Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6627–6632
Marseille, 11–16 May 2020

DaCToR: A Data Collection Tool for the RELATER Project

[†]Juan Hussain, [+†]Oussama Zenkri, [†]Sebastian Stüker, [*†]Alexander Waibel
[†]Institute for Anthropomatics and Robotics
KIT - Karlsruhe Institute of Technology, Germany
firstname.lastname@kit.edu
[+] firstname.lastname@student.kit.edu
[*] Language Technologies Institute
Carnegie Mellon University
firstname.lastname@cmu.edu

Abstract

Collecting domain-specific data for under-resourced languages, e.g., dialects of languages, can be very expensive, potentially financially prohibitive and taking long time. Moreover, in the case of rarely written languages, the normalization of non-canonical transcription might be another time consuming but necessary task. In order to collect domain-specific data in such circumstances in a time and cost-efficient way, collecting read data of pre-prepared texts is often a viable option. In order to collect data in the domain of psychiatric diagnosis in Arabic dialects for the project RELATER, we have prepared the data collection tool DaCToR for collecting read texts by speakers in the respective countries and districts in which the dialects are spoken. In this paper we describe our tool, its purpose within the project RELATER and the dialects which we have started to collect with the tool.

Keywords: speech, data collection tool, under-resourced languages, dialects

1. Introduction

Speech recognition, speech translation and speech synthesis systems have a very wide range of applications today. These systems, which employ modern learning algorithms usually require large amount of data to achieve acceptable performance. However, Obtaining sufficient amount of data is often a problem for low-resource languages, such as dialects or non-written languages, and for domain-specific fields, such as the medical domain. For training speech recognition systems one needs to collect sufficient amounts of spoken data that is transcribed at sentence level. Collecting data in the exact domain to be addressed in the required speaking style, such as spontaneous speech in psychiatric interviews, is often infeasible. Instead, voice data is often collected from the Internet and transcribed manually (Hernandez et al., 2018). Another common source of voice data may be telephone calls (Canavan, Alexandra, David Graff, and George Zipperlen, 1997) or news broadcasts (Walker, Kevin, et al., 2014). These methods require significant transcription effort, especially if no automatic transcription system is available. In the case of rare dialects, another problem can be the ambiguity of the transcription created by different transcribers, which requires additional post-processing effort for normalization. In order to collect data in a cost efficient way and avoid the transcription effort, we record domain-specific texts that are read by appropriate speakers of the language or dialect in question. Consequently, we avoid the problems of non-canonical transcriptions and writing systems. Though read speech is not necessarily the required speaking style, it is possible in this way to collect a variety of speech data stemming from many speakers. This contributes to improved speaker and channel independence in case of multiple recordings of the same text.

As part of the project RELATER (see Section 3.), we have developed a tool [1] to collect speech data for the field of psychiatric interviews in Arabic dialects for Arabic-speaking refugees with insufficient knowledge of German.

The rest of the paper is structured as follows. In Section 3. we will introduce the RELATER project followed by a description of our speech data collection tool in Section 4. Next, we present in section 5. the linguistic background of the acquired language data, followed by an outlook on possible improvements in future work in Section 6.

2. Related Work

AIKUMA (Bird et al., 2014) is an open source Android app which allows recording, re-speaking and oral transcribing from different synchronized mobile phones with text-less user interface. LIG-AIKUMA (Gauthier et al., 2016) is an extension of AIKUMA for collecting parallel data. The re-speaking feature, used to record the same speech more clearly, is adapted for recording the translation. SPICE (Speech Processing - Interactive Creation and Evaluation Toolkit for new Languages) (Schultz et al., 2007) is a web-based toolkit for rapid prototyping of speech and language processing components. An Audio recorder for data collection is embedded as well. Most corpora are built by first, performing automatic segmentation of audio with, for instance, LIUM which is a speaker diarization toolkit (Meignier and Merlin, 2010) and second, using a pre-trained automatic transcription tool. However, this method is not suitable for under-resourced languages or dialects. Further, Arabic Multi-Dialectal Transcription Tool AMADAT (Maamouri et al., 2004) is used for annotating audio in dialectal Arabic with some conventions, since the dialectal Arabic is not standardized. A similar tool to our

[1] Our source code is available on *https://github.com/Juan-hussain/dactor*

work is the *SpeechRecorder* tool [2] offered by the Center for Speech Technology Research-*CSTR* from the University of Edinburgh. The proposed tool allows the user to create a separate recording for each utterance. However, this tool lacks many features, which make it inconvenient for our project. First, the software is only available for Mac OS, which restricts the user base of the tool. Second, the user interface is not intuitive as it requires multiple steps before the recording can start and the transition between utterances requires switching the window to select the next utterance. Third, the text in question must obey a predefined structure before it can be imported.

3. RELATER

With the arrival of large numbers of fugitives from the armed conflicts in the Arabic area, e.g. in Syria, public institutions in Germany are faced with bridging the language barrier in order to communicate with the refugees. Among the many communication scenarios in which the language barrier needs to be bridged are also situations of medical care. One part of the medical care that needs to be provided is the psychiatric and psychotherapeutic diagnosis of refugees. Psychiatry and psychotherapy is especially dependent on successful communication, thus bridging the language barrier in these situations is of particular importance. More so than previous waves of migrants, current refugees in Germany speak languages (largely Levantine and Iraqi Arabic) and come from cultural backgrounds in which the German therapeutic community is not versed.

While skilled interpreters are in principal a good solution, solely relying on them has several problems and drawbacks: a) they are often hard to find, b) not available around the clock, and c) not financed for many caregivers such as hospitals. These problems pose major obstacles in providing the necessary assessment and care.

The goals of the project *Removing language barriers in treating refugees—RELATER*, funded by the German Ministry of Education and Research, is to develop a cross-lingual communication system using the latest advances in automatic spoken language translation to bridge the language barrier in these situations of psychiatric diagnosis. Specifically, we will address situations that use the the M.I.N.I. International Neuropsychiatric Interview (Sheehan et al., 1998), the most widely used psychiatric structured diagnostic interview instrument in the world, employed by mental health professionals and health organizations in more than 100 countries.

In a second step we will then address bridging the language gap in a smartphone based interaction system that enables therapists to stay in contact with patients once they move to new locations as they are being settled in Germany.

We will especially address translating from and to dialects of Arabic, as the native language of a large portion of the refugees in Germany is a dialect of Arabic, such as Syrian Levantine.

The creation of spoken language translation systems for dialects, such as the dialectal variants of Arabic, is one of the current research challenges of automatic speech translation. Often, resources for these dialects are scarce and sometimes they are not even written or no canonical writing conventions for them exist.

In order to be able to train, develop and test speech translation technology for such dialects we will need to collect a validation set, an evaluation set and at least small amounts of adaptation data for the dialects of interest within RELATER.

4. DaCToR

This section introduces the developed speech Data Collection Tool for the RELATER project -*DaCToR*. The purpose of the software is to record pre-labeled data. Manual labeling of data is a very time-consuming task, especially for under-resourced languages, such as the various Arabic dialects for which reliable transcription algorithms are not available. We have, therefore, designed this tool to record texts read by the user who have to signal the transition to the next sentence. Hence, the software records the speech and logs the timestamp at the end of each spoken phrase. In the following sections, we will introduce the main User Interface *UI* (section 4.1.), explain how to use it (section 4.2.), present the output files and their formats (section 4.3.) and finally introduce the used automatic timestamp readjustment function (section 4.4.).

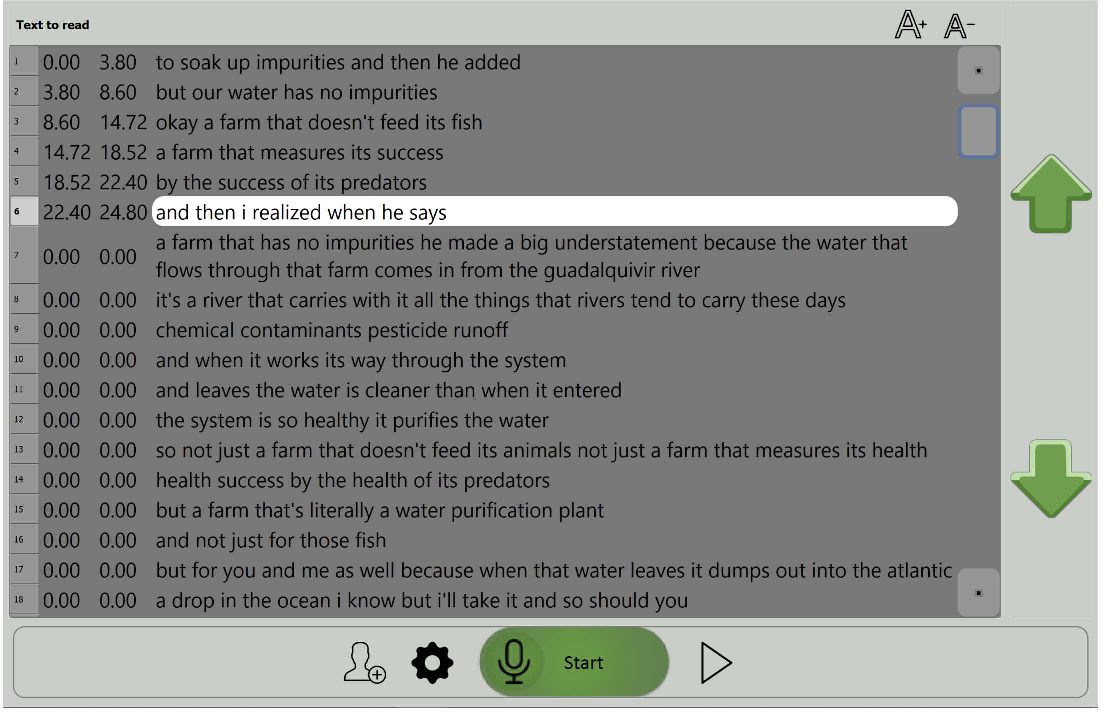
4.1. User Interface

While developing this tool we tried to keep the UI as intuitive as possible through minimalist design and familiar UI-elements. Hence, we arranged the main window into three fields as can be seen in figure 1. The main area consists of two buttons for text-size adjustment and a table containing the text to read and the start and end timestamps of each sentence. As the tool is likewise intended to be used with touch devices, which might lack a physical keyboard, we dedicated a field on the right of the main area which houses two navigation buttons allowing the user to navigate between the different phrases. The last section at the bottom of the interface houses the buttons required to control the software. The function of each key, although not labeled, can be easily determined by its symbol. The central button is for starting and stopping the recording. The right button from the center can be used to play a recorded phrase. The first left button from the center is dedicated to the loading of new text files. The leftmost button serves for switching the account or creating a new one.

4.2. How to Use

When the user starts the software, (s)he is prompted to either login to an existing account or create a new one. The account is not only used to store the individual work made by each user but also to distinguish between the various speakers through considering some helpful information about them. The required information to create a new account are a username, which will be used to log in, the birth year of the speaker, the gender, the country of origin, the spoken dialect and the level of education. After successfully completing the form, the software creates a

[2] SpeechRecorder is available on:
http://www.cstr.ed.ac.uk/research/projects/speechrecorder/



1	0.00	3.80	to soak up impurities and then he added
2	3.80	8.60	but our water has no impurities
3	8.60	14.72	okay a farm that doesn't feed its fish
4	14.72	18.52	a farm that measures its success
5	18.52	22.40	by the success of its predators
6	22.40	24.80	and then i realized when he says
7	0.00	0.00	a farm that has no impurities he made a big understatement because the water that flows through that farm comes in from the guadalquivir river
8	0.00	0.00	it's a river that carries with it all the things that rivers tend to carry these days
9	0.00	0.00	chemical contaminants pesticide runoff
10	0.00	0.00	and when it works its way through the system
11	0.00	0.00	and leaves the water is cleaner than when it entered
12	0.00	0.00	the system is so healthy it purifies the water
13	0.00	0.00	so not just a farm that doesn't feed its animals not just a farm that measures its health
14	0.00	0.00	health success by the health of its predators
15	0.00	0.00	but a farm that's literally a water purification plant
16	0.00	0.00	and not just for those fish
17	0.00	0.00	but for you and me as well because when that water leaves it dumps out into the atlantic
18	0.00	0.00	a drop in the ocean i know but i'll take it and so should you

Figure 1: Main interface of DaCToR after recording the first six phases.

folder named as the username in a sub-folder of the default documents directory. This folder contains two directories, one for the audio files and one for the log files in stm format (section 4.3.). After successfully logging in for the first time, the table in the main area remains empty until the user selects a text file after being prompted to perform this action. The next time the user logs in, the tool restores the session as left when closing the software. This allows the user to listen to her/his recordings, modify them and continue the recording. At this point, the user can start recording by pressing the start button. The phrase to be read is highlighted while the rest of the phrases is deemphasized through a low contrast between the text and the background. After reading the phrase, the user can either pause the recording or move on to the next one, without pausing the recording, by pressing the down arrow. In case the user is unfamiliar with the vocabulary, recording the phrases one by one through hitting the stop button at the end of each phrase might be the more convenient option. In both cases, only one single output audio file is generated for each text file. The keystroke or the stop of the recording is also used to mark the time window of the recorded phrase. If the down arrow has been pressed while recording the last phrase, the recording stops automatically. If the user doubts the quality of a recorded phrase, (s)he is able to listen to that phrase by selecting it with the mouse or the arrow-keys and pressing the play button. Re-recording a recorded phrase is performed through selecting that phrase and pressing the start button. By closing or switching account, a post processing step is performed. This includes rearranging the timestamps of non sequential recordings and removing the gaps caused by re-recording some phrases.

4.3. Produced Data

The tool generates a folder for each user. This folder contains a folder for the recorded audio files, a speaker information file and a second folder for the segment time marked *stm*-like files. The advantage of using such files is the convenient structure for speech recognition and synthesis. The file is structured in columns as follows:
utterance id, speaker name, audio file name, *from* timestamp in second, *to* timestamp in second and the text, for instance:
text1_spk1_00000_01140 spk1 recording1 0.00 1.70 to solve problems
text1_spk1_00000_01140 spk1 recording1 1.70 3.81 think outside the box

4.4. Automatic Timestamp Readjustment

Our software allows the user to navigate between the different phrases using the arrow-keys. While moving on to the following phrase without having to stop recording speeds up the process, most of the users tend to press the arrow key before they completely finished speaking the phrase. This behaviour leads to misalignment issues. In this tool, we solve this problem by adjusting the user-signaled timestamps after each keystroke. The solution we implement is based on the work of (Giannakopoulos, 2009), which aims to automatically segment audio files by removing silent sequences. The proposed method estimates silences in the audio through calculating the signal energy $E(i)$ and the spectral centroid C_i and applying a threshold to each calculated feature vector:

$$E(i) = \frac{1}{N} \cdot \sum_{n=1}^{N} |x_i(n)|^2 \qquad (1)$$

$$C_i = \frac{\sum_{k=1}^{N} (k+1) \cdot X_i(k)}{\sum_{k=1}^{N} X_i(k)}, \qquad (2)$$

where x_i denotes the discrete audio frame with N samples, X_i its Discrete Fourier Transform *DFT* and i the frame index. Both features are calculated for signal frames with a duration of $50ms$ each, for which the speech signal can be seen as stationary.

Signal frames whose feature values are below the thresholds are considered as silence. In case multiple silences occur in the evaluated sequence, only the longest silence will be considered. In our work, we evaluate a 2 seconds long sequence (100 frames) from the audio signal around the user denoted timestamp with the suggested method. The midpoint of the detected silence frame is used as final timestamp. A visualization of the output of the algorithm on a sample audio is shown in figure 2.

5. Data Collection: Arabic Dialects

5.1. Language Background

The Arabic language is the best representation of language Diglossia, introduced by Ferguson (Ferguson, 1959), as the situation where two or more varieties of the same language are used by some speakers. The superposed variety, or high variety *H*, is used in education, news broadcasts, political and religious texts, and the regional dialects (low variety *L*) are used, for instance, in informal daily conversations, social media and in folk literature. In addition, the *H* is standardized, i.e has very stable norms of grammar, vocabulary, orthography and pronunciation. It can be divided into Classical Arabic, which is the language of old literature, and the Modern Standard Arabic *MSA* (section 5.1.2.).

On the other hand, regional dialects are not standardized and have no standard orthographies (section 5.1.1.). In section 5.1.3., we will see examples of lexical, morphological and phonological differences between the high variety and low variety. The details about Tunisian can be read in (Zribi et al., 2013).

5.1.1. Dialect Orthography

Since the Dialects have no standard orthography many works try to find conventions for audio transcription and collected text normalization. An important work performed was the Linguistic Data Consortium guidelines for transcribing Levantine and Iraqi (Maamouri et al., 2004). It suggests a strategy of the transcription for dialectal Arabic by using MSA-based orthographic conventions, since Arab transcribers use their MSA Knowledge for the transcription of Arabic dialects. This includes using both symbols and rules of MSA orthography e.g. writing without short vowels and diacritical marks except for nunation. Inspired by this work, CODA is invented (Habash et al., 2012), a conventional orthography for dialectal Arabic, which is intended to be for general writing purposes and abstracts from phonological variations in sub-dialects in contrast to the previous work. This Work covers Egyptian dialect EGY in details and extended by (Zribi et al., 2014) for Tunisian dialect, by (Saadane and Habash, 2015) for Algerian Arabic and by (Turki et al., 2016) for Maghrebi Arabic. For the data collection, an automatic conversion from spontaneous orthography of EGY to CODA is developed as a freely available tool called CODAFY (Eskander et al., 2013).

Another challenge the data collectors are facing, is the writing system for Arabic using English characters, the so called "Arabizi", which is not a letter-based transliteration (Yaghan, 2008). Instead, short vowels are used similar to sound-to-letter rules of English. To circumvent this problem, (Al-Badrashiny et al., 2014) developed a tool for EGY to convert Arabizi to CODA.

5.1.2. Modern Standard Arabic

As we have seen, the standardized varieties of Arabic are divided into the classic and Modern Standard Arabic *MSA*. According to (Abdelali, 2004), MSA is a modernization of classical Arabic including words for modern phenomena and additions from dialectal Arabic. MSA is the formal language of 21 Arab countries with 290 million speakers (Abdelali, 2004). It is used in education, news, formal speeches, reports, newspapers, and most cartoon, historical and documentary movies. An important study, conducted by (Abdelali, 2004) on the differences of MSA in different Arab countries, collected data from 10 newspapers from 10 different Arab countries. Taken together, 48% of the words in the newsletters were unique to one resource, compared to 52% of common words found at least in two resources; however, this may be expected from Zipf's law. As reported, the main differences are: first, the spelling. However all the examples were about the grapheme Alif with or without Hamza and its position above or beneath the Alif. Second, the different usage of words, for instance, the word *Hjz* [3] (arrest) is used in Morocco and *twqyf* is used in Algeria. Other differences are using the transliteration of foreign words or using loanwords instead of Arabic ones, such as, the word *kwbry* (bridge) used in Egypt and *jsr* used in other countries.

5.1.3. Syrian Levantine Dialect

Levantine dialect is spoken in Syria, Lebanon, Jordan, Palestine and Israel. Accordingly, different variants of Levantine can be considered. However, all variants show a combination of root-and-pattern and affixational morphology (Habash and Rambow, 2006). The **morphological** difference to MSA is enormous in such a manner that a morphological analyser for MSA reaches only 60% coverage of verb forms in Levantine as stated in (Habash and Rambow, 2006). As an example, the MSA word *s+yktb+hA* (will write it) is in Syrian Levantine *bdo yktb+A*, where the future proclitic *s* is mapped to the word *bdo* and *hA* (it) to the suffix *A* (some speakers pronounce *hA* unchanged). **Lexical** differences are mostly restricted in the daily frequent used vocabulary, such as *lyš* (why) in MSA *lmAðA*. **Phonologically**, the consonant /q/[4] in MSA, is realized as the glottal stop /'/ in many words of Syrian Levantine. The consonant /θ/ is either pronounced as /t/, for instance, the MSA *mθl* (similar) is changed to *mtl* or as /s/ as in *msAl* (Example)

[3] Arabic transliteration in alphabetical order (Habash et al., 2007): $Abt\theta jHxd\check{\eth}rzs\check{s}SDT\check{D}\zeta\gamma fqklmnhwy$

[4] the symbols in /./ is from the International Phonetic Alphabet (IPA)

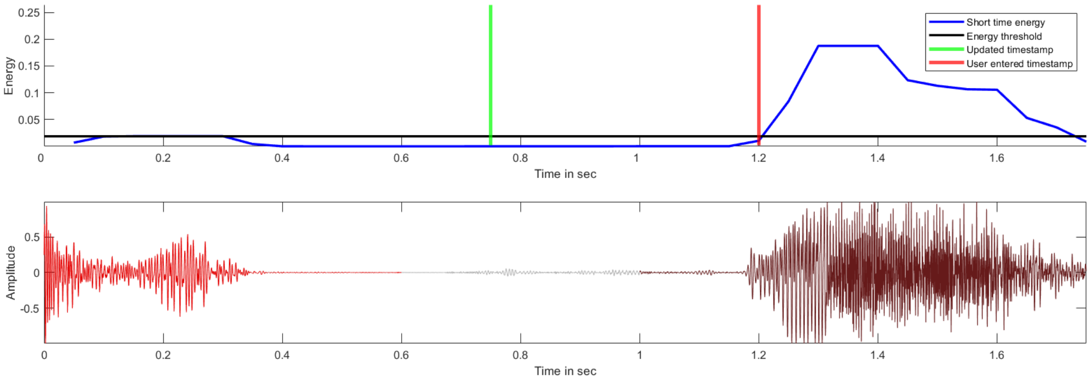


Figure 2: Visualization of the output of the automatic timestamp readjustment. The figure shows the recalculation of the silence midpoint in green.

from the MSA *mθAl*. Similarly, /ð/ is pronounced as /z/ such in *kzb* (lie) from MSA *kðb* and as /d/ as in *dhb* (gold) from MSA ð*hb*. For more Details see (Habash et al., 2012).

5.2. Data Collection

The official data collection phase of the project RELATER has just begun at the time of submitting this paper. So far, we have had volunteers for collecting Tunisian dialect and Syrian Levantine data. One of the participants has already been trained to normalize the collected texts according to the CODA conventions 5.1.1.. For Syrian, the texts are extracted from a novel found on the Web and short stories from whats-app and for Tunisian the texts are from blogs, Facebook posts, short stories and jokes.

So far, two participants recorded about 6 hours of Syrian Levantine until the time of paper submission and one participant recorder about 2 hours of Tunisian dialect.

One obstacle we face is the number of volunteers, especially such that have the necessary hardware for the data collection.

6. Future Work

In the future, we intend to develop a web-based version for desktop as well as for mobile devices, since mobile devices are more available for potential users. Thus, roles need to be assigned, for instance, normal users who read the texts and text publishers who manage these texts, load them online and assign different parts to different users. Furthermore, the ability to input parallel data, i.e. recording audio in a target language as well as transcribing it needs to be implemented. In addition, we intend to include certain tags in the raw text to highlight the words and sentences which should be read in different way to capture prosodic variations, e.g., in intonation, tone, stress, or rhythm.

7. Acknowledgement

The work in this paper was funded by the German Ministry of Education and Science within the project *Removing language barriers in treating refugees—RELATER*, no. 01EF1803B.

8. Bibliographical References

Abdelali, A. (2004). Localization in modern standard arabic. *Journal of the American Society for Information Science and technology*, 55(1):23–28.

Al-Badrashiny, M., Eskander, R., Habash, N., and Rambow, O. (2014). Automatic transliteration of romanized dialectal arabic. In *Proceedings of the eighteenth conference on computational natural language learning*, pages 30–38.

Bird, S., Hanke, F. R., Adams, O., and Lee, H. (2014). Aikuma: A mobile app for collaborative language documentation. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 1–5.

Eskander, R., Habash, N., Rambow, O., and Tomeh, N. (2013). Processing spontaneous orthography. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 585–595.

Ferguson, C. A. (1959). Diglossia. *word*, 15(2):325–340.

Gauthier, E., Blachon, D., Besacier, L., Kouarata, G.-N., Adda-Decker, M., Rialland, A., Adda, G., and Bachman, G. (2016). Lig-aikuma: A mobile app to collect parallel speech for under-resourced language studies.

Giannakopoulos, T. (2009). A method for silence removal and segmentation of speech signals, implemented in matlab. *University of Athens, Athens*, 2.

Habash, N. and Rambow, O. (2006). 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*, pages 681–688.

Habash, N., Soudi, A., and Buckwalter, T. (2007). On arabic transliteration. In *Arabic computational morphology*, pages 15–22. Springer.

Habash, N., Diab, M. T., and Rambow, O. (2012). Conventional orthography for dialectal arabic. In *LREC*, pages 711–718.

Hernandez, F., Nguyen, V., Ghannay, S., Tomashenko, N., and Estève, Y. (2018). Ted-lium 3: Twice as much data and corpus repartition for experiments on speaker adaptation. *Lecture Notes in Computer Science*, page 198–208.

Maamouri, M., Buckwalter, T., and Cieri, C. (2004). Dialectal arabic telephone speech corpus: Principles, tool design, and transcription conventions. In *NEMLAR International Conference on Arabic Language Resources and Tools, Cairo*, pages 22–23.

Meignier, S. and Merlin, T. (2010). Lium spkdiarization: an open source toolkit for diarization.

Saadane, H. and Habash, N. (2015). A conventional orthography for algerian arabic. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 69–79.

Schultz, T., Black, A. W., Badaskar, S., Hornyak, M., and Kominek, J. (2007). Spice: Web-based tools for rapid language adaptation in speech processing systems. In *Eighth Annual Conference of the International Speech Communication Association*.

Sheehan, D., Lecrubier, Y., Sheehan, K., Amorim, P., Janavs, J., Weiller, E., Hergueta, T., Baker, R., and Dunbar, G. (1998). The mini-international neuropsychiatric interview (m.i.n.i.): the development and validation of a structured diagnostic psychiatric interview for dsm-iv and icd-10. *The Journal of Clinical Psychiatry*, 59 (suppl 20):22–23.

Turki, H., Adel, E., Daouda, T., and Regragui, N. (2016). A conventional orthography for maghrebi arabic. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC), Portoroz, Slovenia*.

Yaghan, M. A. (2008). “arabizi”: A contemporary style of arabic slang. *Design issues*, 24(2):39–52.

Zribi, I., Graja, M., Khmekhem, M. E., Jaoua, M., and Belguith, L. H. (2013). Orthographic transcription for spoken tunisian arabic. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 153–163. Springer.

Zribi, I., Boujelbane, R., Masmoudi, A., Ellouze, M., Belguith, L. H., and Habash, N. (2014). A conventional orthography for tunisian arabic. In *LREC*, pages 2355–2361.

9. Language Resource References

Canavan, Alexandra, David Graff, and George Zipperlen. (1997). *CALLHOME American English Speech LDC97S42*. Philadelphia: Linguistic Data Consortium.

Walker, Kevin, et al. (2014). *GALE Phase 2 Arabic Broadcast News Speech Part 1 LDC2014S07*. Philadelphia: Linguistic Data Consortium.