LinTO Audio and Textual Datasets to
Train and Evaluate Automatic Speech Recognition in Tunisian Arabic Dialect

Hedi Naouara, Jean-Pierre Lorré, Jérôme Louradour
LINAGORA

hnaouara@linagora.com, jplorre@linagora.com, jlouradour@linagora.com

Abstract

Developing Automatic Speech Recognition (ASR) systems for Tunisian Arabic Dialect is challenging due to the dialect's linguistic complexity and the scarcity of annotated speech datasets. To address these challenges, we propose the LinTO audio and textual datasets – comprehensive resources that capture phonological and lexical features of Tunisian Arabic Dialect. These datasets include a variety of texts from numerous sources and real-world audio samples featuring diverse speakers and code-switching between Tunisian Arabic Dialect and English or French. By providing high-quality audio paired with precise transcriptions, the LinTO audio and textual datasets aim to provide qualitative material to build and benchmark ASR systems for the Tunisian Arabic Dialect.

Keywords — Tunisian Arabic Dialect, Speech-to-Text, Low-Resource Languages, Audio Data Augmentation

Introduction

In recent years, artificial intelligence has made significant advances in the fields of natural language processing (NLP) and automatic speech recognition (ASR). Pioneering models such as wav2vec (Baevski et al. 2020), OpenAI Whisper (Radford et al. 2022) and Meta Massively Multilingual Speech (Tjandra et al. 2022) boast support for dozens to hundreds of languages. For example, despite massive amounts of training data (680,000 hours of transcribed speech for Whisper), recent models fail to correctly transcribe with Arabic dialects. For these reasons, we have collected data in the Tunisian Arabic dialect to provide high-quality resources for effective training and evaluation of ASR systems. This paper presents our datasets and describes some preliminary results that show that they can be used to improve ASR sytems for Tunisian Arabic Dialect.

Challenges and Existing Initiatives for Tunisian

The linguistic evolution of the Tunisian Arabic dialect has been shaped by a complex historical background, influenced by various civilisations, including Amazigh (Berber) (Tilmatine 1999), Phoenician, Roman and Arab cultures, as well as the Ottoman Empire - now Turkey (Ouerhani 2009), French colonialism and modern globalisation. These diverse influences have contributed to Tunisian lexical richness and distinct phonetic characteristics based on regional varieties (Gibson 1999; Zribi et al. 2017), which pose significant challenges for the development of ASR systems. This task is further hindered by the limited availability of annotated speech datasets for Tunisian Arabic Dialect.

Many Arabic dialects are underrepresented, including Tunisian Arabic. This dialect presents unique challenges due to its phonetic complexity (Gayraud et al. 2018), lack of standardized dictionaries (Saidi 2007), lack of formal writing rules, and frequent code-switching with French and English. Although previous initiatives, such as TunSwitch (Abdallah et al. 2023) and TARIC (Mdhaffar et al. 2024) have made progress, further efforts are needed to build a comprehensive dataset for Tunisian Arabic Dialect. TARIC was collected entirely at a single train station, resulting in limited topics like greetings and ticket prices. TunSwitch is too small to train a model from scratch. And while the Massive Arabic Speech Dataset (MASC) (Al-Fetyani et al. 2021) contains 1000 hours of speech collected from YouTube, it includes only 3 hours of Tunisian.

Contribution

To address these challenges, we present the LinTO datasets for Tunisian Arabic Dialect, comprising a diverse collection of text and audio contents.

The LinTO textual dataset in Tunisian includes content from various sources such as films and TV series, rap lyrics, documentaries, stories, and more. This diverse collection was curated to support the training of language models for the Tunisian Arabic Dialect. We also implemented tailored normalization to address the language's unique characteristics, such as standardizing the spelling of dialect-specific words, handling code-switching with French and English, and transliterating Arabizi. These normalizations enhance the linguistic quality and coverage of the language model, especially when training data is limited.

The LinTO audio (raw and augmented) datasets in Tunisian include a wide range of recordings, such as music, documentaries, podcasts, TV shows, radio broadcasts, educational stories, and narratives of prophets, each accompanied by a transcript. To enhance dataset diversity, we employed data augmentation techniques on raw audio data. Specifically, Voice Conversion Augmentation was used to introduce more speaker variety into datasets with significant amounts of single-speaker recordings (*e.g.* story telling). Noise reduction techniques were applied to improve the clar-

ity and quality of the audio, addressing the challenges posed by the limited availability of annotated speech datasets. Cleaned audio was then be enhanced by adding noise, reverberation and other effects.

Our datasets are designed to support speech recognition tasks for the Tunisian Arabic Dialect and feature code-switching between Tunisian, English, and French. They are meticulously organized into multiple configurations and splits to facilitate a variety of experimental setups, making them valuable resources for research and development in speech processing. We release three datasets under the CC BY 4.0 license on Hugging Face: an audio dataset with transcribed speech in Tunisian[1], an augmentation of this dataset using Voice Conversion Augmentation[2] and a textual dataset gathering several sources of text in Tunisian Arabic Dialect[3].

In the following sections, we provide a detailed overview of the methodologies and procedures used to collect and process our Tunisian Arabic Dialect datasets. First, we discuss the compilation of textual corpora, highlighting the diversity of sources, the challenges and our solutions. We then outline the strategies used for collecting transcribed audio recordings and describe data augmentation techniques. Finally, we carry out some preliminary experiments to provide a first baseline ASR model for the Tunisian Arabic Dialect.

Textual Corpus in Tunisian Arabic Dialect

Language Models (LM) are a crucial element in many ASR systems. They help determine the most likely spelling for a given sequence of phonemes, making their fine-tuning essential for optimal system performance. Unlike Acoustic Models, which predict lattices of phoneme or character probabilities based on audio frames, LMs do not require aligned audio data and can be trained solely on text, which is more cost-effective to collect.

The LinTO Tunisian text dataset was compiled from various sources, including public datasets on Hugging Face:
- TuDiCoI: A Tunisian dialogue dataset built by ARBML (An Arabic Researchers Community).
- Brahim Mohamed: A Tunisian dialect summary dataset created to test Llama2.

several datasets from GitHub:
- T-HSAB: A Tunisian Hate Speech and Abusive Language dataset.
- TSAC: A Tunisian Sentiment Analysis Corpus.
- BARD: A dataset of Arabic book reviews.
- Tunbert: A dataset of daily Tunisian communications.

We include in addition DrejjaToEnglish dataset from Kaggle the TunSwitch dataset trancripts. Finally, we crawled various blogs and websites:
- Stories: Chakhabit, HkayetErwi, TunHistoires, Lbachch.
- Blog posts: TunHistoires, Lbachch.
- TV and film transcripts: ChroniqueChroniyet, KisatiAna.
- RAP lyrics: A Tunisian rap lyrics dataset.
- Tweets: A collection of Tunisian Tweets.

---

[1]https://huggingface.co/datasets/linagora/linto-dataset-audio-ar-tn
[2]https://huggingface.co/datasets/linagora/linto-dataset-audio-ar-tn-augmented
[3]https://huggingface.co/datasets/linagora/linto-dataset-text-ar-tn

language(s)	# lines	# words	# unique words
Arabic	4 269 k	15 964 k	255 k
French or English	207 k	594 k	28 k
code-switching	65 k	385 k	34 k
overall	**4 541 k**	**16 944 k**	**289 k**

Table 1: Composition of the LinTO Tunisian textual corpus.

The links corresponding to each source are provided in the dataset card of the LinTO dataset on Hugging Face.

This effort resulted in over 4.5 million lines of text and approximately 288,000 unique words. The composition of the dataset is shown in Table 1. Notably, the dataset includes sentences with code-switching between Tunisian and French or English, as well as some sentences exclusively in French or English. This inclusion enriches the vocabulary with terms sometimes used in spoken Tunisian.

The Tunisian Arabic Dialect lacks a standardized dictionary, leading to various spellings based on pronunciation (Ouerhani 2009). For instance, هذا (meaning "this" or "that" and often pronounced as "hatha") can appear as هاذا or هذى. This variability complicates modeling and evaluation. To address this issue, we created a normalization dictionary that maps more than 12,500 words to their standardized forms, to have consistent representations of words.

Several materials collected from social media and YouTube include Arabizi, a writing style that mixes Latin letters and numbers to represent Arabic sounds. This challenges our models, which need text in Tunisian Arabic characters. For example, the Arabizi phrase "9alou y9awi sa3dek 9alou taw taw taw" should be transliterated into "قالوا يقوي سعدك قالوا تو تو تو". We found that models for arabizi transliteration such as (Talafha, Abuammar, and Al-Ayyoub 2021) are not suitable for the Tunisian Arabic Dialect, even though they perform well on Modern Standard Arabic. To ensure accurate transcription, we semi-automatically transliterated Arabizi transcripts, by correcting a first automatic translation.

Transcribed Recordings of Spoken Tunisian

The lack of annotated audio datasets is a significant challenge in developing an ASR system for dialects. We focused on collecting clear and intelligible audio, representing diverse speakers and contexts. In this section, we describe the data collected and the main pre-processing steps.

Dataset Composition

Capturing diverse contexts in the training data is essential for effective speech modeling. To enhance diversity, we gathered several existing datasets and collected dozens of additional hours of data by crawling the web. The LinTO audio dataset, whose composition is detailed in Table 2, includes:
- the transcribed part of TunSwitch (Abdallah et al. 2023), which can be divided into two parts: 75% with code-switching (CS) and 25% with Tunisian Arabic only (TO)
- several datasets available in Hugging Face:

source	subset	TRAIN SET			TEST SET	
		audio dur.	VCA aug.	num. words	audio dur.	num. words
TunSwitch	**CS**	10H01	×7.0	75 k	27m	4 k
	TO	3H20	×6.7	18 k	28m	3 k
Hugging Face	**Ameni Kh**	4H05	×1	32 k	3m	0.5 k
	Arbi Houssem	3H50	×1	33 k		
	MA Konyali	3H27	×1	20 k		
YouTube	**story telling**	27H17	×6.2	148 k		
	theme channels	20H12	×1.3	113 k		
	TV & radio	7H48	×8.0	48 k		
	misc. crawled	4H08	×8.8	19 k	19m	1 k
	shorts	3H47	×8.0	28 k		
	MASC	2H53	×7.9	12 k		
websites	**OneStory**	1H33	×7.1	12 k	3m	1 k
	Appr.LeTunisien	0H38	×5.2	1 k	3m	0.15 k
	overall	**92H56**	**466H**	**560 k**	**1H20**	**10 k**
TARIC (not in released data)		7H25	52H	57 k	0H50	7 k

Table 2: Composition of the LinTO Tunisian audio dataset and TARIC. Audio durations are indicated with/without VC Augmentation. Number of transcribed words are also given.

- – Arbi Houssem:[4] TV shows
- – Ameni Kh:[5] TV and radio content
- – MA Konyali:[6] isolated words with augmentation
- the Tunisian subset of MASC (Al-Fetyani et al. 2021), comprising YouTube audios with provided transcripts. We re-extracted audio and transcripts from MASC URLs using our YouTube crawling pipeline.

Our experiments include TARIC (Mdhaffar et al. 2024) but we do not re-distribute it within the LinTO audio dataset due to its non-commercial license.

We also crawled Tunisian educational platforms such as ApprendreLeTunisien[7] and OneStory Media[8], and YouTube videos with curated captions taken as transcripts. All audio recordings are sampled at 16 kHz. The resulting data fall into the following categories:

- **story telling**: Abdel Aziz Erwi, Hkeytet Tounsia Mensia, Lobna Majjedi
- **theme channels**: Bayari Bilionaire (soccer), Hamza Baloumi El Mohakek (crime), QLM media (history)
- **TV and radio**: Carthage plus and Telvza TV (TV) Diwan FM (radio), Mohammed Khammesi (podcast)
- **short videos**: Short videos of radio shows, announcements and some manually selected jokes
- **misc. crawled videos**: This data was manually collected by reviewing content to select annotated audio from diverse channels. We faced challenges because YouTube does not filter audios by specific dialects, so we manually identified Tunisian audios from those tagged as Arabic.

Transcript Curation

Combining existing datasets with online sources highlighted challenges like incorrect audio annotations. Accurate annotations are critical for training effective Speech-to-Text models and must align with our standardization dictionary.

Over 85% of our YouTube dataset was transcribed using YouTube's Auto-Transcribe. While the results were generally acceptable, some issues arose, such as word truncation (e.g., تع and نع for متاع) and the substitution of incorrect words (e.g., معنتها was frequently transcribed as مع). To improve transcription quality, we referred to a standardized dictionary that included both truncated and complete forms of words. For more complex cases, we manually verified the transcripts by listening to the corresponding audio segments.

Segmentation of Long Audio Segments

ASR systems are usually trained on 15 to 30 second audio segments. We thus segmented longer audio files into chunks not exceeding 30 seconds. As described in (Zhu and Zhang 2021), we implemented a forced alignment to accurately align audio segments with their respective transcripts, based on a wav2vec ASR model trained on Modern Standard Arabic (Grosman 2021). This alignment procedure was also used to correct misalignments in YouTube captions.

Augmentation of Audio Data

Several data augmentation techniques are used to improve the robustness of acoustic modeling. These methods expand the training dataset without the need for real speech data, helping the model handle variations in speakers, recording conditions and background noises. Standard techniques include: speed change, volume change, simulated reverberation and background effects. These standards or classical data augmentation techniques are generally applied just before the training phase and are cheap to compute. Contrary to the augmented data described below, such "classical" augmentation is not included in the LinTO audio raw corpus.

Voice Conversion Augmentation (VCA)

Several datasets contain single speaker recordings. To avoid overfitting certain voices, we used voice modification to augment single-speaker recordings and increase diversity in an augmented version of the LinTO audio dataset.

In preliminary experiments, we compared several voice modification models and found that the most realistic results were obtained using SoftVC VITS Voice Conversion (voicepaw et al. 2024), an implementation of Adversarial Training on Soft Speech Units with state-of-the-art methods proposed by (van Niekerk et al. 2022; Kim, Kong, and Son 2021). We used HuBert (Hsu et al. 2021) as a pretrained encoder, and HiFi-GAN (Kong, Kim, and Bae 2020) for the decoder. We used a few dozen minutes of speech from 7 speakers, including Tunisian story tellers as well as one non native speaker, and trained generative models from those voices.

MASC and data crawled on YouTube (subset "misc. crawled") include a lot of musical clips and especially noisy

ASR		YouTube	TunSwitch CS	TO	Apprendre Le Tunisien	TARIC	OneStory
OpenAI's Whisper	`large v3`	82.0	93.8	60.4	66.7	92.6	49.8
	`large v3 turbo`	89.6	117.9	62.9	74.8	110.1	53.1
	`large v2`	78.7	101.2	62.3	66.7	102.8	54.1
LinTO ar-tn Kaldi ASR	`No-Aug`	47.8	58.5	35.5	28.3	25.7	26.9
	`No-VCA`	**34.4**	25.1	**20.6**	26.4	**13.9**	4.8
	`VCA`	36.6	**20.2**	21.9	**23.3**	16.1	**4.5**

Table 3: Word Error Rates (%) of different ASR models across different subsets within the LinTO audio dataset in Tunisian.

recordings. To improve quality, we use a source separation tool called Deezer Spleeter (Hennequin et al. 2020) to isolate speech from noise and music. The augmented dataset includes both original and cleaned audio recordings, and VC augmentation was done on cleaned data.

The augmented dataset includes 466 hours of audio, with details for each subset given in the second column of Table 2.

Experiments: Tunisian Speech-to-Text

This section describes preliminary experiments with our training and evaluation datasets, which leads to the release of a first baseline model for Tunisian ASR.

First, we tested several large versions of Whisper, as shown in the first rows of Table 3. Even if Whisper can recognize some Arabic, it fails to transcribe the Tunisian Arabic Dialect, with Word Error Rates (WER) ranging from 50% to more than 100% (indicating high insertion rates).
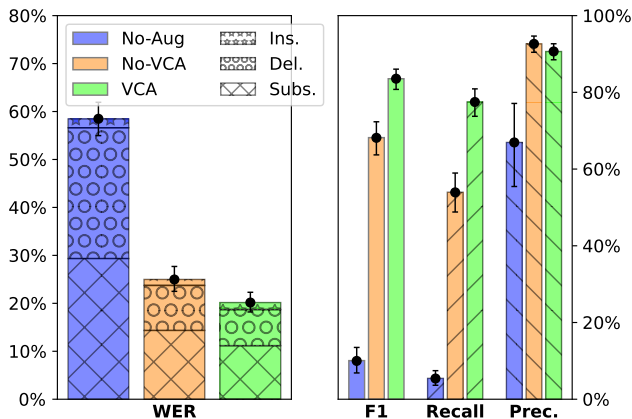

Figure 1: Details of results across three training conditions on TunSwitch Code-Switching (CS) test set. *left:* Word Error Rates (WER) decomposed into insertion (Ins), deletion (Del) and substitution (Subs) rates. *right:* F1, Recall and Precision scores on Latin words. All 95% confidence intervals are computed by performing bootstrap resampling.

To provide a baseline for Tunisian ASR, we trained models using the Kaldi open-source toolkit (Povey et al. 2011), where a triphone GMM-HMM model is trained to align target phonemes with an audio signal. The acoustic model consists of a Time Delay Neural Network (TDNN) (Ped-

dinti, Povey, and Khudanpur 2015), and a Finite-State Transducer (FST) converts sequences of phoneme probabilities into most probable words. The underlying LM is a word n-gram (with n up to 4) trained on the LinTO textual dataset. Training details are available under a GPL license.[9]

There are no good phonetic dictionaries for Arabic dialects, and phoneme-based ASR has failed to outperform treating each Tunisian Arabic character as a phoneme (Ali et al. 2008). We used Kaldi with the Buckwalter transliteration (Habash, Soudi, and Buckwalter 2007), which we extended to encode Latin characters while using only ASCII character pairs to represent Latin and Arabic characters.

Table 3 compares several training dataset mix conditions for LinTO models trained on our data:
- `No-Aug`: trained on 93H of original audio data without any augmentation for 20 epochs;
- `No-VCA`: trained on 4 epochs on $5 \times 93H$ of training data augmented with classical techniques (speed, volume, ...).
- `VCA`'s training data also includes VC augmented data, with a total duration of $5 \times 466H$.

`VCA` has a notable impact on code-switching performance. Figure 1 compares on TunSwitch the models trained on the 3 data mixes. To assess accuracy for recognition of English and French phrases in a code-switching context, we give F1, recall and precision scores on Latin words. We can see that `VCA` significantly improves WER, and in particular the recall on Latin words. These results suggest that `VCA` enhances the model's ability to model rarer phenomena, such as English words mixed with Arabic words.

Conclusion

Our work advances Automatic Speech Recognition (ASR) by releasing public datasets containing audio and textual data in Tunisian. These datasets can be used to train and evaluate ASR systems. Through our experiments, we demonstrate the feasibility of training a high-quality ASR model using traditional methods on this dataset. We establish the first baseline ASR model for the Tunisian Arabic Dialect, highlighting its capabilities, including initial support for code-switching with French and English. This baseline is trained exclusively from scratch on our open datasets. Future work involves fine-tuning models like Whisper and conformers on the open Arabic spoken data we have released.

---

[9]https://github.com/linagora-labs/ASR_train_kaldi_tunisian

References

Abdallah, A. A. B.; Kabboudi, A.; Kanoun, A.; and Zaiem, S. 2023. Leveraging Data Collection and Unsupervised Learning for Code-switched Tunisian Arabic Automatic Speech Recognition. arXiv:2309.11327.

Al-Fetyani, M.; Al-Barham, M.; Abandah, G.; Alsharkawi, A.; and Dawas, M. 2021. MASC: Massive Arabic Speech Corpus.

Ali, M.; Elshafei, M.; Al-Ghamdi, M.; Al-Muhtaseb, H.; and Al-Najjar, A. 2008. Generation of arabic phonetic dictionaries for speech recognition. In *2008 International Conference on Innovations in Information Technology*, 59–63.

Baevski, A.; Zhou, H.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. arXiv:2006.11477.

Gayraud, F.; Barkat-Defradas, M.; Lahrouchi, M.; and Ben Hamed, M. 2018. Development of phonetic complexity in Arabic, Berber, English and French. *Canadian Journal of Linguistics / Revue canadienne de linguistique*, 63(4): 527–555.

Gibson, M. L. 1999. *Dialect contact in Tunisian Arabic: sociolinguistic and structural aspects.* Ph.D. thesis, University of Reading.

Grosman, J. 2021. Fine-tuned XLSR-53 large model for speech recognition in Arabic. https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-arabic.

Habash, N.; Soudi, A.; and Buckwalter, T. 2007. *On Arabic Transliteration*, 15–22. Dordrecht: Springer Netherlands. ISBN 978-1-4020-6046-5.

Hennequin, R.; Khlif, A.; Voituret, F.; and Moussallam, M. 2020. Spleeter: a fast and efficient music source separation tool with pre-trained models. *Journal of Open Source Software*, 5(50): 2154. Deezer Research.

Hsu, W.-N.; Bolte, B.; Tsai, Y.-H. H.; Lakhotia, K.; Salakhutdinov, R.; and Mohamed, A. 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. arXiv:2106.07447.

Kim, J.; Kong, J.; and Son, J. 2021. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech. *CoRR*, abs/2106.06103.

Kong, J.; Kim, J.; and Bae, J. 2020. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. arXiv:2010.05646.

Mdhaffar, S.; Bougares, F.; de Mori, R.; Zaiem, S.; Ravanelli, M.; and Estève, Y. 2024. TARIC-SLU: A Tunisian Benchmark Dataset for Spoken Language Understanding. In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 15606–15616. Torino, Italia: ELRA and ICCL.

Ouerhani, B. 2009. Interférence entre le dialectal et le littéral en Tunisie: Le cas de la morphologie verbale. *Synergies Tunisie n*, 1: 75–84.

Peddinti, V.; Povey, D.; and Khudanpur, S. 2015. A time delay neural network architecture for efficient modeling of long temporal contexts. In *Interspeech*.

Povey, D.; Ghoshal, A.; Boulianne, G.; Burget, L.; Glembek, O.; Goel, N.; Hannemann, M.; Motlíček, P.; Qian, Y.; Schwarz, P.; Silovský, J.; Stemmer, G.; and Vesel, K. 2011. The Kaldi speech recognition toolkit. *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. arXiv:2212.04356.

Saidi, D. 2007. Typology of motion event in Tunisian Arabic. *Proceedings of LingO*, 196: 23.

Talafha, B.; Abuammar, A.; and Al-Ayyoub, M. 2021. Atar: Attention-based LSTM for Arabizi transliteration. *International Journal of Electrical and Computer Engineering*, 11: 2327–2334.

Tilmatine, M. 1999. Substrat et convergences: le berbère et l'arabe nord-africain. *EDNA, Estudios de dialectología norteafricana y andalusí*, 4: 99–119.

Tjandra, A.; Singhal, N.; Zhang, D.; Kalinli, O.; Mohamed, A.; Le, D.; and Seltzer, M. L. 2022. Massively Multilingual ASR on 70 Languages: Tokenization, Architecture, and Generalization Capabilities. arXiv:2211.05756.

van Niekerk, B.; Carbonneau, M.-A.; Zaidi, J.; Baas, M.; Seute, H.; and Kamper, H. 2022. A Comparison of Discrete and Soft Speech Units for Improved Voice Conversion. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

voicepaw et al. 2024. SoftVC VITS Singing Voice Conversion Fork. Accessed: YYYY-MM-DD.

Zhu, J.; and Zhang, C. 2021. Performing forced alignment with Wav2vec 2.0. *The Journal of the Acoustical Society of America*, 150: A357–A357.

Zribi, I.; Ellouze, M.; Belguith, L. H.; and Blache, P. 2017. Morphological disambiguation of Tunisian dialect. *Journal of King Saud University - Computer and Information Sciences*, 29(2): 147–155. Arabic Natural Language Processing: Models, Systems and Applications.