# A Survey of Code-switched Arabic NLP: Progress, Challenges, and Future Directions

**Injy Hamed,**[λ] **Caroline Sabty,**[ξ] **Slim Abdennadher,**[ξ]
**Ngoc Thang Vu,**[σ] **Thamar Solorio,**[λδ] **Nizar Habash**[λμ]
[λ]MBZUAI [ξ]German International University, Egypt [σ]University of Stuttgart
[δ]University of Houston [μ]New York University Abu Dhabi
{injy.hamed,thamar.solorio}@mbzuai.ac.ae
{caroline.sabty,slim.abdennadher}@giu-uni.de
thang.vu@ims.uni-stuttgart.de nizar.habash@nyu.edu

## Abstract

Language in the Arab world presents a complex diglossic and multilingual setting, involving the use of Modern Standard Arabic, various dialects and sub-dialects, as well as multiple European languages. This diverse linguistic landscape has given rise to code-switching, both within Arabic varieties and between Arabic and foreign languages. The widespread occurrence of code-switching across the region makes it vital to address these linguistic needs when developing language technologies. In this paper, we provide a review of the current literature in the field of code-switched Arabic NLP, offering a broad perspective on ongoing efforts, challenges, research gaps, and recommendations for future research directions.

## 1 Introduction

Code-switching (CSW), the act of alternating between multiple languages in the same discourse, is a common linguistic phenomenon in multilingual societies, including Arab countries. With CSW's global prevalence, it has become essential to equip language technologies to effectively handle it to build inclusive and user-friendly tools that cater to the needs of multilingual communities. While there exist several survey papers on CSW (Çetinoğlu et al., 2016; Sitaram et al., 2019; Doğruöz et al., 2021; Winata et al., 2023), this paper narrows the focus to the Arabic language, offering more in-depth knowledge and insights for this specific language setup. We provide this review, discussing current literature, challenges, and research gaps, with the aim of guiding future research and accelerating progress in this area.

The paper is organized as follows: §2 provides an overview on the linguistic landscape in the Arab world and historical factors giving rise to CSW; §3 discusses the types of CSW; §4 describes the paper categorization process; §5 presents overall statistics on the current literature in CSW Arabic NLP; §6 and §7 give further insights into the efforts on data collection, modeling, and guidelines and annotation tools that are useful in the context of CSW; and §8 suggests future research directions.

## 2 Language in the Arab World

Historical events have long influenced the patterns of CSW between Arabic and other languages. During the 7th and 8th centuries, the Islamic conquests brought Arabic into extensive contact with several languages, including Persian (Khan et al., 2011) and Spanish (Thomas and Sayahi, 2019), leading to profound linguistic exchanges. Later, from the 16th until the early 20th century, Turkish and Arabic intermingled during the Ottoman Empire, leaving lasting imprints on both languages. Afterwards, during the 19th and 20th centuries, several Arab countries were impacted by the spread of European colonialism across the region. According to Duranti (2008) and Cotterell et al. (2014), colonization played a main role in shaping the language in the region, where the local languages were overlain by European languages. These events and cultural interactions led to linguistic exchanges, giving rise to CSW between Arabic and numerous languages. In contemporary times, globalization and international businesses and education have further intensified this phenomenon, with Arabic speakers increasingly using English and other languages in social and professional settings.

Besides code-switching to foreign languages, with Arabic being a diglossic language (Ferguson, 1959), Arabs also switch between formal Arabic and dialects. The formal Arabic variant, Modern Standard Arabic (MSA), serves as a lingua franca across Arab countries and is typically used in formal contexts. Dialects (and sub-dialects), belonging to each country, are used in everyday conversations and informal writings. This gives rise to two main types of CSW in the region: (1) CSW



*Proceedings of the 31st International Conference on Computational Linguistics*, pages 4561–4585
January 19–24, 2025. 

| CSW Type | Example |
|---|---|
| **Inter-sentential** | أنا عندي إمتحان قريب. I need to prepare for it.<br>*I have an exam coming up.* I need to prepare for it. |
| **Extra-sentential** | Okay هاشتغل على المشروع بكرا .<br>Okay *I'll work on the project tomorrow.* |
| **Intra-sentential** | المشروع بتاعي is in the field of computer vision<br>*My project* is in the field of computer vision. |
| **Morphological CSW** | لازم ن+evaluate ال+models النهارده before the deadline<br>*We have to* evaluate *the* models *today* before the deadline. |

Table 1: Examples of different CSW types followed by their English translation. The originally Arabic phrases are italicized in the English translation.

between MSA and dialects, and (2) CSW between Arabic and foreign languages. Following Adouane et al. (2018a), we refer to the former type as *diglossic code-switching* and the latter as *bilingual code-switching*. Bilingual CSW is seen across the region, including Arabic-English in Egypt (Abu-Melhim, 1991), Jordan (Mustafa and Al-Khatib, 1994), Palestine (Mkahal, 2016), Saudi Arabia (Omar and Ilyas, 2018), and UAE (Khuwaileh, 2003). A high level of multilingualism, with the mixing of Arabic, English, and French, is found in Morocco (Samih and Maier, 2016a), Algeria (Baya Essayahi and Kerras, 2016), Lebanon (Bacha and Bahous, 2011), and Tunisia (Baoueb, 2009).

While Arabic CSW occurs with a wide range of languages, we restrict the scope of this paper to those so far addressed within the field of computational linguistics. We also restrict our scope to CSW language alternation, excluding studies looking into the origin of words in languages influenced by others, such as the investigation presented in Micallef et al. (2024) for Maltese.

## 3 Types of Code-switching

According to Poplack (1980), there are three main types of CSW: inter-sentential, extra-sentential, and intra-sentential. In the case of morphologically rich languages, morphological code-switching (Stefanich et al., 2019) also occurs. We elaborate on each type below, providing examples in Table 1:

- **Inter-sentential CSW** involves switching of languages on the sentence-level.
- **Extra-sentential CSW** (or **tag-switching**) involves using tag elements from another language such as fillers, interjections, tags, and idiomatic expressions.
- **Intra-sentential CSW** involves word-level switching, where CSW segments must conform to the syntactic rules of both languages.
- **Morphological CSW** (or **intra-word CSW**) involves switching on the morpheme-level. Given that Arabic is a morphologically rich language, Arabic-speakers attach Arabic clitics and affixes to foreign words.

Another type of language alternation is referred to as **borrowing**, where loanwords are embedded into sentences without the need for grammatical considerations. According to Poplack (1980), borrowing is not considered as a type of CSW. Myers-Scotton (1997), however, considers any embedded segment from the secondary language as a case of CSW. Given that automatically identifying whether an embedded word is a case of borrowing or CSW is a challenging task, the majority of the surveyed efforts in this paper consider any presence of foreign tokens as CSW. Within the scope of the paper, CSW is used to refer to all the mentioned types of language alternation, including borrowing.

## 4 Paper Categorization Process

We conduct our search for relevant papers on Google Scholar using keywords that involve *code-switch*, *code-mix*, and *Arabic*. Afterwards, we categorize the collected papers, where we take inspiration from the guidelines presented in Winata et al. (2023). We annotate the papers for multiple categories, as shown in Table 2, including the year, venue, and language pairs. For the language pairs, we use five categories specifying CSW between MSA, dialectal Arabic (DA), and foreign language(s): MSA-DA, MSA-Foreign, DA-Foreign, Arabic-Foreign, MSA-DA-Foreign. 'Arabic-Foreign' is used when the variant of Arabic (MSA or DA) is not explicitly stated. Empirical papers are annotated for the NLP tasks and methodological approaches. Resource papers are annotation for the domain of collected data and the NLP tasks that they support.

| Category | Options |
|---|---|
| **Language Pairs** | MSA-DA, MSA-Foreign, DA-Foreign, Arabic-Foreign, MSA-DA-Foreign |
| **Venues** | Conference, Workshop, Symposium, Book, Journal, ArXiv, Thesis |
| **Methods** | Rule/Linguistic Constraint, Statistical Model, Neural Network, Pre-trained Model |
| **NLP Tasks** | **Text:** Word-level Language Identification, Sentence-level Language Identification, CSW Point Prediction, Dialectness Level Estimation, Dependency Parsing, Fake News Detection, Humor Detection, Humor Generation, Emotion Detection, Language Modeling, Lemmatization, Machine Translation, Micro-Dialect Identification, Named Entity Recognition, Natural Language Entailment, Natural Language Understanding, Part-of-Speech Tagging, Question Answering, Sarcasm Detection, Sentiment Analysis, Semantic Parsing, Spelling Correction and Text Normalization, Summarization, CSW Text Generation, Tokenization, Topic Modeling, Transliteration, Word Analogy, Abusive Language Detection<br>**Speech:** Word-level Language Identification, Sentence-level Language Identification, Automatic Speech Recognition, Speech Synthesis, Speech Translation, Sentence Boundary Detection, Text-to-Speech |

Table 2: The categories used in the annotation process.

## 5 Overview on Literature Review

### 5.1 Number of Papers Across the Years

While CSW has been thoroughly studied by linguists since 1980s (Poplack, 1980), it started receiving considerable attention from the Arabic NLP community in 2014. The work on CSW has been greatly motivated by the *Workshop on Computational Approaches to Linguistic Code-Switching*, including shared tasks (ST) which took place in the years of 2014, 2016, 2018, and 2021, reflected in the peaks shown in Figure 1. In general, on average, there are 12 papers per year since 2014, reflecting the need for increased attention in this area.

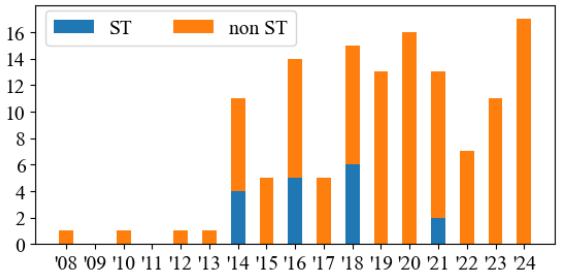


Figure 1: Number of Shared Task (ST) and non-Shared Task (non-ST) papers.

### 5.2 Distribution of Papers Across Venue Types

The majority of the papers are published at conferences (50%), followed by workshops (36%, with 13% absolute belonging to shared tasks), followed by journals (7%), arXiv papers (4%) that have not yet been published in venues, and theses (3%).

### 5.3 Evolution in Methodology Methods

We show the evolution in methodology methods in CSW Arabic NLP by plotting the distribution of methods across the years in Figure 2. Comparing these timelines to those presented in Winata et al. (2023) for CSW overall work, we observe a 3-4 year delay in adopting new methodologies. This is observed for neural-based approaches, which are applied in CSW Arabic NLP in 2016, three years after their utilization in overall work on CSW. Similarly, the use of pretrained models in CSW Arabic NLP started in 2020, compared to 2016 for CSW overall. Despite this delay, we observe alignment with the global NLP trends, showing interest in the Arabic community in advancing CSW research.

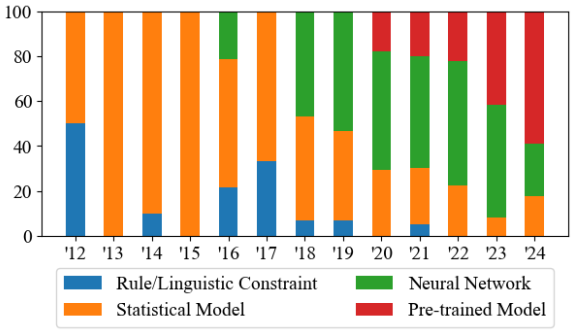


Figure 2: Distribution of papers based on the methods.

### 5.4 Investigated Language Pairs

We analyze the papers by language pair categories following our annotation guidelines, specifying Arabic variants and foreign languages when indicated. Figure 3 shows these distributions. We observe that current research is primarily focused on CSW DA-foreign and MSA-DA, which is inline with their prevalent use in real-life. For diglossic CSW, MSA-Egyptian is the most studied language pair, while Egyptian-English and Algerian-French are leading for bilingual CSW. Several studies cover more than three languages, however, their count might be exaggerated, as some papers lack sentence-level CSW statistics, making it infeasible to understand the extent of mixing.

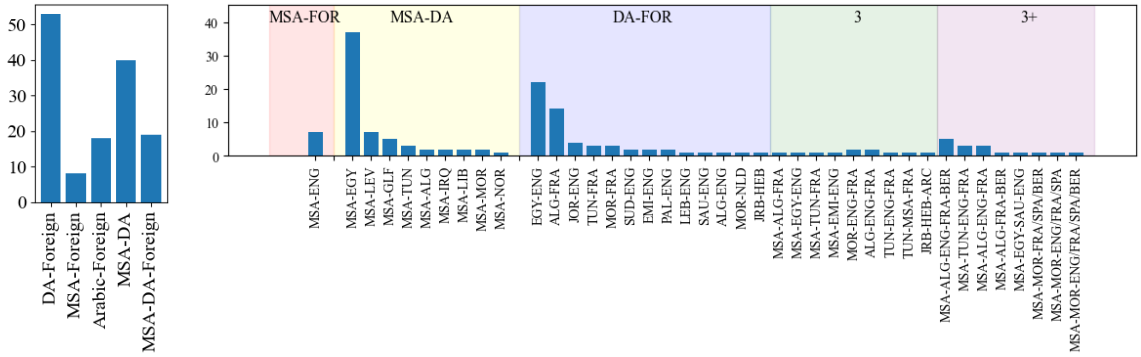


Figure 3: The number of papers covering the different CSW language setups. We present the distribution as per the annotation guidelines (left) with further language specification (right). Language codes are provided in Appendix A.

## 6 Insights on Language Resources

**Dataset Sources** We report the following distribution of datasets' sources: *social media* (36), *transcriptions* (21), *speech recordings* (18), *internet forms and blogs* (11), *news* (8), dialogue (6), *songs* (3), *government documents* (2), and *books* (1). The domination of social media is expected as it offers a non-expensive source of text, where CSW is likely to occur. However, it is to be noted that CSW phenomena occurring in text could be more restricted than in natural speech (Hamed et al., 2022c). For example, in the case of Arabic bilingual CSW, given that Arabic and foreign languages use different scripts, users may be dissuaded to code-switch, especially in the case of intra-word CSW. Users may opt for romanization, known as Arabizi (Bies et al., 2014), where Arabic words are represented by Latin letters and numerals. However, users have different attitudes and preferences regarding the use of Arabizi (Alsulami, 2019), where initiatives, such as *BilArabi* (meaning "*in Arabic*"), have even been established to promote the use of MSA over social media as part of efforts in preserving the Arabic language (Taha Thomure, 2019).

**NLP Tasks' Coverage** In Table 3, we present statistics on the NLP tasks supported by the collected datasets. The complete list of papers is provided in Appendix C. We note that the task of language modeling is supported by all collected textual corpora. Otherwise, only *word-level LID* contains a reasonable number of corpora (22). This is followed by *ASR* (17), *MT* (7), *transliteration* (7), and *sentiment analysis* (6). The remaining tasks are not well supported, highlighting a significant gap in resources. Further efforts are needed to extend existing datasets, offering greater diversity in represented sources, dialects, and tasks.

| Task | E | R |
|---|---|---|
| Text: Word-level Language Identification (LID) | 32 | 22 |
| Text: Named Entity Recognition (NER) | 15 | 3 |
| Text: Machine Translation (MT) | 13 | 7 |
| Text: Sentiment Analysis | 10 | 6 |
| Text: Part-of-Speech Tagging | 6 | 3 |
| Text: Sentence-level Language Identification | 5 | 3 |
| Text: Transliteration | 5 | 7 |
| Text: Language Modeling | 4 | 49 |
| Text: Abusive Language Detection | 2 | 3 |
| Text: Sentence-level Micro-Dialect Identification | 2 | 2 |
| Text: Spelling Correction and Text Normalization | 2 | 1 |
| Text: Dependency Parsing | 2 | 1 |
| Text: Tokenization | 1 | 3 |
| Text: Fake News Detection | 1 | 1 |
| Text: Word Analogy | 1 | 1 |
| Text: Topic Modeling | 1 | 1 |
| Text: Question Answering | 1 | 0 |
| Text: Lemmatization | 0 | 1 |
| Text: Dialectness Level Estimation | 0 | 2 |
| Text: Emotion Detection | 0 | 1 |
| Speech: Automatic Speech Recognition (ASR) | 22 | 17 |
| Speech: Speech Translation | 2 | 1 |
| Speech: Word-level Language Identification | 1 | 4 |
| Speech: Sentence Boundary Detection | 1 | 11 |
| Speech: Sentence-level Language Identification | 0 | 1 |

Table 3: The coverage of NLP tasks across empirical (E) and resource (R) papers.

## 7 Insights on NLP Tasks

In Table 3, we present the distribution of empirical papers across tasks (complete list in Appendix B). The most studied tasks are *word-level LID in text*, *ASR*, *NER*, and *MT*. We believe the distribution is greatly affected by Shared Tasks (Solorio et al., 2014; Molina et al., 2016; Aguilar et al., 2018; Chen et al., 2021) and PhD theses (Elfardy, 2017; Samih, 2017; Amazouz, 2019; Adouane, 2020; Al-Ghamdi, 2020; Sabty, 2024; Hamed, 2024). In this Section, we provide insights into research efforts and challenges in a selection of the NLP tasks.

### 7.1 Language Identification (LID)

**Word-level LID** Word-level LID is a sequence-to-sequence classification task, where the goal is to assign a language label to each word. Only few researchers have tackled this task in the speech domain, covering CSW Algerian-French (Amazouz et al., 2017) and MSA-DA (Chowdhury et al., 2020). The remaining efforts were conducted in the text domain, with social media platforms being the prominent source of text. In the text-domain, the majority of empirical papers covered MSA-DA (19/32), followed by romanized Arabic-foreign (9/32) and Arabic-Arabized foreign (6/32) CSW.

One challenge in this task is the presence of words sharing lexical forms across languages, that are cognates/faux-amis, having similar/different meanings. This issue is more pronounced in diglossic CSW due to shared vocabulary between MSA and DA. This introduces ambiguity, especially with the lack of acoustic cues, challenging automatic as well as manual annotations. These challenges were discussed in Solorio et al. (2014) and Aguilar et al. (2020), where MSA-DA was found to be the hardest language pair across the explored languages, including Nepali-English, Mandarin-English, Spanish English, and Hindi-English. Another challenge that arises, especially in the case of bilingual CSW, is distinguishing between borrowings and CSW. Given the complexity of such a distinction, many researchers do not address this, where the decision is left to the annotators' judgment. Alternatively, a language label for borrowed words can be used as proposed in Adouane and Dobnik (2017). Morphological CSW also adds another layer of complexity. While most researchers have handled it using a 'mixed' language label, Sabty et al. (2021a) further expand this label specifying the languages on the morpheme-level in morphological CSW words.

**Sentence-level LID** For sentence-level LID, we discuss two tasks. The first involves identifying whether a sentence includes CSW, which was explored for MSA-DA (Altamimi, 2020) and romanized Arabic-foreign (Shehadi and Wintner, 2022).

The second task is *dialect identification*, which involves identifying the Arabic variant of a sentence. This task has been well studied, where it has mostly been defined as a single-label classification task (Bouamor et al., 2018; Abdelali et al., 2021; Abdul-Mageed et al., 2023). Lately, Abdul-Mageed et al. (2024) modified the task definition into multi-label classification, overcoming the limitations of single labels, including sentences being acceptable under several Arabic variants (Keleg and Magdy, 2023). In the scope of CSW, the task has been investigated in the contexts of diglossic (Al-Badrashiny et al., 2015; El-Haj et al., 2018) and bilingual (Abdul-Mageed et al., 2020; Abainia, 2018) CSW, where challenges are introduced due to the presence of foreign words and the vocabulary overlap between MSA and DA. We note, however, that the current task definition falls short in handling CSW, and we recommend refining it to identify mixed variants within a sentence.

### 7.2 Dialectness Level Identification

This task has been defined as both a regression (Keleg et al., 2023) and classification (Abdul-Mageed et al., 2024) problem. However, both efforts do not take CSW into account. Efforts involving Arabic CSW are limited to guidelines' development and data annotation. Badawi (1973) defines five levels for Egyptian Arabic based on the use of MSA, dialect, as well as foreign languages, reflecting on the educational levels of the participants of each group. Habash et al. (2008) accounts for diglossic CSW when defining the following five levels of dialectness: perfect MSA, imperfect MSA, MSA-dialectal CSW, dialect with MSA incursions, and pure dialect. The guidelines were adopted by Hamed et al. (2024), where transcriptions having a mixture of MSA, Gulf and Egyptian Arabic, and English were annotated accordingly.

### 7.3 Transliteration

Transliterating romanized Arabic words into Arabic script is a challenging task as romanization is non-standard and unlike HSB transliteration scheme (Habash et al., 2007), does not have a direct one-to-one mapping to Arabic words. When compounded with CSW, another layer of complexity is added as foreign words should remain unchanged. Previous approaches (Darwish, 2014; Eskander et al., 2014) involved a two-step process of Arabic/foreign word-level LID followed by script conversion. Shazal et al. (2020) presented the first effort for a unified model for both; Arabizi detection and transliteration. In terms of the explored language pairs, the main focus of research has been on romanized Arabic-English, where we still lack research on Arabic-Arabized foreign. Little exploration has also been conducted on transliterating Judeo-Arabic with code-switched Hebrew and Aramaic into Arabic (Mitelman et al., 2024).

### 7.4 Machine Translation (MT)

Despite CSW Arabic MT being only addressed in 13 papers, efforts cover a broad range of dialects and translation directions. Chen et al. (2021) introduced a shared task, covering two-way translation between CSW MSA-Egyptian and English. Research on translating from CSW Arabic-foreign to the foreign language has been conducted for several language pairs, covering MSA-English, Egyptian-English, Jordanian-English, Palestinian-English, Algerian-French, Moroccan-French, and Tunisian-French, as outlined in Appendix B. Less research has focused on translating CSW Arabic-foreign to Arabic as well as both directions (Hamed et al., 2022c; Heakl et al., 2024). Translating to the primary language is particularly useful, as it could help bilingual speakers bridge language gaps, given that bilinguals are driven towards CSW in cases of lack of language proficiency (Heredia and Altarriba, 2001). Another interesting translation direction is translating from monolingual to CSW, which has been slightly explored in the context of CSW data augmentation (Hamed et al., 2023b) and ChatGPT evaluation (Khondaker et al., 2023).

### 7.5 Automatic Speech Recognition (ASR)

CSW presents challenges to ASR on all three fronts; data collection, modeling and evaluation. In terms of *data collection*, given the costly and tedious process of building CSW speech corpora, they are usually scarce. CSW also usually occurs in only a portion of naturally-occurring speech, further limiting the amount of collected data. Challenges in *modeling* include the scarcity/lack of CSW corpora and its imbalance with the vaster amounts of MSA and dialectal Arabic data that are usually used to boost performance, in addition to the mixed acoustic dynamics that can arise with using different languages (Hamed et al., 2020; Mustafa et al., 2022). In terms of methodologies, both hybrid (Elfahal et al., 2020; Chowdhury et al., 2020; Ali et al., 2021; Hussein et al., 2023) and end-to-end (Mubarak et al., 2021; Chowdhury et al., 2021; Hamed et al., 2022c, 2023a; Hussein et al., 2024) ASR architectures have been investigated. Both architectures were compared in Hamed et al. (2022a), where comparable, yet complementary, performance was reported. Pretrained models have mostly been used for benchmarking purposes (Heakl et al., 2024; Al Ali and Aldarmaki, 2024; Abdelali et al., 2024), where more research is needed to explore how to optimally use these models to achieve further advancements, as conducted in Kulkarni et al. (2023). In terms of *evaluation*, limitations of WER and CER have been discussed (Chowdhury et al., 2021; Hamed et al., 2022a; Abdallah et al., 2024). Researchers have utilized transliteration as well as phonological and semantic similarities to alleviate cross-transcription issue, improving correlations with human judgments (Hamed et al., 2023c; Kadaoui et al., 2024).

### 7.6 Guidelines and Annotation Tools

Guidelines have been developed for CSW Arabic data collection for several tasks, including transcriptions (Hamed et al., 2020, 2024), ASR minimal post-editing (Hamed et al., 2023c), translations (Hamed et al., 2022c), annotations for word-level LID in text (Samih and Maier, 2016b; Diab et al., 2019) and speech (Chowdhury et al., 2020) as well as dialectness level identification (Badawi, 1973; Habash et al., 2008). With regards to useful tools, several tools have been developed for facilitating data collection (Benajiba and Diab, 2010; Voss et al., 2014; Al-Shargi and Rambow, 2015; Diab et al., 2019; Elwy and Sabty, 2024).

## 8 Research Gaps and Future Directions

This survey highlights a clear need to expand research in CSW Arabic NLP to address a wider coverage of languages and tasks. Efforts are needed across data collection, modeling, and evaluation, ultimately supporting the development of more inclusive language technologies that can effectively handle the linguistic diversity within the Arab world.

**Building CSW Arabic Benchmarks** Benchmarks are essential for standardizing evaluation, tracking progress, and guiding research. However, CSW Arabic remains underrepresented in current benchmarks. In the scope of CSW, the two main benchmarks are LinCE (Aguilar et al., 2020) and GLUECoS (Khanuja et al., 2020). Arabic is only represented in LinCE, covering diglossic CSW for LID and NER tasks. In the scope of Arabic NLP, Nagoudi et al. (2023) and Abdelali et al. (2024) offer benchmarks that cover a variety of dialects and tasks, however, CSW evaluations only cover ASR and MT tasks. The existing benchmarks are insufficient to fully reflect the current performance of models, highlighting the need for broader benchmarks that encompass a wider range of language pairs and tasks to better guide future research.

**Understanding and Improving the Capabilities of Pretrained Models for Arabic** As large language models (LLMs) and large speech models gain popularity, understanding their capabilities for CSW Arabic is crucial. Researchers evaluated pretrained models for CSW languages (Winata et al., 2021) and Arabic NLP (Khondaker et al., 2023; Nagoudi et al., 2023; Abdelali et al., 2024; Heakl et al., 2024; Al Ali and Aldarmaki, 2024). However, CSW Arabic is only minimally addressed, where the limited evaluations reveal varying performance across tasks and language pairs. In ASR, word error rates reported on CSW speech corpora range from 28% to 54% (Abdelali et al., 2024; Heakl et al., 2024; Kadaoui et al., 2024; Lachemat et al., 2024). In MT, LLMs' performance varies across different corpora having different language pairs and translation directions, with reported BLEU scores ranging from 38 to 87 (Khondaker et al., 2023; Heakl et al., 2024). Despite these models surpassing state-of-the-art models (Abdelali et al., 2024; Heakl et al., 2024), there remains substantial room for improvement. Additionally, the ability of LLMs to generate CSW Arabic text has not yet been thoroughly evaluated. Such an evaluation was conducted for South East Asian languages where varying performance was reported across languages (Yong et al., 2023). For Arabic, Khondaker et al. (2023) prompted ChatGPT to convert English sentences to CSW Arabic-English. While the generated translations were found to be fluent and faithful to the original meaning, human annotators gave low scores for its CSW ability, defined as 'how accurately the translated text includes code-switching'. We highlight this as another area worthy of investigation.

**Building CSW User-facing Applications** As highlighted in Doğruöz et al. (2021), there is a need for developing applications that interact with users in CSW language. This is an important capability of NLP systems in order to build user-friendly and human-like language technologies. Such applications are still lacking for Arabic, as in other languages. One important aspect to consider here is the dynamic behaviour of CSW, where it is affected by multiple factors. In the Arab region, CSW behaviour is found to be affected by external factors (e.g. topic, communication channel, and participants' roles and relationships) as well as user-related factors, including demographics (e.g. age and education) and personality (e.g. extraversion and neuroticism traits) (Post, 2015; Alaiyed, 2018; Aljasir, 2020; Hamed et al., 2021, 2022b). Accordingly, applications such as chatbots should be able to identify the appropriate language settings across situations, ranging from monolingual MSA and DA to diglossic and bilingual CSW. Given that the factors affecting CSW may differ across countries, it is important to tackle this problem from the perspective of the Arabic language and culture.

**Generating Personalized CSW Text** The importance of accommodating the different linguistic styles of users in user-facing applications was emphasized in Bawa et al. (2020), where users' preferences regarding CSW chatbots were affected by their attitudes and enthusiasm towards CSW. However, despite CSW text generation receiving considerable attention across languages, research exploring the personalization aspect remains limited (Sengupta et al., 2023; Mondal et al., 2022). In the scope of Arabic, CSW data augmentation has been explored for improving performance on NER, MT, ASR, and speech translation tasks (Sabty et al., 2021b; Hamed et al., 2023a,b; Hussein et al., 2023, 2024), however, we still lack research on the personalization front. A major bottleneck in this pursuit is the lack of diverse datasets annotated with users' meta-data, involving diverse users' sociological and psychological profiles, as well as diverse external settings. Therefore, we encourage researchers to include users' meta-data as part of data collection. Also, to facilitate large-scale projects, unified guidelines are needed for collecting users' meta-data. These guidelines could benefit from the vast amount of research investigating the factors influencing CSW in the Arab world (Bentahila, 1983; Albirini, 2011; Hafez, 2015; Post, 2015; Ali and Mohamed, 2018; El Bolock et al., 2020; Alsamhan and Almutrafi, 2022).

**Exploring the Effectiveness of User-adaptive NLP Models** Previous research, focused on other languages, has demonstrated benefits of adapting NLP systems to users' CSW behaviour. In Vu et al. (2013), it was shown that clustering speakers based on their CSW attitude and adapting the language model accordingly can enhance language modeling and ASR performance. In Rallabandi et al. (2018), features extracted from acoustics were used to distinguish between different CSW styles, where style-specific language models showed reduction in perplexity. Given the limited but promising studies, it would be interesting to extend this line of research to improve CSW Arabic NLP models.

**CSW in Medical and Educational Domains** CSW serves various purposes in medical and educational domains, as investigated by a number of studies in healthcare and educational institutes in the Arab world (Almathkuri, 2016; Alkhudair, 2019; Alhamami, 2020; Alkhlaifat et al., 2020; Yang, 2021; Zaghlool and Altamimi, 2023; Dooly and Bakri, 2024). In the medical domain, CSW was reported to facilitate bridging lexical gaps, addressing sensitive topics, maintaining interpersonal relationships, and signaling power and hierarchical dynamics. It was reported, on the other hand, that the misuse of CSW may evoke negative feelings, including suspicion among patients or concealment of information and disrespect among colleagues. In the educational domain, reported CSW functions include facilitating language acquisition, supporting better and faster comprehension, increasing student-teacher interaction, and enlivening the class atmosphere. It is to be noted though that CSW behaviour is highly dynamic in these domains, affected by factors such as educational instruction language, expatriate populations (being high in some Arab countries such as the UAE), disparities in education levels and language proficiency across urban and rural cities, and disparities in socio-economic backgrounds across private and public sectors. All these functions and nuances of CSW need to be taken into account when developing language technologies to ensure they align with the functional and cultural needs of users.

**Evaluating Naturalness of Generated CSW data** As mentioned in Winata et al. (2023), automatically evaluating the quality of generated CSW data is an area that is still understudied. The task has only been tackled by Kodali et al. (2024) and Kuwanto et al. (2024) for non-Arabic language pairs. In the scope of Arabic, the complexity of naturalness evaluation for CSW Arabic-English was demonstrated in Hamed et al. (2023b), where human inter-annotator agreement only reached fair agreement on pairwise Cohen Kappa. This is due to CSW being a speaker-dependent behaviour (Vu et al., 2013), where annotators' judgments may be biased towards their own CSW styles and those surrounding them. In the scope of Arabic bilingual CSW, human annotations have been collected for naturalness evaluation (Hussein et al., 2023; Hamed et al., 2023a,b), however, only involving a small number of annotators. In future work, given the subjectivity of the task, we encourage diversifying the pool of annotators to minimize cultural bias, in line with recommendations from Hershcovich et al. (2022).

**Handling Morphological CSW** Morphological CSW adds further complexity to NLP tasks, where performance reductions are demonstrated in MT and ASR tasks for this type of CSW (Gaser et al., 2023; Hamed et al., 2022a). Given that morphological CSW typically occurs in a subset of the corpus samples, we recommend that researchers report evaluation results on it separately to fully assess the CSW capabilities of the models. In the context of CSW text generation, producing morphological CSW is challenging, as it cannot be generated through rule-based approaches. For example, in Egyptian Arabic, the suffix ات *At* 'feminine plural form' can be attached to 'event' (which has a masculine gender in Arabic), while cannot be attached to 'school' (which has a feminine gender in Arabic). This limits the applicability of certain text generation methods, such as lexical replacements. Back-translation, on the other hand, was found to be capable of generating sentences with correct morphological CSW (Hamed et al., 2023b). With the growing interest in CSW Arabic text generation, morphological CSW deserves more attention. Broader studies are needed to capture the diverse range of patterns varying between dialects, individuals, and domains (text versus speech), where linguistic studies on morphological CSW patterns can providing valuable support for this research direction (Farid, 2019; Kniaź and Zawrotna, 2021).

**Assessing Evaluation Metrics** Besides CSW introducing challenges in modeling, it also poses challenges in evaluation. This has been discussed in ASR for Arabic bilingual CSW (Hamed et al., 2023c), where the limitations of metrics relying on string-based matching are highlighted in the case of unstandardized orthography and cross-transcription (where words are transcribed and recognized in different scripts). The authors show that word and character error rates, despite being the widely-used metrics in ASR, are not adequate under these conditions. We still lack similar research assessing evaluation metrics for other tasks within CSW Arabic NLP. One foreseeable issue in bilingual CSW is the limitation of string-based matching metrics used in language generation tasks, such as transliteration and summarization, where models should not be penalized for different script choices in case of borrowed words.

**CSW Privacy and Ethical Considerations** There are important privacy and ethical concerns that need to be addressed when collecting CSW corpora. Given that the phenomenon occurs frequently in informal communication, primary sources for obtaining CSW data include speech, social media, and chat messages. This highlights the need for proper anonymization and privacy-preserving practices to protect users' identities. While researchers have worked on data anonymization in text (Zhou et al., 2008; Beigi and Liu, 2020; Sotolář et al., 2021) and speaker anonymization in speech (Meyer et al., 2022, 2023a,b, 2024) for monolingual data, there is a notable gap in addressing CSW contexts. We identify this as an interesting direction, where challenges introduced by CSW would need to be tackled, such as current limitations of text-to-speech systems in handling CSW with Arabic dialects. Furthermore, from a data collection stand-point, we advise researchers to obtain informed consent from participants on the public release of collected data, as a high percentage of non-public CSW data has been highlighted in Winata et al. (2023), which hinders research advancements in the field. Finally, we emphasize the importance of developing corpora covering a wide range of languages and dialects, including underrepresented CSW language setups, to avoid marginalization of minority languages, such as Saidi Arabic (upper Egyptian dialect), Fellahi Arabic (Levantine peasant dialects), Amazigh languages, Kurdish, and Shehri/Jibbali (a modern South Arabian language).

**Exploring Unexplored Tasks** As shown in Table 3, only half of the NLP tasks identified in Table 2 for annotation have been explored empirically. This leaves us with a significant number of unexplored tasks, including CSW point prediction (Solorio and Liu, 2008), natural language entailment, natural language understanding, question answering, humor detection and generation, sarcasm detection, semantic parsing, dialectness level estimation, summarization, speech synthesis, and text-to-speech. Moreover, among the tasks that have been covered, half are addressed in only 1-2 papers. In this large pool of understudied tasks, while word-level classification tasks are important, we recommend prioritizing high-level tasks such as question answering, text-to-speech, and speech translation, as these are essential for enhancing user-facing applications that require effective handling of CSW.

**Improving State-of-the-art CSW Arabic Models** We highlight that even for the well-investigated NLP tasks, there is still room for improvement. For word-level LID, the performance of models reached an F1 score of 91.9 for diglossic CSW (Attia et al., 2019) and 95.0 for bilingual CSW (Shehadi and Wintner, 2022). As for NER, the models reached an F1 score of 85.2 for diglossic CSW (Winata et al., 2021) and 79.4 for bilingual CSW (Sabty et al., 2020). For MT, a BLEU score of 87.2 (Heakl et al., 2024) is reported when translating from CSW sentences to Arabic. BLEU scores ranging from 23.1 to 53.6 are reported when translating to the foreign language across different CSW language pairs. For ASR, WER in the range of 24.8-53.8% is reported across different CSW Arabic-foreign speech corpora. For ST, BLEU scores of 31.1 and 17.0 are reported, translating to Arabic and the foreign language, respectively (Hamed et al., 2022c, 2023a). While comprehensive benchmarks are required, as previously discussed, to assess the current state-of-art, we acknowledge that further work is needed to improve the CSW capabilities of existing models, especially in the areas of text generation and speech processing.

## 9 Summary and Outlook

In this paper, we review current efforts in code-switched Arabic NLP. Arabic code-switching presents a wide range of challenges for the NLP community, that are compounded by the inherent complexities of Arabic itself, including diglossia, romanization, and unstandardized orthographies in dialects. We discuss existing literature, highlight research gaps and challenges, and identify directions for future work. We hope this paper guides researchers to further advance this research area, with the goal of developing language technologies that meet the linguistic needs of the Arab world.

## Limitations

We acknowledge that our survey is limited by a number of factors. First, the scope is confined to available literature, which may exclude emerging research and recent advancements. Second, the survey primarily focuses on major Arabic dialects and may not fully represent less-studied varieties or minority languages in the Arab World. Finally, the diverse linguistic and cultural contexts across the Arab world can influence code-switching patterns, which may not be uniformly covered.

## Ethics Statement

This paper did not involve human annotation collection or the creation of new datasets. We acknowledge that research in code-switched language processing, like many NLP tasks, could lead to the development of tools that might be misused for user profiling or generating non-standard language. Our focus is on advancing nuanced understanding of language use in real-world contexts to enhance applications such as speech recognition, machine translation, and text rewriting, while adhering to ethical guidelines and mitigating potential misuse.

## References


Kheireddine Abainia. 2018. Detecting Algerian subdialects of on-line commentators in social media networks. In *Proceedings of the Third International Conference on Pattern Analysis and Intelligent Systems (PAIS)*, pages 1–8.

Kheireddine Abainia. 2019. DZDC12: a new multipurpose parallel Algerian Arabizi–French code-switched corpus. *Language Resources and Evaluation*, pages 1–37.

Ahmed Amine Ben Abdallah, Ata Kabboudi, Amir Kanoun, and Salah Zaiem. 2024. Leveraging data collection and unsupervised learning for code-switched Tunisian Arabic automatic speech recognition. In *Proceedings of ICASSP*, pages 12607–12611.

Ahmed Abdelali, Hamdy Mubarak, Shammur Absar Chowdhury, Maram Hasanain, Basel Mousi, Sabri Boughorbel, Samir Abdaljalil, Yassine El Kheir, Daniel Izham, Fahim Dalvi, et al. 2024. LAraBench: Benchmarking Arabic AI with large language models. In *Proceedings of EACL*, pages 487–520.

Ahmed Abdelali, Hamdy Mubarak, Younes Samih, Sabit Hassan, and Kareem Darwish. 2021. QADI: Arabic dialect identification in the wild. In *Proceedings of the sixth Arabic natural language processing workshop*, pages 1–10.

Hocini Abdelouahab and Kamel Smaïli. 2024. Detecting fake news: Exploring key features in multilingual Arabic dialect corpus. In *Proccedings of the International Conference on Arabic Language Processing*.

Muhammad Abdul-Mageed, Abdelrahim Elmadany, Chiyu Zhang, Houda Bouamor, Nizar Habash, et al. 2023. NADI 2023: The fourth nuanced Arabic dialect identification shared task. In *Proceedings of the First Arabic Natural Language Processing Conference*, pages 600–613.

Muhammad Abdul-Mageed, Amr Keleg, AbdelRahim Elmadany, Chiyu Zhang, Injy Hamed, Walid Magdy, Houda Bouamor, and Nizar Habash. 2024. NADI 2024: The fifth nuanced Arabic dialect identification shared task. In *Proceedings of the Second Arabic Natural Language Processing Conference*.

Muhammad Abdul-Mageed, Chiyu Zhang, Abdelrahim Elmadany, and Lyle Ungar. 2020. Toward micro-dialect identification in diaglossic and code-switched environments. In *Proceedings of EMNLP*, pages 5855–5876.

Karima Abidi and Kamel Smaïli. 2017. An empirical study of the Algerian dialect of social network. In *Proceedings of the International Conference on Natural Language, Signal and Speech Processing (ICNLSSP)*.

Abdel-Rahman Abu-Melhim. 1991. Code-switching and linguistic accommodation in Arabic. In *Perspectives on Arabic Linguistics III: Papers from the Third Annual Symposium on Arabic Linguistics*, volume 80, pages 231–250. John Benjamins Publishing.

Wafia Adouane. 2020. *Natural Language Processing for Low-resourced Code-switched Colloquial Languages–The Case of Algerian Language*. Phd thesis, University of Gothenburg.

Wafia Adouane and Jean-Philippe Bernardy. 2020. When is multi-task learning beneficial for low-resource noisy code-switched user-generated Algerian texts? In *Proceedings of the Fourth Workshop on Computational Approaches to Code Switching*, pages 17–25.

Wafia Adouane, Jean-Philippe Bernardy, and Simon Dobnik. 2018a. Improving neural network performance by injecting background knowledge: Detecting code-switching and borrowing in Algerian texts. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 20–28.

Wafia Adouane, Jean-Philippe Bernardy, and Simon Dobnik. 2019. Normalising non-standardised orthography in Algerian code-switched user-generated data. In *Proceedings of the Fifth Workshop on Noisy User-generated Text*, pages 131–140.

Wafia Adouane and Simon Dobnik. 2017. Identification of languages in Algerian Arabic multilingual documents. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 1–8.

Wafia Adouane, Simon Dobnik, Jean-Philippe Bernardy, and Nasredine Semmar. 2018b. A comparison of character neural language model and bootstrapping for language identification in multilingual noisy texts. In *Proceedings of the Second Workshop on Subword/Character LEvel Models*, pages 22–31.

Wafia Adouane, Samia Touileb, and Jean-Philippe Bernardy. 2020. Identifying sentiments in Algerian code-switched user-generated comments. In *Proceedings of LREC*, pages 2698–2705.

Mohamed Aghzal and Asmaa Mourhir. 2021. Distributional word representations for code-mixed text in Moroccan Darija. In *Proceedings of the Fifth International Conference on Arabic Computational Linguistics (ACLing)*, pages 266–273.

Gustavo Aguilar, Fahad AlGhamdi, Victor Soto, Mona Diab, Julia Hirschberg, and Thamar Solorio. 2018. Overview of the CALCS 2018 shared task: Named

entity recognition on code-switched data. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*.

Gustavo Aguilar, Sudipta Kar, and Thamar Solorio. 2020. LinCE: A centralized benchmark for linguistic code-switching evaluation. In *Proceedings of LREC*, pages 1803–1813.

Maryam Al-Ali. 2024. Enhancing automatic speech recognition for Emirati-English code-switched speech. Master's thesis, Mohamed bin Zayed University of Artificial Intelligence.

Maryam Khalifa Al Ali and Hanan Aldarmaki. 2024. Mixat: A data set of bilingual Emirati-English speech. In *Proceedings of the Third Annual Meeting of the Special Interest Group on Under-resourced Languages*, pages 222–226.

Mohamed Al-Badrashiny and Mona Diab. 2016a. The George Washington University system for the code-switching workshop shared task 2016. In *Proceedings of The Second Workshop on Computational Approaches to Code Switching*, pages 108–111.

Mohamed Al-Badrashiny and Mona Diab. 2016b. Lili: A simple language independent approach for language identification. In *Proceedings of COLING*, pages 1211–1219.

Mohamed Al-Badrashiny, Heba Elfardy, and Mona Diab. 2015. Aida2: A hybrid approach for token and sentence level dialect identification in Arabic. In *Proceedings of the 19th Conference on Computational Natural Language Learning*, pages 42–51.

Mohamed Al-Badrashiny, Ramy Eskander, Nizar Habash, and Owen Rambow. 2014. Automatic transliteration of romanized dialectal Arabic. In *Proceedings of the 18th Conference on Computational Natural Language Learning*, pages 30–38.

Faisal Al-Shargi and Owen Rambow. 2015. DIWAN: A dialectal word annotation tool for Arabic. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 49–58.

Majedah Abdullah Saleh Alaiyed. 2018. *Diglossic code-switching between Standard Arabic and Najdi Arabic in religious discourse*. Ph.D. thesis, Durham University.

Abdulkafi Albirini. 2011. The sociolinguistic functions of codeswitching between standard Arabic and dialectal Arabic. *Language in society*, 40(5):537–562.

Fahad AlGhamdi. 2020. *Towards a Unified Framework for Computational Processing of Linguistic Code Switching*. Ph.D. thesis, The George Washington University.

Fahad AlGhamdi and Mona Diab. 2019. Leveraging pretrained word embeddings for part-of-speech tagging of code switching data. *arXiv preprint arXiv:1905.13359*.

Fahad AlGhamdi, Giovanni Molina, Mona Diab, Thamar Solorio, Abdelati Hawwari, Victor Soto, and Julia Hirschberg. 2016. Part of speech tagging for code switched data. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 98–107.

Munassir Alhamami. 2020. Switching of language varieties in Saudi multilingual hospitals: insiders' experiences. *Journal of Multilingual and Multicultural Development*, 41(2):175–189.

Ali Alhazmi, Rohana Mahmud, Norisma Idris, Mohamed Elhag Mohamed Abo, and Christopher Ifeanyi Eke. 2024. Code-mixing unveiled: Enhancing the hate speech detection in Arabic dialect tweets using machine learning models. *Plos one*, 19(7):e0305657.

Ahmed Ali, Shammur Chowdhury, Amir Hussein, and Yasser Hifny. 2021. Arabic code-switching speech recognition using monolingual data. In *Proceedings of Interspeech*, pages 3475–3479.

Mustafa Ali and Duaa Mohamed. 2018. Investigation of linkage factors affecting code-switching in Arabic-English speakers. Master's thesis, University of Toledo.

Nouf Aljasir. 2020. Arabic-English code-switching in Saudi Arabia: Exploring bilinguals' behavior and the individual factors influencing it. *Arab Journal for Scientific Publishing (AJSP) ISSN*, 2663:5798.

Etaf Alkhlaifat, Ping Yang, and Mohamed Moustakim. 2020. Code-switching between Arabic and English in Jordanian GP consultations. *Alkhlaifat, E., Yang, P., & Moustakim, M.(2020). Code-switching between Arabic and English in Jordanian GP consultations. Crossroads: A Journal of English Studies*, 30(3):4–22.

Raghad Y Alkhudair. 2019. Professors' and undergraduate students' perceptions and attitudes toward the use of code-switching and its function in academic classrooms. *International Journal of English Linguistics*, 9(6):160–171.

Tasneem S Almasah, Gamal A Ebrahim, and Marwa A Abdelaal. 2023. A code-switched Arabic-English sentiment analysis approach based on deep-learning. In *Proceedings of the International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC)*, pages 452–457.

Jalal Haris Almathkuri. 2016. *Investigating the motivation behind language alternation in the multilingual medical workplace: a study of language practices at King Abdul Aziz Specialist Hospital, Saudi Arabia*. Ph.D. thesis, University of Southampton.

Nada Saleh Alsamhan and Fatimah Almutrafi. 2022. Codeswitching in online written communication among Arabic English bilinguals: A sociolinguistic perspective. *International Journal of Applied Linguistics and English Literature*, 11(2):79–90.

Ashwaq Alsulami. 2019. A sociolinguistic analysis of the use of Arabizi in social media among Saudi Arabians. *International Journal of English Linguistics*, 9(6):257–270.

Mohammed Altamimi, Osama Alruwaili, and William J Teahan. 2018. BTAC: A twitter corpus for Arabic dialect identification. In *Proceedings of the Sixth conference on computer-mediated communication (CMC) and social media corpora (CMC-corpora 2018)*.

Mohammed Hamed R Altamimi. 2020. *Categorisation of Arabic Twitter Text*. Ph.D. thesis, Bangor University.

Djegdjiga Amazouz. 2019. *Linguistic and phonetic investigations of French-Algerian Arabic code-switching: Large corpus studies using automatic speech*. Ph.D. thesis, University of Texas at Austin, United States.

Djegdjiga Amazouz, Martine Adda-Decker, and Lori Lamel. 2017. Addressing code-switching in French/Algerian Arabic speech. In *Proceedings of Interspeech*, pages 62–66.

Djegdjiga Amazouz, Martine Adda-Decker, and Lori Lamel. 2018. The French-Algerian code-switching triggered audio corpus (FACST). In *Proceedings of LREC*, pages 1468–1473.

Mohammed Attia, Younes Samih, Ali Elkahky, Hamdy Mubarak, Ahmed Abdelali, and Kareem Darwish. 2019. POS tagging for improving code-switching identification in Arabic. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 18–29.

Mohammed Attia, Younes Samih, and Wolfgang Maier. 2018. GHHT at CALCS 2018: Named entity recognition for dialectal Arabic using neural networks. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 98–102.

Nahla Nola Bacha and Rima Bahous. 2011. Foreign language education in lebanon: A context of cultural and curricular complexities. *Journal of Language Teaching and Research*, 2(6):1320.

As-Said Muhammad Badawi. 1973. *Mustawayat Al-Arabiyya Al-Mu'asira fi Misr*. Dar al-maarif.

Mohamed Balabel, Injy Hamed, Slim Abdennadher, Ngoc Thang Vu, and Özlem Çetinoğlu. 2020. Cairo student code-switch (CSCS) corpus: An annotated Egyptian Arabic-English corpus. In *Proceedings of LREC*, pages 3973–3977.

Lamia Bach Baoueb. 2009. Social factors for code-switching in tunisian business companies: A case study.

Kfir Bar, Nachum Dershowitz, Lior Wolf, Yackov Lubarsky, and Yaacov Choueka. 2015. Processing Judeo-Arabic texts. In *Proceedings of the First International Conference on Arabic Computational Linguistics (ACLing)*, pages 138–144.

Rizky Elzandi Barik and Dessi Puji Lestari. 2019. Text corpus and acoustic model addition for Indonesian-Arabic code-switching in automatic speech recognition system. In *Proceedings of the International Conference of Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, pages 1–6.

Anshul Bawa, Pranav Khadpe, Pratik Joshi, Kalika Bali, and Monojit Choudhury. 2020. Do multilingual users prefer chat-bots that code-mix? let's nudge and find out! *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–23.

Moulay Lahssan Baya Essayahi and Nassima Kerras. 2016. A sociolinguistic study of the Algerian language. *Arab World English Journal (AWEJ) Special Issue on CALL*, 3.

Ghazaleh Beigi and Huan Liu. 2020. A survey on privacy in social media: Identification, mitigation, and applications. *ACM Transactions on Data Science*, 1(1):1–38.

Yassine Benajiba and Mona Diab. 2010. A web application for dialectal Arabic text annotation. In *Editors & Workshop Chairs*, page 91.

Abdelali Bentahila. 1983. Motivations for code-switching among Arabic-French bilinguals in Morocco. *Language & communication*, 3(3):233–243.

Ann Bies, Zhiyi Song, Mohamed Maamouri, Stephen Grimes, Haejoong Lee, Jonathan Wright, Stephanie Strassel, Nizar Habash, Ramy Eskander, and Owen Rambow. 2014. Transliteration of arabizi into Arabic orthography: Developing a parallel annotated Arabizi-Arabic script SMS/chat corpus. In *Proceedings of the Workshop on Arabic Natural Language Processing*, pages 93–103.

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 LREC*, pages 3387–3396.

Naaima Boudad, Rdouan Faizi, and Rachid Oulad Haj Thami. 2023. Multilingual, monolingual and mono-dialectal transfer learning for Moroccan Arabic sentiment classification. *Social Network Analysis and Mining*, 14(1):3.

Laurie Burchell, Alexandra Birch, Robert Thompson, and Kenneth Heafield. 2024. Code-switched language identification is harder than you think. In *Proceedings of EACL*, pages 646–658.

Özlem Çetinoğlu, Sarah Schulz, and Ngoc Thang Vu. 2016. Challenges of computational processing of code-switching. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 1–11.

Shuguang Chen, Gustavo Aguilar, Anirudh Srinivasan, Mona Diab, and Thamar Solorio. 2021. CALCS 2021 shared task: Machine translation for code-switched data. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*.

Song Chen, Dana Fore, Stephanie Strassel, Haejoong Lee, and Jonathan Wright. 2017. BOLT Egyptian Arabic SMS/chat and transliteration LDC2017T07. *Philadelphia: Linguistic Data Consortium*.

Mohamed Amine Cheragui, Abdelhalim Hafedh Dahou, and Amin Abdedaiem. 2023. Exploring BERT models for part-of-speech tagging in the Algerian dialect: A comprehensive study. In *Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023)*, pages 140–150.

Parul Chopra, Sai Krishna Rallabandi, Alan W Black, and Khyathi Raghavi Chandu. 2021. Switch point biased self-training: Re-purposing pretrained models

for code-switching. In *Findings of EMNLP*, pages 4389–4397.

Shammur A Chowdhury, Younes Samih, Mohamed Eldesouki, and Ahmed Ali. 2020. Effects of dialectal code-switching on speech modules: A study using Egyptian Arabic broadcast speech. In *Proceedings of Interspeech*, pages 2382–2386.

Shammur Absar Chowdhury, Amir Hussein, Ahmed Abdelali, and Ahmed Ali. 2021. Towards one model to rule all: Multilingual strategy for dialectal code-switching Arabic ASR. In *Proceedings of Interspeech*.

Ryan Cotterell, Adithya Renduchintala, Naomi Saphra, and Chris Callison-Burch. 2014. An Algerian Arabic-French code-switched corpus. In *Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools Workshop Programme*, page 34.

Kareem Darwish. 2014. Arabizi detection and conversion to arabic. In *Proceedings of the Workshop on Arabic Natural Language Processing*, pages 217–224.

Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Nada AlMarwani, and Mohamed Al-Badrashiny. 2019. Creating a large multi-layered representational repository of linguistic code switched Arabic data. *arXiv preprint arXiv:1909.13009*.

A Seza Doğruöz, Sunayana Sitaram, Barbara Bullock, and Almeida Jacqueline Toribio. 2021. A survey of code-switching: Linguistic and social perspectives for language technologies. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1654–1666.

Melinda Dooly and Ola Bakri. 2024. Code-switching in a legal English class for Egyptian learners: A conversation analysis case study. *Arab World English Journal (AWEJ)*, 15(3):297–320.

Alessandro Duranti. 2008. *A companion to linguistic anthropology*. John Wiley & Sons.

Alia El Bolock, Injy Hamed, Yomna Abdelrahman, Ngoc Thang Vu, Cornelia Herbert, and Slim Abdennadher. 2020. Who, when and why: The 3 ws of code-switching. In *Proceedings of the 18th International Conference on Practical Applications of Agents and Multi-Agent Systems, in the third Workshop on Character Computing (C2)*, pages 160–170.

Mahmoud El-Haj, Paul Rayson, and Mariam Aboelezz. 2018. Arabic dialect identification in the context of bivalency and code-switching. In *Proceedings of LREC*, pages 3622–3627.

Mohammed O Elfahal, Mohammed Mustafa, Mohammed Elhafiz Mustafa, and Rashid A Saeed. 2020. A framework for Sudanese Arabic–English mixed speech processing. In *Proceedings of the International Conference on Computing and Information Technology (ICCIT-1441)*, pages 1–6.

Mohammed Osman Eltayeb Elfahal. 2019. *Automatic Recognition and Identification for Mixed Sudanese Arabic–English Languages Speech*. Ph.D. thesis, Sudan University of Science & Technology.

Heba Elfardy, Mohamed Al-Badrashiny, and Mona Diab. 2013. Code switch point detection in Arabic. In *Proceedings of the International Conference on Application of Natural Language to Information Systems*, pages 412–416.

Heba Elfardy, Mohamed Al-Badrashiny, and Mona Diab. 2014. AIDA: Identifying code switching in informal Arabic text. In *Proceedings of The First Workshop on Computational Approaches to Code Switching*, pages 94–101.

Heba Elfardy and Mona Diab. 2012. Token level identification of linguistic code switching. In *Proceedings of COLING*, pages 287–296.

Hebatallah Elfardy. 2017. *Perspective identification in informal text*. Phd thesis, Columbia University.

Abdelrahim Elmadany, Muhammad Abdul-Mageed, et al. 2021. Investigating code-mixed Modern Standard Arabic-Egyptian to English machine translation. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 56–64.

Abdelrahim Elmadany, Muhammad Abdul-Mageed, et al. 2022. AraT5: Text-to-text transformers for arabic language generation. In *Proceedings of ACL*, pages 628–647.

Iyad Elwy and Caroline Sabty. 2024. Takween: A comprehensive Arabic collection and annotation tool. In *Proceedings of the 6th International Conference on AI in Computational Linguistics*, pages 186–193.

Ramy Eskander, Mohamed Al-Badrashiny, Nizar Habash, and Owen Rambow. 2014. Foreign words and the automatic processing of Arabic social media text written in Roman script. In *Proceedings of The First Workshop on Computational Approaches to Code Switching*, pages 1–12.

Ramy Eskander, Shubhanshu Mishra, Sneha Mehta, Sofia Samaniego, and Aria Haghighi. 2022. Towards improved distantly supervised multilingual named-entity recognition for tweets. In *Proceedings of the Second Workshop on Multi-lingual Representation Learning (MRL)*, pages 115–124.

Salma Mohamed Farid. 2019. A case study of syntactic patterns of Egyptian colloquial Arabic-English code-switching. Master's thesis, The American University in Cairo.

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

Marwa Gaser, Manuel Mager, Injy Hamed, Nizar Habash, Slim Abdennadher, and Ngoc Thang Vu. 2023. Exploring segmentation approaches for neural machine translation of code-switched Egyptian Arabic-English text. In *Proceedings of EACL*, pages 3523–3538.

Parvathy Geetha, Khyathi Chandu, and Alan W Black. 2018. Tackling code-switched NER: Participation of CMU. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 126–131.

Imane Guellil, Ahsan Adeel, Faical Azouaou, Mohamed Boubred, Yousra Houichi, and Akram Abdelhaq Moumna. 2021. Sexism detection: The first corpus in Algerian dialect with a code-switching in Arabic/French and English. *arXiv preprint arXiv:2104.01443*.

Nizar Habash, Owen Rambow, Mona Diab, and Reem Kanjawi-Faraj. 2008. Guidelines for annotation of arabic dialectness. In *Proceedings of the LREC Workshop on HLT & NLP within the Arabic world*, pages 49–53.

Nizar Habash, Abdelhadi Soudi, and Timothy Buckwalter. 2007. On Arabic transliteration. *Arabic computational morphology: Knowledge-based and empirical methods*, pages 15–22.

Rana Medhat Hafez. 2015. Factors affecting code switching between Arabic and English. M.A. thesis, American University in Cairo.

Soufiane Hajbi, Younes Chihab, Rachid Ed-Dali, and Redouan Korchiyne. 2022. Natural language processing based approach to overcome Arabizi and code switching in social media Moroccan dialect. In *Proceedings of Advances in Information, Communication and Cybersecurity (ICI2C'21)*, pages 57–66. Springer.

Injy Hamed. 2024. *Neural-based NLP systems for code-switched Arabic-English speech*. Ph.D. thesis, University of Stuttgart.

Injy Hamed, Pavel Denisov, Chia-Yu Li, Mohamed Elmahdy, Slim Abdennadher, and Ngoc Thang Vu. 2022a. Investigations on speech recognition systems for low-resource dialectal Arabic–English code-switching speech. *Computer Speech & Language*, 72:101278.

Injy Hamed, Alia El Bolock, Cornelia Herbert, Slim Abdennadher, and Ngoc Thang Vu. 2022b. The who in code-switching: A case study for predicting Egyptian Arabic–English code-switching levels based on character profiles. *International Journal of Asian Language Processing*, 32(01):2250010.

Injy Hamed, Alia El Bolock, Nader Rizk, Cornelia Herbert, Slim Abdennadher, and Ngoc Thang Vu. 2021. Predicting user code-switching level from sociological and psychological profiles. In *Proceedings of the International Conference on Asian Language Processing (IALP)*, pages 395–400.

Injy Hamed, Mohamed Elmahdy, and Slim Abdennadher. 2017. Building a first language model for code-switch Arabic-English. In *Proceedings of the International Conference on Arabic Computational Linguistics*, pages 208–216.

Injy Hamed, Mohamed Elmahdy, and Slim Abdennadher. 2018a. Collection and analysis of code-switch egyptian arabic-english speech corpus. In *Proceedings of LREC*, pages 208–216.

Injy Hamed, Mohamed Elmahdy, and Slim Abdennadher. 2018b. Expanding n-grams for code-switch language models. In *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics*, pages 221–229.

Injy Hamed, Fadhl Eryani, David Palfreyman, and Nizar Habash. 2024. ZAEBUC-spoken: A multilingual multidialectal Arabic-English speech corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17770–17782.

Injy Hamed, Nizar Habash, Slim Abdennadher, and Ngoc Thang Vu. 2022c. ArzEn-ST: A three-way speech translation corpus for code-switched Egyptian Arabic-English. In *Proceedings of the Arabic Natural Language Processing Workshop (WANLP)*, pages 119–130.

Injy Hamed, Nizar Habash, Slim Abdennadher, and Ngoc Thang Vu. 2023a. Investigating lexical replacements for Arabic-English code-switched data augmentation. In *Proceedings of the Sixth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT)*, pages 86–100.

Injy Hamed, Nizar Habash, and Thang Vu. 2023b. Data augmentation techniques for machine translation of code-switched texts: A comparative study. In *Findings of EMNLP*, pages 140–154.

Injy Hamed, Amir Hussein, Oumnia Chellah, Shammur Chowdhury, Hamdy Mubarak, Sunayana Sitaram, Nizar Habash, and Ahmed Ali. 2023c. Benchmarking evaluation metrics for code-switching automatic speech recognition. In *Proceedings of SLT*, pages 999–1005.

Injy Hamed, Ngoc Thang Vu, and Slim Abdennadher. 2020. ArzEn: A speech corpus for code-switched Egyptian Arabic-English. In *Proceedings of LREC*, pages 4237–4246.

Injy Hamed, Moritz Zhu, Mohamed Elmahdy, Slim Abdennadher, and Ngoc Thang Vu. 2019. Code-switching language modeling with bilingual word embeddings: A case study for Egyptian Arabic-English. In *Proceedings of the International Conference on Speech and Computer*, pages 160–170.

Ahmed Heakl, Youssef Zaghloul, Mennatullah Ali, Rania Hossam, and Walid Gomaa. 2024. ArzEn-LLM: Code-switched Egyptian Arabic-English translation and speech recognition using LLMs. *arXiv preprint arXiv:2406.18120*.

Roberto R Heredia and Jeanette Altarriba. 2001. Bilingual language mixing: Why do bilinguals code-switch? *Current directions in psychological science*, 10(5):164–168.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, et al. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of ACL*, pages 6997–7013.

Amir Hussein, Shammur Absar Chowdhury, Ahmed Abdelali, Najim Dehak, Ahmed Ali, and Sanjeev Khudanpur. 2023. Textual data augmentation for Arabic-English code-switching speech recognition. In *Proceedings of SLT*, pages 777–784.

Amir Hussein, Dorsa Zeinali, Ondřej Klejch, Matthew Wiesner, Brian Yan, Shammur Chowdhury, Ahmed Ali, Shinji Watanabe, and Sanjeev Khudanpur. 2024. Speech collage: code-switched audio generation by collaging monolingual corpora. In *Proceedings of ICASSP*, pages 12006–12010.

Manal A Ismail. 2015. The sociolinguistic dimensions of code-switching between Arabic and English by saudis. *International Journal of English Linguistics*, 5(5):99.

Aaron Jaech, George Mulcaire, Mari Ostendorf, and Noah A Smith. 2016. A neural model for language identification in code-switched tweets. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 60–64.

Devanshu Jain, Maria Kustikova, Mayank Darbari, Rishabh Gupta, and Stephen Mayhew. 2018. Simple features for strong performance on named entity recognition in code-switched twitter data. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 103–109.

Naman Jain and Riyaz Ahmad Bhat. 2014. Language identification in code-switching scenario. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 87–93.

Florian Janke, Tongrui Li, Eric Rincón, Gualberto A Guzman, Barbara Bullock, and Almeida Jacqueline Toribio. 2018. The University of Texas system submission for the code-switching workshop shared task 2018. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 120–125.

Mohamed Amine Jerbi, Hadhemi Achour, and Emna Souissi. 2019. Sentiment analysis of code-switched Tunisian dialect: Exploring RNN-based techniques. In *Proceedings of the International Conference on Arabic Language Processing*, pages 122–131.

Karima Kadaoui, Maryam Ali, Hawau Toyin, Ibrahim Mohammed, and Hanan Aldarmaki. 2024. PolyWER: A holistic evaluation framework for code-switched speech recognition. In *Findings of EMNLP*, pages 6144–6153.

Tom Kalkman. 2024. Detecting and analyzing code-switching behaviour in a Moroccan-Dutch dataset using transformer architectures. Master's thesis, Utrecht University.

Amr Keleg, Sharon Goldwater, and Walid Magdy. 2023. ALDi: Quantifying the Arabic level of dialectness of text. In *Proceedings of EMNLP*, pages 10597–10611.

Amr Keleg and Walid Magdy. 2023. Arabic dialect identification under scrutiny: Limitations of single-label classification. In *Proceedings of the Arabic Natural Language Processing Conference*, pages 385–398.

Geoffrey Khan, Michael P Streck, and Janet CE Watson. 2011. *The Semitic languages: An international handbook*, volume 36. Walter de Gruyter.

Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. 2020. GLUECoS: An evaluation benchmark for code-switched nlp. In *Proceedings of ACL*, pages 3575–3585.

Md Tawkat Islam Khondaker, Numaan Naeem, Fatimah Khan, Abdelrahim Elmadany, and Muhammad Abdul-Mageed. 2024. Benchmarking LLaMA-3 on Arabic language generation tasks. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 283–297.

Md Tawkat Islam Khondaker, Abdul Waheed, Muhammad Abdul-Mageed, et al. 2023. GPTAraEval: A comprehensive evaluation of ChatGPT on Arabic NLP. In *Proceedings of EMNLP*, pages 220–247.

Abdullah A Khuwaileh. 2003. Code switching and multilingualism in a small multi-ethnic group society (uae). *Journal of Language for International Business*, 14(2):32–49.

Levi King, Eric Baucom, Timur Gilmanov, Sandra K$^{\ddot{\text{u}}}$bler, Daniel Whyatt, Wolfgang Maier, and Paul Rodrigues. 2014. The IUCL+ system: Word-level language identification via extended markov models. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 102–106.

Małgorzata Kniaź and Magdalena Zawrotna. 2021. Embedded English verbs in Arabic-English code-switching in egypt. *International Journal of Bilingualism*, 25(3):622–639.

Prashant Kodali, Anmol Goel, Likhith Asapu, Vamshi Krishna Bonagiri, Anirudh Govil, Monojit Choudhury, Manish Shrivastava, and Ponnurangam Kumaraguru. 2024. From human judgements to predictive models: Unravelling acceptability in code-mixed sentences. *arXiv preprint arXiv:2405.05572*.

Atharva Kulkarni, Ajinkya Kulkarni, Miguel Couceiro, and Hanan Aldarmaki. 2023. Adapting the adapters for code-switching in multilingual ASR. *arXiv preprint arXiv:2310.07423*.

Garry Kuwanto, Chaitanya Agarwal, Genta Indra Winata, and Derry Tanti Wijaya. 2024. Linguistics theory meets LLM: Code-switched text generation via equivalence constrained large language models. *arXiv preprint arXiv:2410.22660*.

Houssam Eddine-Othman Lachemat, Akli Abbas, Nourredine Oukas, Yassine El Kheir, Samia Haboussi, and Absar Showdhury Shammur. 2024. CAFE: a novel code switching dataset for Algerian dialect French and English. *arXiv preprint arXiv:2411.13424*.

Wuwei Lan, Yang Chen, Wei Xu, and Alan Ritter. 2020. An empirical study of pre-trained transformers for Arabic information extraction. In *Proceedings of EMNLP*, pages 4727–4734.

Chu-Cheng Lin, Waleed Ammar, Lori Levin, and Chris Dyer. 2014. The CMU submission for the shared task on language identification in code-switched data. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 80–86.

Khaled Lounnas, Mourad Abbas, and Mohamed Lichouri. 2021. Towards phone number recognition

for code switched Algerian dialect. In *Proceedings of the Fourth International Conference on Natural Language and Speech Processing (ICNLSP 2021)*, pages 290–294.

M'hamed Mataoui, Omar Zelmati, and Madiha Boumechache. 2016. A proposed lexicon-based sentiment analysis approach for the vernacular Algerian Arabic. *Research in Computing Science*, 110(1):55–70.

Mohamed Amine Menacer, David Langlois, Denis Jouvet, Dominique Fohr, Odile Mella, and Kamel Smaïli. 2019. Machine translation on a parallel code-switched corpus. In *Proceedings of the Canadian Conference on Artificial Intelligence*, pages 426–432. Springer.

Sarina Meyer, Florian Lux, Pavel Denisov, Julia Koch, Pascal Tilli, and Ngoc Thang Vu. 2022. Speaker anonymization with phonetic intermediate representations. In *Proceedings of Interspeech*, pages 4925–4929.

Sarina Meyer, Florian Lux, Julia Koch, Pavel Denisov, Pascal Tilli, and Ngoc Thang Vu. 2023a. Prosody is not identity: A speaker anonymization approach using prosody cloning. In *Proceedings of ICASSP*, pages 1–5.

Sarina Meyer, Florian Lux, and Ngoc Thang Vu. 2024. Probing the feasibility of multilingual speaker anonymization. In *Proceedings of Interspeech*, pages 4448–4452.

Sarina Meyer, Pascal Tilli, Pavel Denisov, Florian Lux, Julia Koch, and Ngoc Thang Vu. 2023b. Anonymizing speech with generative adversarial networks to preserve speaker privacy. In *Proceedings of SLT*, pages 912–919.

Kurt Micallef, Nizar Habash, Claudia Borg, Fadhl Eryani, and Houda Bouamor. 2024. Cross-lingual transfer from related languages: Treating low-resource Maltese as multilingual code-switching. In *Proceedings of EACL*, pages 1014–1025.

Daniel Weisberg Mitelman, Nachum Dershowitz, and Kfir Bar. 2024. Code-switching and back-transliteration using a bilingual model. In *Findings of EACL*, pages 1501–1511.

Iyad Ahmad Hamdan Mkahal. 2016. Code switching as a linguistic phenomenon among Palestinian English Arabic bilinguals with reference to translation. Master's thesis, An-Najah National University.

Djegdjiga Mohdeb-Amazouz, Adda-Decker Martine, and Lori Lamel. 2016. Arabic-French code-switching across Maghreb Arabic dialects: a quantitative analysis. In *Proceedings of the Corpus-driven Studies of Heterogeneous and Multilingual Corpora Workshop*.

Giovanni Molina, Fahad AlGhamdi, Mahmoud Ghoneim, Abdelati Hawwari, Nicolas Rey-Villamizar, Mona Diab, and Thamar Solorio. 2016. Overview for the second shared task on language identification in code-switched data. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 40–49.

Sneha Mondal, Shreya Pathak, Preethi Jyothi, Aravindan Raghuveer, et al. 2022. CoCoa: An encoder-decoder model for controllable code-switched generation. In *Proceedings of EMNLP*, pages 2466–2479.

Hamdy Mubarak, Amir Hussein, Shammur Absar Chowdhury, and Ahmed Ali. 2021. QASR: QCRI Aljazeera speech resource a large scale annotated Arabic speech corpus. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2274–2285.

Benjamin Muller, Benoit Sagot, and Djamé Seddah. 2020. Can multilingual language models transfer to an unseen dialect? a case study on north african arabizi. *arXiv preprint arXiv:2005.00318*.

Mumtaz Begum Mustafa, Mansoor Ali Yusoof, Hasan Kahtan Khalaf, Ahmad Abdel Rahman Mahmoud Abushariah, Miss Laiha Mat Kiah, Hua Nong Ting, and Saravanan Muthaiyah. 2022. Code-switching in automatic speech recognition: The issues and future directions. *Applied Sciences*, 12(19):9541.

Zahra Mustafa and Mahmoud Al-Khatib. 1994. Code-mixing of Arabic and English in teaching science. *World Englishes*, 13(2):215–224.

Carol Myers-Scotton. 1997. *Duelling languages: Grammatical structure in codeswitching*. Oxford University Press.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, Ahmed El-Shangiti, and Muhammad Abdul-Mageed. 2023. Dolphin: A challenging and diverse benchmark for Arabic NLG. In *Findings of EMNLP*, pages 1404–1422.

Abdulfattah Omar and Mohammed Ilyas. 2018. The sociolinguistic significance of the attitudes towards code-switching in Saudi Arabia academia. *International Journal of English Linguistics*, 8(3).

Shana Poplack. 1980. Sometimes i'll start a sentence in spanish y termino en espanol: toward a typology of code-switching. *Linguistics*, 18(7-8):581–618.

Rebekah Elizabeth Post. 2015. *The impact of social factors on the use of Arabic-French code-switching in speech and IM in Morocco*. Ph.D. thesis, The University of Texas at Austin.

SaiKrishna Rallabandi, Sunayana Sitaram, and Alan W Black. 2018. Automatic detection of code-switching style from acoustics. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 76–81.

Houda Saadane, Hosni Seffih, Christian Fluhr, Khalid Choukri, and Nasredine Semmar. 2018. Automatic identification of Maghreb dialects using a dictionary-based approach. In *Proceedings of LREC*, pages 3638–3644.

Hadeel Saadany and Constantin Orasan. 2020. Is it great or terrible? preserving sentiment in neural machine translation of Arabic reviews. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 24–37.

Caroline Sabty. 2024. Computational approaches to Arabic-English code-switching. *arXiv preprint arXiv:2410.13318*.

Caroline Sabty, Mohamed Elmahdy, and Slim Abdennadher. 2019a. Named entity recognition on Arabic-English code-mixed data. In *Proceedings of the 13th International Conference on Semantic Computing (ICSC)*, pages 93–97.

Caroline Sabty, Mohamed Islam, and Slim Abdennadher. 2020. Contextual embeddings for Arabic-English code-switched data. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 215–225.

Caroline Sabty, Islam Mesabah, Özlem Çetinoğlu, and Slim Abdennadher. 2021a. Language identification of intra-word code-switching for Arabic–English. *Array*, 12:100104.

Caroline Sabty, Islam Omar, Fady Wasfalla, Mohamed Islam, and Slim Abdennadher. 2021b. Data augmentation techniques on Arabic data for named entity recognition. In *Proceedings of the International Conference on AI in Computational Linguistics*, volume 189, pages 292–299.

Caroline Sabty, Ahmed Sherif, Mohamed Elmahdy, and Slim Abdennadher. 2019b. Techniques for named entity recognition on Arabic-English code-mixed data. *International Journal of Transdisciplinary AI*, 1(1):44–63.

Younes Samih. 2017. *Dialectal Arabic processing using deep learning*. Ph.D. thesis, Dissertation, Düsseldorf, Heinrich-Heine-Universität, 2017.

Younes Samih, Suraj Maharjan, Mohammed Attia, Laura Kallmeyer, and Thamar Solorio. 2016. Multilingual code-switching identification via LSTM recurrent neural networks. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 50–59.

Younes Samih and Wolfgang Maier. 2016a. An Arabic-Moroccan Darija code-switched corpus. In *Proceedings of LREC*, pages 4170–4175.

Younes Samih and Wolfgang Maier. 2016b. Detecting code-switching in Moroccan Arabic social media. *SocialNLP@ IJCAI-2016, New York*.

Djamé Seddah, Farah Essaidi, Amal Fethi, Matthieu Futeral, Benjamin Muller, Pedro Javier Ortiz Suárez, Benoît Sagot, and Abhishek Srivastava. 2020. Building a user-generated content North-African Arabizi treebank: Tackling hell. In *Proceedings of ACL*, pages 1139–1150.

Ayan Sengupta, Md Shad Akhtar, and Tanmoy Chakraborty. 2023. Persona-aware generative model for code-mixed language. *arXiv preprint arXiv:2309.02915*.

Ali Shazal, Aiza Usman, and Nizar Habash. 2020. A unified model for Arabizi detection and transliteration using sequence-to-sequence models. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop (WANLP)*, pages 167–177.

Safaa Shehadi and Shuly Wintner. 2022. Identifying code-switching in Arabizi. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 194–204.

Ahmed Sherif and Caroline Sabty. 2024. Sentiment analysis for Egyptian Arabic-English code-switched data using traditional neural models and advanced language models. In *International Conference on Speech and Computer*, pages 54–69.

Prajwol Shrestha. 2016. Codeswitching detection via lexical features in conditional random fields. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 121–126.

Sunayana Sitaram, Khyathi Raghavi Chandu, Sai Krishna Rallabandi, and Alan W Black. 2019. A survey of code-switched speech and language processing. *arXiv preprint arXiv:1904.00784*.

Kamel Smaïli, Hamza Anissa, Langlois David, and Amazouz Djegdjiga. 2024. BOUTEF: Bolstering our understanding through an elaborated fake news corpus. In *Proceedings of the Eighth International Conference on Arabic Language Processing*.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, et al. 2014. Overview for the first shared task on language identification in code-switched data. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 62–72.

Thamar Solorio and Yang Liu. 2008. Learning to predict code-switching points. In *Proceedings of EMNLP*, pages 973–981.

Ondřej Sotolář, Jaromír Plhák, and David Šmahel. 2021. Towards personal data anonymization for social messaging. In *International Conference on Text, Speech, and Dialogue*, pages 281–292.

Sara Stefanich, Jennifer Cabrelli, Dustin Hilderman, and John Archibald. 2019. The morphophonology of intraword codeswitching: Representation and processing. *Frontiers in Communication*, 4:54.

Hanada Taha Thomure. 2019. Arabic language education in the UAE: Choosing the right drivers. *Education in the United Arab Emirates: Innovation and Transformation*, pages 75–93.

Taghreed Tarmom, William Teahan, Eric Atwell, and Mohammad Ammar Alsalka. 2020. Compression versus traditional machine learning classifiers to detect code-switching in varieties and dialects: Arabic as a case study. *Natural Language Engineering*, pages 1–14.

Juan Antonio Thomas and Lotfi Sayahi. 2019. Language contact phenomena in three Aljamiado texts: Religion as a sociolinguistic factor. *eHumanista: Journal of Iberian Studies*, 41:142–154.

Taha Tobaili, Miriam Fernandez, Harith Alani, Sanaa Sharafeddine, Hazem Hajj, and Goran Glavaš. 2019. Senzi: A sentiment analysis lexicon for the Latinised Arabic (Arabizi). In *Proceedings of the International*

*Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1203–1211.

Samia Touileb and Jeremy Barnes. 2021. The interplay between language similarity and script on a novel multi-layer Algerian dialect corpus. In *Findings of ACL-IJCNLP*, pages 3700–3712.

Enes Yavuz Ugan, Christian Huber, Juan Hussain, and Alexander Waibel. 2022. Language-agnostic code-switching in sequence-to-sequence speech recognition. *arXiv preprint arXiv:2210.08992*.

Clare R Voss, Stephen Tratz, Jamal Laoudi, and Douglas M Briesch. 2014. Finding Romanized Arabic dialect in code-mixed tweets. In *Proceedings of LREC*, pages 2249–2253.

Ngoc Thang Vu, Heike Adel, and Tanja Schultz. 2013. An investigation of code-switching attitude dependent language modeling. In *Proceedings of the International Conference on Statistical Language and Speech Processing*, pages 297–308.

Changhan Wang, Kyunghyun Cho, and Douwe Kiela. 2018. Code-switched named entity recognition with embedding attention. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 154–158.

Genta Winata, Alham Fikri Aji, Zheng Xin Yong, and Thamar Solorio. 2023. The decades progress on code-switching research in NLP: A systematic survey on trends and challenges. In *Findings of ACL*, pages 2936–2978.

Genta Indra Winata, Samuel Cahyawijaya, Zihan Liu, Zhaojiang Lin, Andrea Madotto, and Pascale Fung. 2021. Are multilingual models effective in code-switching? In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 142–153.

Ping Yang. 2021. *A Sociolinguistic Study of Doctor-Patient Interaction in Healthcare Settings: A Jordanian Perspective*. Ph.D. thesis, Sydney University.

Zheng Xin Yong, Ruochen Zhang, Jessica Forde, Skyler Wang, Arjun Subramonian, Holy Lovenia, Samuel Cahyawijaya, Genta Winata, Lintang Sutawika, Jan Christian Blaise Cruz, et al. 2023. Prompting multilingual large language models to generate code-mixed texts: The case of South East Asian languages. In *Proceedings of the 6th Workshop on Computational Approaches to Linguistic Code-Switching*, pages 43–63.

Zuhair D Zaghlool and Norah M Altamimi. 2023. Saudi EFL teachers' and students' perceptions towards using English-Arabic code switching as a teaching and learning strategy. *Journal of Language Teaching and Research*, 14(4):1049–1057.

Bin Zhou, Jian Pei, and WoShun Luk. 2008. A brief survey on anonymization techniques for privacy preserving publishing of social network data. *ACM Sigkdd Explorations Newsletter*, 10(2):12–22.

Inès Zribi, Mariem Ellouze, Lamia Hadrich Belguith, and Philippe Blache. 2015. Spoken Tunisian Arabic corpus "STAC": transcription and annotation. *Research in computing science*, 90:123–135.

Inès Zribi, Inès Kammoun, Mariem Ellouze, L Belguith, and Philippe Blache. 2016. Sentence boundary detection for transcribed Tunisian Arabic. *Bochumer Linguistische Arbeitsberichte*, 323.

## A Language and Dialect Codes

We outline the codes we use for languages and dialects in Table 4.

## B Overview on Empirical Papers

In Tables 5-8, we provide a list of the empirical papers, specifying the covered language pairs, for text-based and speech-based NLP tasks.

## C Overview on Resource Papers

In Tables 9-11, we provide a list of the resource papers. For each text-based and speech-based NLP task, we specify the language pairs covered by the collected resources presented in the papers.

| Language/Dialect | Code |
|---|---|
| Arabic | ARA |
| Dialectal Arabic | DA |
| Modern Standard Arabic | MSA |
| Algerian Arabic | ALG |
| Egyptian Arabic | EGY |
| Emirati Arabic | EMI |
| Iraqi Arabic | IRQ |
| Jordanian Arabic | JOR |
| Lebanese Arabic | LEB |
| Levantine Arabic | LEV |
| Libyian Arabic | LIB |
| Moroccan Arabic | MOR |
| North African Arabic | NOR |
| Palestinian Arabic | PAL |
| Saudi Arabian Arabic | SAU |
| Sudanese Arabic | SUD |
| Tunisian Arabic | TUN |
| Foreign Language | FOR |
| Aramaic | ARC |
| Berber | BER |
| Dutch | NLD |
| English | ENG |
| French | FRA |
| German | DEU |
| Hebrew | HEB |
| Indonesian | IND |
| Judeo Arabic | JRB |
| Spanish | SPA |

Table 4: The list of codes we use for languages and dialects throughout the paper.

| | DA-Foreign | MSA-Foreign | ARA-Foreign | MSA-DA | MSA-DA-Foreign | Language Details |
|---|---|---|---|---|---|---|
| Word-level Language Identification | | | | | | |
| (Eskander et al., 2014) | ✓ | | | | | EGY-ENG |
| (Voss et al., 2014) | ✓ | | | | | MOR-ENG-FRA |
| (Bar et al., 2015) | ✓ | | | | | JRB-HEB |
| (Aghzal and Mourhir, 2021) | ✓ | | | | | MOR-ENG-FRA |
| (Mitelman et al., 2024) | ✓ | | | | | JRB-HEB/ARC |
| (Kalkman, 2024) | ✓ | | | | | MOR-NLD |
| (Darwish, 2014) | | | ✓ | | | ARA-ENG |
| (Sabty et al., 2021a) | | | ✓ | | | ARA-ENG |
| (Shehadi and Wintner, 2022) | | | ✓ | | | ARA-ENG-FRA |
| (Elfardy and Diab, 2012) | | | | ✓ | | MSA-EGY, MSA-LEV |
| (Elfardy et al., 2013) | | | | ✓ | | MSA-EGY |
| (Elfardy et al., 2014) | | | | ✓ | | MSA-EGY |
| (Solorio et al., 2014) | | | | ✓ | | MSA-EGY |
| (King et al., 2014) | | | | ✓ | | MSA-EGY |
| (Lin et al., 2014) | | | | ✓ | | MSA-EGY |
| (Jain and Bhat, 2014) | | | | ✓ | | MSA-EGY |
| (Al-Badrashiny et al., 2015) | | | | ✓ | | MSA-EGY |
| (Al-Badrashiny and Diab, 2016a) | | | | ✓ | | MSA-EGY |
| (Molina et al., 2016) | | | | ✓ | | MSA-EGY |
| (Samih et al., 2016) | | | | ✓ | | MSA-EGY |
| (Shrestha, 2016) | | | | ✓ | | MSA-EGY |
| (Jaech et al., 2016) | | | | ✓ | | MSA-EGY |
| (Attia et al., 2019) | | | | ✓ | | MSA-EGY |
| (Aguilar et al., 2020) | | | | ✓ | | MSA-EGY |
| (Al-Badrashiny and Diab, 2016b) | ✓ | | | ✓ | | MSA-EGY/LEV/GLF, DA-ENG |
| (Saadane et al., 2018) | ✓ | | | ✓ | | MSA, ALG/TUN/MOR/EGY, ENG/FRA |
| (Samih and Maier, 2016b) | | | | | ✓ | MSA-MOR-FRA/SPA/BER |
| (Adouane and Dobnik, 2017) | | | | | ✓ | MSA-ALG-ENG-FRA-BER |
| (Adouane et al., 2018b) | | | | | ✓ | MSA-ALG-ENG-FRA-BER |
| (Adouane et al., 2018a) | | | | | ✓ | MSA-ALG-ENG-FRA-BER |
| (Adouane and Bernardy, 2020) | | | | | ✓ | MSA-ALG-ENG-FRA-BER |
| (Tarmom et al., 2020) | | | | | ✓ | MSA-EGY-SAU-ENG |
| Named Entity Recognition | | | | | | |
| (Sabty et al., 2019a) | ✓ | ✓ | | | | MSA-ENG, EGY-ENG |
| (Sabty et al., 2019b) | ✓ | ✓ | | | | MSA-ENG, EGY-ENG |
| (Sabty et al., 2020) | ✓ | ✓ | | | | MSA-ENG, EGY-ENG |
| (Sabty et al., 2021b) | ✓ | ✓ | | | | MSA-ENG, EGY-ENG |
| (Aguilar et al., 2018) | | | | ✓ | | MSA-EGY |
| (Wang et al., 2018) | | | | ✓ | | MSA-EGY |
| (Jain et al., 2018) | | | | ✓ | | MSA-EGY |
| (Geetha et al., 2018) | | | | ✓ | | MSA-EGY |
| (Janke et al., 2018) | | | | ✓ | | MSA-EGY |
| (Attia et al., 2018) | | | | ✓ | | MSA-EGY |
| (Aguilar et al., 2020) | | | | ✓ | | MSA-EGY |
| (Chopra et al., 2021) | | | | ✓ | | MSA-EGY |
| (Winata et al., 2021) | | | | ✓ | | MSA-EGY |
| (Eskander et al., 2022) | | | | ✓ | | MSA-EGY |
| (Adouane and Bernardy, 2020) | | | | | ✓ | MSA-ALG-ENG-FRA-BER |

Table 5: List of empirical papers for the stated text-based tasks.

| | DA-Foreign | MSA-Foreign | ARA-Foreign | MSA-DA | MSA-DA-Foreign | Language Details |
|---|---|---|---|---|---|---|
| Machine Translation | | | | | | |
| (Hamed et al., 2022c) | ✓ | | | | | EGY-ENG |
| (Hamed et al., 2023b) | ✓ | | | | | EGY-ENG |
| (Gaser et al., 2023) | ✓ | | | | | EGY-ENG |
| (Hamed et al., 2023a) | ✓ | | | | | EGY-ENG |
| (Khondaker et al., 2023) | ✓ | | | | | ALG-FRA, JOR-ENG |
| (Nagoudi et al., 2023) | ✓ | | | | | ALG/MOR/TUN-FRA, EGY/JOR/PAL-ENG |
| (Khondaker et al., 2024) | ✓ | | | | | ALG/MOR/TUN-FRA, EGY/JOR/PAL-ENG |
| (Heakl et al., 2024) | ✓ | | | | | EGY-ENG |
| (Menacer et al., 2019) | | ✓ | | | | MSA-ENG |
| (Elmadany et al., 2022) | ✓ | ✓ | | | | ALG-FRA, JOR-ENG, (synthetic) MSA-ENG/FRA |
| (Saadany and Orasan, 2020) | | | | ✓ | | MSA-DA (DA mostly EGY) |
| (Elmadany et al., 2021) | | | | ✓ | | MSA-EGY |
| (Chen et al., 2021) | | | | ✓ | | MSA-EGY |
| Sentiment Analysis | | | | | | |
| (Tobaili et al., 2019) | ✓ | | | | | LEB-ENG |
| (Touileb and Barnes, 2021) | ✓ | | | | | ALG-FRA |
| (Boudad et al., 2023) | ✓ | | | | | MOR-FOR |
| (Sherif and Sabty, 2024) | ✓ | | | | | EGY-ENG |
| (Sabty et al., 2020) | ✓ | ✓ | | | | MSA-ENG, EGY-ENG |
| (Almasah et al., 2023) | | | ✓ | | | ARA-FOR |
| (Jerbi et al., 2019) | | | | | ✓ | MSA-TUN-ENG-FRA |
| (Mataoui et al., 2016) | | | | | ✓ | MSA-ALG-FRA |
| (Adouane et al., 2020) | | | | | ✓ | MSA-ALG-FRA-BER |
| (Adouane and Bernardy, 2020) | | | | | ✓ | MSA-ALG-ENG-FRA-BER |
| Part-of-Speech Tagging | | | | | | |
| (Muller et al., 2020) | ✓ | | | | | ALG-FRA |
| (Seddah et al., 2020) | ✓ | | | | | ALG-FRA |
| (Touileb and Barnes, 2021) | ✓ | | | | | ALG-FRA |
| (Cheragui et al., 2023) | ✓ | | | | | ALG-FOR |
| (AlGhamdi et al., 2016) | | | | ✓ | | MSA-EGY |
| (AlGhamdi and Diab, 2019) | | | | ✓ | | MSA-EGY/LEV |
| Sentence-level Language Identification | | | | | | |
| (Shehadi and Wintner, 2022) | | | ✓ | | | ARA-ENG-FRA |
| (Al-Badrashiny et al., 2015) | | | | ✓ | | MSA-EGY |
| (Burchell et al., 2024) | | | | ✓ | | MSA-EGY |
| (El-Haj et al., 2018) | | | | ✓ | | MSA-GLF/LEV/TUN/EGY |
| (Altamimi, 2020) | | | | ✓ | | MSA-EGY/GLF/IRQ MSA-LEV/Maghrebi |

Table 6: List of empirical papers for the stated text-based tasks.

| | DA-Foreign | MSA-Foreign | ARA-Foreign | MSA-DA | MSA-DA-Foreign | Language Details |
|---|---|---|---|---|---|---|
| **Transliteration** | | | | | | |
| (Al-Badrashiny et al., 2014) | ✓ | | | | | EGY-ENG |
| (Eskander et al., 2014) | ✓ | | | | | EGY-ENG |
| (Shazal et al., 2020) | ✓ | | | | | EGY-ENG |
| (Mitelman et al., 2024) | ✓ | | | | | JRB-HEB/ARC |
| (Darwish, 2014) | | | ✓ | | | ARA-ENG |
| **Language Modeling** (Excluding language modeling conducted as part of ASR efforts) | | | | | | |
| (Hamed et al., 2019) | ✓ | | | | | EGY-ENG |
| (Hamed et al., 2017) | | ✓ | | | | MSA-ENG |
| (Hamed et al., 2018b) | | ✓ | | | | MSA-ENG |
| (Lan et al., 2020) | | | ✓ | | | ARA-ENG |
| **Abusive Language Detection** | | | | | | |
| (Guellil et al., 2021) | ✓ | | | | | ALG-ENG-FRA |
| (Alhazmi et al., 2024) | | | ✓ | | | ARA-FOR |
| **Sentence-level Micro-Dialect Identification** | | | | | | |
| (Abainia, 2018) | ✓ | | | | | ALG-FRA |
| (Abdul-Mageed et al., 2020) | | | ✓ | | | ARA-FOR |
| **Spelling Correction and Text Normalization** | | | | | | |
| (Adouane et al., 2019) | | | | | ✓ | MSA-ALG-ENG-FRA-BER |
| (Adouane and Bernardy, 2020) | | | | | ✓ | MSA-ALG-ENG-FRA-BER |
| **Dependency Parsing** | | | | | | |
| (Muller et al., 2020) | ✓ | | | | | ALG-FRA |
| (Seddah et al., 2020) | ✓ | | | | | ALG-FRA |
| **Tokenization** | | | | | | |
| (Gaser et al., 2023) | ✓ | | | | | EGY-ENG |
| **Fake News Detection** | | | | | | |
| (Abdelouahab and Smaïli, 2024) | | | | | ✓ | MSA-TUN/ALG-ENG-FRA |
| **Word Analogy** | | | | | | |
| (Aghzal and Mourhir, 2021) | ✓ | | | | | MOR-ENG-FRA |
| **Topic Modeling** | | | | | | |
| (Touileb and Barnes, 2021) | ✓ | | | | | ALG-FRA |
| **Question Answering** | | | | | | |
| (Sabty et al., 2020) | ✓ | ✓ | | | | MSA-ENG, EGY-ENG |

Table 7: List of empirical papers for the stated text-based tasks.

| | DA-Foreign | MSA-Foreign | ARA-Foreign | MSA-DA | MSA-DA-Foreign | Language Details |
|---|---|---|---|---|---|---|
| Automatic Speech Recognition | | | | | | |
| (Elfahal, 2019) | ✓ | | | | | SUD-ENG |
| (Elfahal et al., 2020) | ✓ | | | | | SUD-ENG |
| (Lounnas et al., 2021) | ✓ | | | | | ALG-FRA |
| (Hamed et al., 2022a) | ✓ | | | | | EGY-ENG |
| (Hamed et al., 2022c) | ✓ | | | | | EGY-ENG |
| (Hamed et al., 2023a) | ✓ | | | | | EGY-ENG |
| (Abdallah et al., 2024) | ✓ | | | | | TUN-ENG-FRA |
| (Heakl et al., 2024) | ✓ | | | | | EGY-ENG |
| (Al Ali and Aldarmaki, 2024) | ✓ | | | | | EMI-ENG |
| (Al-Ali, 2024) | ✓ | | | | | EMI-ENG |
| (Lachemat et al., 2024) | ✓ | | | | | ALG-ENG-FRA |
| (Barik and Lestari, 2019) | | | ✓ | | | ARA-IND |
| (Ugan et al., 2022) | | | ✓ | | | ARA-DEU |
| (Ali et al., 2021) | | | ✓ | | | ARA-ENG/FRA |
| (Kulkarni et al., 2023) | | | ✓ | | | ARA-ENG |
| (Hussein et al., 2023) | | | ✓ | | | ARA-ENG |
| (Hussein et al., 2024) | | | ✓ | | | ARA-ENG |
| (Kadaoui et al., 2024) | | | ✓ | | | ARA-ENG |
| (Chowdhury et al., 2020) | | | | ✓ | | MSA-EGY |
| (Chowdhury et al., 2021) | | | ✓ | ✓ | | MSA-EGY, ARA-ENG/FRA |
| (Mubarak et al., 2021) | | | ✓ | ✓ | | MSA-DA, ARA-ENG/FRA |
| (Abdelali et al., 2024) | | | ✓ | ✓ | | ARA-ENG/FRA, MSA-DA (DA: EGY, GLF, LEV, NOR) |
| Speech Translation | | | | | | |
| (Hamed et al., 2023a) | ✓ | | | | | EGY-ENG |
| (Hamed et al., 2022c) | ✓ | | | | | EGY-ENG |
| Word-level Language Identification | | | | | | |
| (Chowdhury et al., 2020) | | | | ✓ | | MSA-EGY |
| Sentence Boundary Detection | | | | | | |
| (Zribi et al., 2016) | | | | | ✓ | MSA-TUN-FRA |

Table 8: List of empirical papers for the stated speech-based tasks.

| | DA-Foreign | MSA-Foreign | ARA-Foreign | MSA-DA | MSA-DA-Foreign | Language Details |
|---|---|---|---|---|---|---|
| Word-level Language Identification | | | | | | |
| (Voss et al., 2014) | ✓ | | | | | MOR-ENG-FRA |
| (Cotterell et al., 2014) | ✓ | | | | | ALG-FRA |
| (Amazouz et al., 2017) | ✓ | | | | | ALG-FRA |
| (Amazouz et al., 2018) | ✓ | | | | | ALG-FRA |
| (Abainia, 2019) | ✓ | | | | | ALG-ENG/FRA |
| (Seddah et al., 2020) | ✓ | | | | | ALG-FRA |
| (Kalkman, 2024) | ✓ | | | | | MOR-NLD |
| (Darwish, 2014) | | | ✓ | | | ARA-ENG |
| (Sabty et al., 2021a) | | | ✓ | | | ARA-ENG |
| (Shehadi and Wintner, 2022) | | | ✓ | | | ARA-ENG-FRA |
| (Habash et al., 2008) | | | | ✓ | | MSA-DA |
| (Elfardy and Diab, 2012) | | | | ✓ | | MSA-EGY, MSA-LEV |
| (Solorio et al., 2014) | | | | ✓ | | MSA-EGY |
| (Molina et al., 2016) | | | | ✓ | | MSA-EGY |
| (Chowdhury et al., 2020) | | | | ✓ | | MSA-EGY |
| (Samih and Maier, 2016a) | | | | | ✓ | MSA-MOR-ENG/FRA/SPA/BER |
| (Samih and Maier, 2016b) | | | | | ✓ | MSA-MOR-FRA/SPA/BER |
| (Abidi and Smaïli, 2017) | | | | | ✓ | MSA-ALG-ENG-FRA |
| (Adouane and Dobnik, 2017) | | | | | ✓ | MSA-ALG-ENG-FRA-BER |
| (Diab et al., 2019) | | | | | ✓ | MSA-EGY-FOR |
| (Tarmom et al., 2020) | | | | | ✓ | MSA-EGY-SAU-ENG |
| (Hajbi et al., 2022) | | | | | ✓ | MSA-MOR-ENG/FRA/SPA |
| Machine Translation | | | | | | |
| (Bies et al., 2014) | ✓ | | | | | EGY-ENG |
| (Abainia, 2019) | ✓ | | | | | ALG-ENG/FRA |
| (Seddah et al., 2020) | ✓ | | | | | ALG-FRA |
| (Hamed et al., 2022c) | ✓ | | | | | EGY-ENG |
| (Nagoudi et al., 2023) | ✓ | | | | | ALG/MOR/TUN-FRA, EGY/JOR/PAL-ENG |
| (Menacer et al., 2019) | | ✓ | | | | MSA-ENG |
| (Elmadany et al., 2022) | ✓ | ✓ | | | | ALG-FRA, JOR-ENG, (synthetic) MSA-ENG/FRA |
| Transliteration | | | | | | |
| (Al-Badrashiny et al., 2014) | ✓ | | | | | EGY-ENG |
| (Bies et al., 2014) | ✓ | | | | | EGY-ENG |
| (Chen et al., 2017) | ✓ | | | | | EGY-FOR |
| (Abainia, 2019) | ✓ | | | | | ALG-ENG/FRA |
| (Touileb and Barnes, 2021) | ✓ | | | | | ALG-FRA |
| (Darwish, 2014) | | | ✓ | | | AR-ENG |
| (Alhazmi et al., 2024) | | | ✓ | | | ARA-FOR |
| Sentiment Analysis | | | | | | |
| (Tobaili et al., 2019) | ✓ | | | | | LEB-ENG |
| (Sherif and Sabty, 2024) | ✓ | | | | | EGY-ENG |
| (Almasah et al., 2023) | | | ✓ | | | ARA-FOR |
| (Mataoui et al., 2016) | | | | | ✓ | MSA-ALG-FRA |
| (Adouane et al., 2020) | | | | | ✓ | MSA-ALG-FRA-BER |
| (Touileb and Barnes, 2021) | ✓ | | | | | ALG-FRA |

Table 9: List of resource papers supporting the stated text-based tasks.

| | DA-Foreign | MSA-Foreign | ARA-Foreign | MSA-DA | MSA-DA-Foreign | Language Details |
|---|---|---|---|---|---|---|
| Named Entity Recognition | | | | | | |
| (Sabty et al., 2019a) | ✓ | ✓ | | | | MSA-ENG, EGY-ENG |
| (Sabty et al., 2019b) | ✓ | ✓ | | | | MSA-ENG, EGY-ENG |
| (Aguilar et al., 2018) | | | | ✓ | | MSA-EGY |
| Sentence-level Language Identification | | | | | | |
| (Shehadi and Wintner, 2022) | | | ✓ | | | ARA-ENG-FRA |
| (Altamimi et al., 2018) | | | | ✓ | | MSA-EGY/GLF/IRQ/LEV, MSA-MOR/TUN/LIB/ALG |
| (Chowdhury et al., 2020) | | | | ✓ | | MSA-EGY |
| Tokenization | | | | | | |
| (Balabel et al., 2020) | ✓ | | | | | EGY-ENG |
| (Seddah et al., 2020) | ✓ | | | | | ALG-FRA |
| (Gaser et al., 2023) | ✓ | | | | | EGY-ENG |
| Part-of-speech Tagging | | | | | | |
| (Balabel et al., 2020) | ✓ | | | | | EGY-ENG |
| (Seddah et al., 2020) | ✓ | | | | | ALG-FRA |
| (Diab et al., 2019) | | | | | ✓ | MSA-EGY-FOR |
| Abusive Language Detection | | | | | | |
| (Abainia, 2019) | ✓ | | | | | ALG-ENG/FRA |
| (Guellil et al., 2021) | ✓ | | | | | ALG-ENG/FRA |
| (Alhazmi et al., 2024) | | | ✓ | | | ARA-FOR |
| Sentence-level Micro-Dialect Identification | | | | | | |
| (Abainia, 2018) | ✓ | | | | | ALG-FRA |
| (Abdul-Mageed et al., 2020) | | | ✓ | | | ARA-FOR |
| Dialectness Level Estimation | | | | | | |
| (Habash et al., 2008) | | | | ✓ | | MSA-DA |
| (Hamed et al., 2024) | | | | | ✓ | MSA-EMI/EGY-ENG |
| Spelling Correction and Text Normalization | | | | | | |
| (Adouane et al., 2019) | | | | | ✓ | MSA-ALG-ENG-FRA-BER |
| Dependency Parsing | | | | | | |
| (Seddah et al., 2020) | ✓ | | | | | ALG-FRA |
| Fake News Detection | | | | | | |
| (Smaïli et al., 2024) | | | | | ✓ | MSA-TUN/ALG-ENG-FRA |
| Word Analogy | | | | | | |
| (Aghzal and Mourhir, 2021) | ✓ | | | | | MOR-ENG-FRA |
| Lemmzatization | | | | | | |
| (Balabel et al., 2020) | ✓ | | | | | EGY-ENG |
| Emotion Detection | | | | | | |
| (Abainia, 2019) | ✓ | | | | | ALG-ENG/FRA |
| Topic Modeling | | | | | | |
| (Touileb and Barnes, 2021) | ✓ | | | | | ALG-FRA |
| Text Corpora (without annotations) | | | | | | |
| (Hamed et al., 2019) | ✓ | | | | | EGY-ENG |
| (Aghzal and Mourhir, 2021) | ✓ | | | | | MOR-ENG-FRA |
| (Hamed et al., 2017) | | ✓ | | | | MSA-ENG |
| (Adouane et al., 2018b) | | | | | ✓ | MSA-ALG-ENG-FRA-BER |

Table 10: List of resource papers supporting the stated text-based tasks.

| | DA-Foreign | MSA-Foreign | ARA-Foreign | MSA-DA | MSA-DA-Foreign | Language Details |
|---|---|---|---|---|---|---|
| Automatic Speech Recognition | | | | | | |
| (Ismail, 2015) | ✓ | | | | | SAU-ENG |
| (Amazouz et al., 2017) | ✓ | | | | | ALG-FRA |
| (Amazouz et al., 2018) | ✓ | | | | | ALG-FRA |
| (Hamed et al., 2018a) | ✓ | | | | | EGY-ENG |
| (Elfahal, 2019) | ✓ | | | | | SUD-ENG |
| (Hamed et al., 2020) | ✓ | | | | | EGY-ENG |
| (Lounnas et al., 2021) | ✓ | | | | | ALG-FRA |
| (Hamed et al., 2022c) | ✓ | | | | | EGY-ENG |
| (Abdallah et al., 2024) | ✓ | | | | | TUN-ENG-FRA |
| (Al Ali and Aldarmaki, 2024) | ✓ | | | | | EMI-ENG |
| (Lachemat et al., 2024) | ✓ | | | | | ALG-ENG-FRA |
| (Ali et al., 2021) | | | ✓ | | | ARA-ENG/FRA |
| (Ugan et al., 2022) | | | ✓ | | | ARA-DEU |
| (Chowdhury et al., 2020) | | | | ✓ | | MSA-EGY |
| (Mubarak et al., 2021) | | | ✓ | ✓ | | MSA-DA, ARA-ENG/FRA |
| (Zribi et al., 2015) | | | | | ✓ | MSA-TUN-FRA |
| (Hamed et al., 2024) | | | | | ✓ | MSA-EMI/EGY-ENG |
| Word-level Language Identification | | | | | | |
| (Mohdeb-Amazouz et al., 2016) | ✓ | | | | | ALG/TUN/MOR-FRA |
| (Amazouz et al., 2017) | ✓ | | | | | ALG-FRA |
| (Amazouz et al., 2018) | ✓ | | | | | ALG-FRA |
| (Chowdhury et al., 2020) | | | | ✓ | | MSA-EGY |
| Sentence-level Language Identification | | | | | | |
| (Chowdhury et al., 2020) | | | | ✓ | | MSA-EGY |
| Speech Translation | | | | | | |
| (Hamed et al., 2022c) | ✓ | | | | | EGY-ENG |

Table 11: List of resource papers supporting the stated speech-based tasks.