash. 2018. Fine-grained arabic dialect identification. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 1332–1344, Santa Fe, New Mexico, USA.

Yves Scherrer and Philipp Stoeckle. 2016. A quantitative approach to Swiss German dialectometric analyses and comparisons of linguistic levels. *Dialectologia et Geolinguistica*, 24(1):92–125.

Abdulhadi Shoufan and Sumaya Al-Ameri. 2015. Natural language processing for dialectical Arabic: A survey. In *Proceedings of the Workshop for Arabic Natural Language Processing (WANLP)*, page 36, Beijing, China.

Christoph Tillmann, Saab Mansour, and Yaser Al-Onaizan. 2014. Improved Sentence-Level Arabic Dialect Classification. In *Proceedings of the Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 110–119, Dublin, Ireland.

Bert Vaux and Scott Golder. 2003. The Harvard Dialect Survey. *Cambridge, MA: Harvard University Linguistics Department*.

Janet CE Watson. 2007. *The Phonology and Morphology of Arabic*. Oxford University Press.

David Joseph Wrisley. 2016. Linguistic Landscapes of Beirut Project. http://llbeirut.org.

David Joseph Wrisley. 2017. Locating medieval French, or why we collect and visualize the geographic information of texts. *Speculum*, 92(S1):S145–S169.

David Joseph Wrisley. 2019. "Aggregate map." Visualizing Medieval Places. http://vmp.djwrisley.com/map/.

Omar Zaidan and Chris Callison-Burch. 2014. Arabic dialect identification. *Computational Linguistics*, 40(1):171–202.

Omar F Zaidan and Chris Callison-Burch. 2011. The Arabic Online Commentary Dataset: an Annotated Dataset of Informal Arabic With High Dialectal Content. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, pages 37–41.

Marcos Zampieri, Shervin Malmasi, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer, and Noëmi Aepli. 2017. Findings of the VarDial Evaluation Campaign 2017. In *Proceedings of the Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 1–15, Valencia, Spain.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Ahmed Ali, Suwon Shon, James Glass, Yves Scherrer, Tanja Samardžić, Nikola Ljubešić, Jörg Tiedemann, Chris van der Lee, Stefan Grondelaers, Nelleke Oostdijk, Antal van den Bosch, Ritesh Kumar, Bornini Lahiri, and Mayank Jain. 2018. Language identification and morphosyntactic tagging: The second VarDial evaluation campaign. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, Santa Fe, USA.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F. Zaidan, and Chris Callison-Burch. 2012. Machine Translation of Arabic Dialects. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 49–59, Montréal, Canada.