# TARIC-SLU: A Tunisian Benchmark Dataset For Spoken Language Understanding

**Salima Mdhaffar**[1,5], **Fethi Bougares**[2], **Renato De Mori**[1,3], **Salah Zaiem**[4], **Mirco Ravanelli**[5], **Yannick Estève**[1]

[1] LIA, Avignon University, France
[2] Elyadata, Paris, France
[3] McGill, Montréal, Canada
[4] LTCI, Télécom Paris, Institut Polytechnique de Paris, France
[5] Mila-Quebec AI Institute, Université de Montréal, Concordia University, Canada

salima.mdhaffar@univ-avignon.fr,fethi.bougares@elyadata.com

**Abstract**

In recent years, there has been a significant increase in interest in developing Spoken Language Understanding (SLU) systems. SLU involves extracting a list of semantic information from the speech signal. A major issue for SLU systems is the lack of sufficient amount of bi-modal (audio and textual semantic annotation) training data. Existing SLU resources are mainly available in high-resource languages such as English, Mandarin and French. However, one of the current challenges concerning low-resourced languages is data collection and annotation. In this work, we present a new freely available corpus, named TARIC-SLU, composed of railway transport conversations in Tunisian dialect that is continuously annotated in dialogue acts and slots. We describe the semantic model of the dataset, the data and experiments conducted to build ASR-based and SLU-based baseline models. To facilitate its use, a complete recipe, including data preparation, training and evaluation scripts, has been built and will be integrated into SpeechBrain, a popular open-source conversational AI toolkit based on PyTorch.

**Keywords:** Tunisian Dialect, TARIC, Spoken Language Understanding, low-resource languages, dataset

## 1. Introduction

In recent years, the field of natural language processing has experienced significant advancements in developing datasets and models to enhance spoken language understanding (SLU) systems (Bastianelli et al., 2020; Lugosch et al., 2021; Tomasello et al., 2023). However, many of these resources predominantly concentrate on widely spoken languages, resulting in an insufficient focus on dialectal and minority languages. Within the context of the Tunisian dialect, a language replete with intricate linguistic nuances, limited resources have hindered the progress of voice-based applications and services for the Tunisian population. This paper introduces an innovative and exhaustive dataset of spoken Tunisian dialect, meticulously annotated with semantic information, to address this crucial gap and promote advancements in SLU technology for Tunisian Arabic.

In the context of a conventional dialogue system, information is typically represented through a semantic frame structure (Tur and Mori, 2011). For each utterance, constructing the semantic representation primarily involves (i) classifying the user's utterance in terms of 'speech acts' (SA) (Searle, 1969) or 'intents' and (ii) slot filling (Wang et al., 2005). Slot filling is a natural language processing and information extraction technique that entails the identification and extraction of specific pieces of information or attributes, referred to as 'slots,' from unstructured text or spoken language. These slots are typically associated with predefined categories or entities and find common usage in various NLP applications, including chatbots, virtual assistants, question-answering systems, and information retrieval. The process of annotating data for training and evaluating systems or models designed to tackle these tasks is often arduous and time-consuming.

This paper introduces the TARIC-SLU dataset, marking a significant milestone as the first publicly accessible dataset tailored for the field of Tunisian Spoken Language Understanding (SLU). The TARIC-SLU dataset is a comprehensive compilation encompassing six hours of agent-client interactions, meticulously annotated with dual-tiered semantic content, focusing on dialogue acts and underlying concepts.

The primary contributions of this paper can be succinctly summarized as follows:

- The release of the TARIC-SLU corpus[1], representing the very first resource of its kind designed for SLU within the Tunisian dialect, and its availability for the broader research community.

[1] https://github.com/elyadata/TARIC-SLU



*LREC-COLING 2024*, pages 15606–15616
20-25 May, 2024. 

- A comprehensive exposition of the semantic representations and annotation methodologies employed in the construction of the TARIC-SLU corpus.
- The release of an open-source SpeechBrain recipe [2], following the framework proposed by (Ravanelli et al., 2021), specifically tailored for the training and evaluation of an end-to-end neural model, purpose-built for the semantic tasks inherent to the TARIC-SLU dataset.
- Additionally, this paper provides baseline results for Automatic Speech Recognition (ASR), Natural Language Understanding (NLU), and Spoken Language Understanding (SLU) tasks, all derived from the TARIC-SLU corpus.

The subsequent sections of this paper are organized as follows. Section 2 presents an in-depth examination of the prior works and relevant literature in the field. In Section 3, we delve into the specific characteristics and idiosyncrasies associated with the Tunisian dialect, highlighting its distinctive attributes. Section 4 expounds on the data collection procedures and the intricacies involved in the annotation process for the TARIC-SLU corpus. Section 5 presents the baseline results.

## 2. Related works

The importance of spoken language understanding (SLU) tasks has grown steadily with the increase in the number of virtual assistant users. The majority of SLU systems are predominantly developed for the English language, leading to a high quality of SLU services in this language. The availability of English SLU systems has been driven, among other factors, by the availability of annotated resources such as ATIS (Hemphill et al., 1990), SNIPS (Coucke et al., 2018), Spoken SQuAD (Li et al., 2018), MultiWOZ (Budzianowski et al., 2018), the Fluent Speech Command (FSC) corpus (Lugosch et al., 2019), SLURP (Bastianelli et al., 2020), Timers and such (Lugosch et al., 2021), STOP (Tomasello et al., 2023), etc.

In contrast to English, SLU systems in other languages lag behind in performance, especially for low-resourced languages.

Recently, there has been a considerable amount of work and effort to collect resources and develop systems for low-resourced languages. The research community's interest in this area is evident through the *Low Resource Spoken Language Understanding* special sessions held in the last two editions of the Interspeech conference.

[2] https://github.com/speechbrain/speechbrain/tree/develop/recipes/TARIC

Notably, the number of freely accessible Arabic datasets for SLU is very limited in number, size, availability and dialects coverage.

Table 1 offers an overview on the existing datasets in Arabic languages, including Modern Standard Arabic (MSA), Algerian dialect (ALG), Egyptian dialect (EGY), and Tunisian dialect (TUN). Unfortunately, most of these datasets are not publicly accessible to the research community. For the Tunisian dialect, an attempt to create an SLU dataset is discussed in (Graja et al., 2013), but this dataset remains non-public. Table 1 also presents two Spanish (SPA) datasets, DIHANA (Alcácer et al., 2005; Benedı et al., 2006) and Basurde (Trias-Sanz and Marino, 2002), both related to train ticket reservation dialogues, similar to the TARIC-SLU dataset. To the best of our knowledge, these are the only other accessible datasets in the domain of train ticket reservation.

The TARIC-SLU dataset is annotated to cover two essential SLU tasks: (1) slot filling detection and (2) Speech Act classification. From a technical perspective, there are two main approaches for addressing the slot filling detection task in the context of speech: (1) the pipeline approach and (2) the end-to-end approach. In the pipeline approach, the initial module is an automatic speech recognition system (ASR) responsible for converting the input speech signals into transcriptions. Subsequently, a second module, the natural language understanding module (NLU), generates a list of semantic concept hypotheses from the transcribed user's utterrance (Béchet et al., 2004; Ghannay et al., 2021; Mdhaffar et al., 2022b).

Recent advancements in the design of dialogue systems increasingly favor replacing the traditional architecture with end-to-end architectures that rely on deep neural networks. In these architectures, semantic concepts are directly generated from the user's speech signal. These end-to-end deep learning approaches have demonstrated strong performance by benefiting from the joint optimization of speech transcription and semantic tagging, and from the limitation of error propagation between the automatic speech recognition (ASR) and natural language understanding (NLU) modules (Haghani et al., 2018; Ghannay et al., 2018; Dinarelli et al., 2020; Tomashenko et al., 2020).

For speech act classification, the task can be considered as an utterance classification problem. The classifier is built by extracting low level acoustic speech parameters and applying a machine learning algorithm (Deb et al., 2023; Adiani et al., 2023). Given the inherent connection between the speech classification task and the slot filling task, recent works in the literature jointly implement these two tasks within a single model (Arora et al., 2023; Zhang and Wang, 2016; Liu and Lane,

2016).

Regardless of the SLU task, effectively using the described techniques requires annotated audio data which are domain and language dependent. This data is hardly available and collecting new datasets is expensive and time consuming. This constraint limits the development of such systems across multiple languages and tasks.

## 3. The Tunisian dialect

### 3.1. Generalities about the Tunisian Dialect

More than 400 million people use Arabic language in 25 countries where Arabic is the official language[3]. Arabic is classified into three categories (Guellil et al., 2021):

- Classical Arabic (CA), refers to the form of the Arabic language in which the literary texts and Quran and the holy book of Islam are written.
- Modern Standard Arabic (MSA) serves as the standardized formal language used for written formal communication and education.
- Dialectal Arabic (DA) is used in daily spoken communication, informal exchanges, etc. DA mostly divided into six main groups: (1) Egyptian, (2) Levantine, (3) Gulf, (4) Iraqi, (5) Maghrebi and (6) Others containing the remaining dialect (Zaidan and Callison-Burch, 2014). These dialects can vary significantly in terms of pronunciation, vocabulary, and grammar. They often reflect the historical and cultural diversity of the Arab world.

Tunisian Dialect (TUN), known as “Tounsi” or “Derja”, is a part of the Maghrebi dialects. It is the primary spoken language in Tunisia and is used extensively in various media platforms, including television, social networks and radio. Tunisian dialect Arabic lexicon contains words from Tamazight, French, Turkish, Italian and other languages.

### 3.2. Specificities of the Tunisian dialect

Tunisian dialect’s specificities present challenges and opportunities for natural language processing systems. Here are some considerations when dealing with the Tunisian dialect in speech processing:

**(1) Lack of Orthography Standardization** Unlike MSA, which has a standardized form for writing and media, Tunisian Arabic lacks such standardization.

[3] https://worldpopulationreview.com/country-rankings/arabic-speaking-countries

**(2) Intra-sentential code-switching** which includes code mixing within a phrase, a clause or a sentence boundary. Tunisian native speakers use Tunisian as well as French and English in a single conversation. For instance, “première classe زُوزْ بْلَايِصْ” (two places first class) is composed from the Arabic expression ” زُوزْ بْلَايِصْ ” (two places in English) and the French expression ”première classe” (first class in English).

**(3) Intra-word code-switching** which is the code mixing within a single word. For instance, Tunisian native speakers frequently mix elements (e.g., a root and an affix/suffix) from different languages within a single word. An example from TARIC-SLU dataset is the word “رَازَارْفِيلِي”. This word comprises a root from the French word “réserver” (to book in English) and the Tunisian dialect possessive determiner suffix.

**(4) Derivatives of the Tunisian dialect:** The Tunisian dialect, like many Arabic dialects, has several derivatives or regional variations that differ in vocabulary, pronunciation, and some grammatical aspects (Gibson, 1999).

**(5) Non-conventional syntactic structures** Speech does not follow, in certain cases, well-formed, canonical syntactic structures. For instance there is no clear agreement of TUN syntactic phenomena such as clause structure, word order or subject-verb agreement.

**(6) Grammatical gender reversals** In the context of TUN, there are instances where the gender of foreign or code-switched words is reversed. For example, the French word ’ticket’ which is masculine in its original language, becomes feminine when used in TUN.

In summary, the specificities of the Tunisian dialect present several challenges for natural language processing systems, particularly in the areas of text and speech processing, language understanding, and translation. Overcoming these challenges requires the development of dialect-specific models, resources, and tools, as well as extensive data collection and adaptation efforts.

## 4. TARIC-SLU

Our dataset for SLU is sourced from the TARIC dataset (Masmoudi et al., 2014). TARIC dataset was dedicated to training and evaluating Tunisian Dialect Automatic Speech Recognition in the context of human-to-human dialogues for train reservation task. In this section, we describe the orig-

| Dataset | Domain | #Utt | #Spk | #Duration | Avail. | Lang | task | | recipe |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | SA | C/V | |
| (Lhioui et al., 2013) | Multi | 140 | 10 | n/a | ✗ | MSA | ✗ | C | ✗ |
| (Lichouri et al., 2022) | Univ | 508 | 4 | 24min | ✗ | MSA | ✓ | ✗ | ✗ |
| (Lichouri et al., 2023) | voice | 1000 | 10 | n/a | ✗ | ALG | ✓ | ✗ | ✗ |
| (Elmadany et al., 2015) | Multi | 4727 | n/a | 2h | ✓ | EGY | ✗ | ✓ | ✗ |
| (Dbabis et al., 2012) | News | 4727 | n/a | 2h | ✗ | MSA | ✓ | ✗ | ✗ |
| TuDiCoI | Train | 6533 | n/a | n/a | ✗ | TUN | ✗ | ✓ | ✗ |
| **TARIC-SLU** | Train | 17816 | n/a | 8h | ✓ | TUN | ✓ | ✓ | ✓ |
| DIHANA | Train | 6278 | 225 | 5.5h | ✓ | SPA | ✓ | ✓ | ✗ |
| Basurde | Train | 565 | n/a | n/a | ✗ | SPA | ✓ | ✗ | ✗ |

Table 1: Overview of existing SLU datasets. "Utt" denotes utterance. "Spk" denotes speaker. "Avail" denotes if the dataset is available for the research community. "Lang" denotes language. "SA" denotes Speech Act. C/V denotes Concepts and Values.

inal TARIC dataset, we present our semantic labelling definitions, and we detail the undertaken annotation process.

### 4.1. TARIC dataset

The acquisition of the TARIC dataset was carried out in some train stations in Tunisia. During TARIC corpus analysis, a large part of its recordings were unusable because of the poor acoustic conditions. To remedy this, these dialogues were manually transcribed and acted by native speakers. The clean subset of TARIC was kept to allow for the testing using real acoustic conditions. Overall, TARIC dataset comprises 4,000 oral dialogue recordings from 108 speakers along with their manual transcriptions. Transcribing and annotating TARIC are based on the annotation conventions of CODA[4] (Conventional Orthography for Dialectal Arabic) (Habash et al., 2012). The audio files of the TARIC dataset are supposed to be available for online download, which is not the case at the moment. The present work was undertaken with the desire, among other things, to distribute this dataset to advance future research in Tunisian Dialect processing and enable a fair comparison across papers and systems. Compared to its initial version, the newly distributed dataset will include :

**(1) Dialogue separation:** the new distribution of the data will include the information about the beginning and the end of each dialogue (which is not the case in the ASR TARIC dataset).

**(2) Better Train-Dev-Test split:** unlike the initial splitting, we propose a new partitioning that take into consideration the significant portion of acted dialogues. The proposed split pays attention as well to speakers distribution between training, dev and test sets.

**(3) Semantic labelling :** TARIC dataset will be augmented with a semantic annotation. As detailed below, semantic annotation includes labelling with speech acts and slots/values.

### 4.2. Semantic labelling

The semantic annotation of a corpus requires the definition of a semantic representation adapted to the application domain. The semantic representation is dependent on the targeted task of the software application.

We annotated the TARIC-SLU dataset with two levels of labels: (1) speech act and (2) slot-value labels.

**Speech act** TARIC-SLU has been meticulously annotated with three distinct speech acts, signifying a comprehensive understanding of the main communicative intentions within the text.

1. **Directive query** is used when a user asks for information or makes a request. Common examples include asking questions, giving orders, and offering advice.

2. **Directive answer** is the speech act where the speaker responds to a query or request for information by providing a specific answer or solution.

3. **Politeness** represents greetings at the beginning or during the conversation. It also includes apologizing, congratulating, thanking, commiserating, and expressing gratitude or good wishes.

**Slots labels** The slot label annotation scheme aligns with common principles employed in other annotated speech corpora, that encompasses

[4]CODA is designed to develop computational models of Arabic dialects. First, it is defined for Egyptian Arabic and then extended to other Arabic dialects.

- Age
- Age_request
- Age_ticket
- An
- Answer
- Arrival_time
- Card_type
- Card_price
- City_name_arrival
- City_name_departure
- City_name_before
- City_name_direction
- Class_number
- Class_type
- Command_task
- Comparative_age
- Comparative_distance
- Comparative_price
- Comparative_time
- Coreference_city
- Coreference_departure
- Date
- Day
- Departure_time
- Discount_value
- Discount_percent
- Duration_request
- Duration
- Existence
- Existence_request
- Hour_request
- Money_exchange
- Month
- Negation
- Number
- Number_of_train
- Number_request
- Object
- Option
- Other_transport
- Part_price
- Part_time
- Period_day
- Period_year
- Person_name
- Price_request
- Rank
- Reference_object
- Reference_person
- Reference_time
- Relative_day
- Relative_time
- State
- Tarif
- Task
- Ticket_number
- Ticket_price
- Ticket_type
- Time
- Train_type

Figure 1: Slots labels used to annotate the TARIC-SLU dataset

finer task-specific details (Bonneau-Maynard et al., 2006).

The semantic representation employs a slot-value structure. A semantic segment is represented by a pair which contains the name of the slot and a sequence of words considered as the value to be assigned to the slot. The slot name represents the meaning of the sequence of words. The proposed annotation scheme considers 60 slots listed in Figure 1.

### 4.3. Annotation process

The annotation guidelines were developed through an iterative process. Initially, a draft was created, encompassing general guidelines and challenging scenario examples. Two annotators employed these guidelines to annotate a first part. The resulting annotations were compared to the desired annotations, highlighting issues and inconsistencies. These findings were then used to refine the guidelines. The final version of the guidelines created as part of this work will be distributed to the research community.

Based on the semantic representation described above and the guidelines, the semantic annotation of the TARIC-SLU corpus has been performed by three annotators.

Table 2 presents an example of annotation for one dialogue between an agent and a client.

### 4.4. Inter-annotator agreement

During the semantic annotation process, Inter-annotator agreements are measured for quality assurance. For this purpose, we randomly selected several sets of dialogues that were annotated twice by two annotators. The annotators were unaware that certain dialogues had been assigned to different individuals to assess agreement reliability.

Speech act annotations in SLU typically occur at the utterance level, meaning the entire user query or statement is assigned a single speech act label representing the user's primary goal. We can directly use straightforward inter-annotator agreement measures, such as Cohen's Kappa.

Slot annotations in SLU involve identifying and labeling specific slots within the user's utterance. These slots can have various values and categories, and their boundaries need to be precisely defined. To compute agreement annotations, there's a must transition from the utterance level to a more granular token level. This transition is achieved by employing the BIO (Begin, Inside, Outside) labeling format proposed by (Ramshaw and Marcus, 1999). The agreement for slot annotations has been computed at the token level. An example of BIO annotation is illustrated in Figure 2.

Table 3 presents the agreement scores. The inter-annotator Kappa for this set is 0.73 for speech acts and 0.8 for slots. We report also the kappa score for slots without the label 'O'.

### 4.5. Dataset statistics

A part of the TARIC dataset (8 hours of speech) has been annotated as described in previous sections. Table 4 provides an overview of the dataset, including the total number of dialogs, the average number of utterances per dialog, the average number of tokens per dialog and the average number of slots per dialog.

## 5. Tasks and Evaluation

In this work, we address two important problems of SLU, (1) speech act classification and (2) slot filling, following the annotation described in section 4.2. Those two problems can be approached using one model for each, referred to

| | dialogue | SA | slot<value> |
|---|---|---|---|
| A | TUN: سَلَامُ عَلَيْكُمْ<br>Buckwalter: salaAmu Ealayokumo<br>English: Hello | P | - |
| C | TUN: أَعْطِينِي زُوزْ تِكَايَاتْ صْفَاقِسْ قَابِسْ<br>Buckwalter: >EoTiyniy zuwzo tikaAyaAto SofaAqiso GAbiso<br>English: please give me two tickets from Sfax to Gabes | Q | command_task<give_me><br>number_of_tickets<two><br>object<ticket><br>city_name_departure<Sfax><br>city_name_arrival<Gabes> |
| A | TUN: قَابِسْ<br>Buckwalter: GAbiso<br>EN: Gabes | A | city_name_arrival<Gabes> |
| C | TUN: قَدَّاشْ مَزَالْتُو وَقْتْ<br>Buckwalter: qad∼aA$o mazaAtoluw waqoto<br>EN: How much time left | Q | duration_req<how much time> |
| A | TUN: الأَرْبْعَةْ وُ نُصْفْ يُخْرِجْ<br>Buckwalter: Al>aroboEapo wu nuSofo yuxorijo<br>EN: At four and half the departure | A | departure_time<four and half> |
| C | TUN: بْقَدَّاشْ التِكَايْ<br>Buckwalter: boqad∼aA$o AltikaAyo<br>EN: How much the ticket | Q | price_req<how much><br>object<ticket> |
| A | TUN: سَبْعَةْ آلَافْ<br>Buckwalter: saboEapo laAfo<br>EN: Seven dinars | A | price_ticket <seven dinars > |

Table 2: Example of conversation between a client (C) and an agent (A) from TARIC-SLU dataset. The third column presents the semantic annotation slot/value. Column SA represents the Speech Acts, which are either Q (Query), A (Answer) or P (Politeness). Buckwalter transliteration of TUN and English translation are provided.

(a) سَلَامُ عَلَيْكُمْ أَعْطِينِي زُوزْ تِكَايَاتْ أَلَايْ رُتُورْ صْفَاقِسْ بِيرْ بُورَقْبَةْ

(b) English: Hello give me two tickets two ways from Sfax to Bir Bouregba, Buckwalter: salaAmu Ealayokumo >aEoTiyniy zuwzo tikaAyaAto >alaAyo rutuwro SofaAqiso biyro buwraqobapo"

(c) salaAmu Ealayokumo **<command_task** >aEoTiyniy **>** **<number_of_tickets** zuwzo **>**
**<object** tikaAyaAto **>** **<ticket_type** >alaAyo rutuwro **>**
**<city_name_departure** SofaAqiso **>** **<city_name_arrival** biyro buwraqobapo **>**

(d) (salaAmu,O), (Ealayokumo,O), (>aEoTiyniy,B_command_task),
(zuwz,B_number_tickets),(tikaAyaAt,B_object),(>alaAyo,B_ticket_type),
(rutuwro,I_ticket_type), (SofaAqiso,B_city_name_departure),
(biyro,B_city_name_arrival) , (buwraqobapo,I_city_name_arrival)

Figure 2: An example of a TARIC-SLU dataset sample. (a) corresponds to the transcribed sentence. (b) the same sample with English translation and Buckwalter transliteration (c) the same sample with its additional semantic tags. (d) the same sample with the BIO model, we represent the sentence by a sequence of pairs (w=word,l=label). Here, '**<command_task** ' is an opening tag starting the support word sequence '>aEoTiyniy' and expressing that this word sequence is associated with the *command_task* semantic concept. The character '*>*' represents the closing tag and it is used to close all concept tags.

as the "pipeline approach", or one model for both, commonly known as "end-to-end" approach.

In this section, we offer baseline results from various models trained on the TARIC-SLU annotated corpus. We first outline the data division into training, development, and test sets. Subsequently, we share baseline results we got for both ASR and SLU tasks from the TARIC-SLU corpus to show how complex the tasks are and to create a standard for future research that uses this corpus. The ASR task is evaluated through the

| | Kappa |
|---|---|
| Speech Acts | 0.73 |
| Slots | 0.8 |
| Slots without the label 'O' | 0.74 |

Table 3: Kappa values

| Type | Count |
|---|---|
| #Dialogs | 2043 |
| #Utterences | 17816 |
| Avg. # of utterances per dialog | 10.17 |
| Avg. # of tokens per dialog | 143.63 |
| Avg. # of slots per dialog | 23.54 |

Table 4: TARIC-SLU corpus statistics

classical word error rate (WER). The SLU speech act recognition is evaluated in terms of speech act error rate (SAER), which is a standard classification error rate. Lastly, the SLU slot filling task is evaluated in terms of concept error rate (COER). COER is computed similarly to WER by taking into account only the semantic labels in the reference and hypothesis annotations and is used for both SLU and named entity recognition (Ghannay et al., 2021; Mdhaffar et al., 2022a). For the calculation of this metric, we have drawn on the detailed description in (Laperrière et al., 2022).

### 5.1. Data Partition

As presented earlier, TARIC-SLU dataset has a total of 2043 dialogues. We used 103 dialogues for the development set, 173 dialogues for the test set and the rest 1767 dialogues for the training set. Given that the TARIC data set is a mix of a real conditions and acted dialogues (see section 4.1 for more details), we partitioned the TARIC data set in such a way that all the real conditions Agent-Client exchanges are split over dev and test sets. Table 5 gives more details about speakers and gender distribution over the three sets.

| | Train | Dev | Test |
|---|---|---|---|
| #utterance | 15751 | 771 | 1294 |
| #speaker | 5 | 39 | 38 |
| #speakers not in Train | - | 37 | 35 |
| #male | 2 | 23 | 24 |
| #female | 3 | 16 | 14 |

Table 5: TARIC-SLU data set split into Train, Dev and Test

### 5.2. ASR Task

Speech Recognition is the first stage of the pipeline approach for SLU. The output texts of this stage are intended to be used as input in the next pipeline stage, namely the NLU module.

The chosen ASR is an end-to-end Wav2Vec2 multilingual model, developed using self-supervised learning in 128 languages, following previous approaches in Tunisian ASR (Abdallah et al., 2023). In addition to the large Wav2Vec2 model, we incorporate an extra layer with 1024 neurons and LeakyReLU as the activation function, followed by a fully-connected layer and a final $40$-dimensional softmax layer, each dimension corresponding to a character. This neural network architecture comprises a total of $316.5$ million trainable parameters. The weights of the two added layers were randomly initialized, while the weights of the Wav2Vec2 part of the neural architecture were initialized using the pre-trained weights. Finally, the convolutional front-end weights are frozen, following common practice (Baevski et al., 2020).

For fine-tuning this self-supervised model with the additional layers, we employed a batch size of 6 samples, distributed across 4 NVIDIA V100 32GB GPU cards. We utilized two optimizers: Adadelta for updating the additional layers' weights and Adam for fine-tuning the self-supervised learning (SSL) model. The initial learning rate for Adadelta was $1.0$, while for Adam, this value was $0.0001$. The maximum number of epochs was chosen to be 100: the best model on the validation is obtained at epoch 83, with a word error rate of 37.87%. During the fine-tuning process, a SpecAugmentation data augmentation technique was applied to the audio signal (Park et al., 2019).

The performance of this system is illustrated in Table 6, showing the difficulty of the task, especially on the development corpus.

| | Dev | Test |
|---|---|---|
| WER | 37.87 | 30.06 |

Table 6: Error Rate of TARIC-SLU dev and test data obtained with the ASR model trained in the context of the pipeline-based SLU approach

### 5.3. SLU Tasks

**Pipeline SLU Task** SLU is typically performed through a cascading of an Automatic Speech Recognition (ASR) and Natural Language Understanding (NLU) module. Even though, it is well known that errors at the ASR stage have a negative impact on the NLU performance, we decided to start by following this conventional pipeline approach. Pipeline-based SLU system will also help us to assess the performance of the NLU component by feeding it with the reference

| Type of Cascade system | COER Dev | COER Test | SAER Dev | SAER Test |
|---|---|---|---|---|
| NLU Multilingual Bert "Gold transcription" | 20.9 | 23.8 | 15.6 | 19.7 |
| NLU Multilingual Bert "Automatic transcription" | 34.1 | 32.8 | 22.4 | 21.2 |

Table 7: Pipeline results (%)

| | COER Dev | COER Test | SAER Dev | SAER Test |
|---|---|---|---|---|
| End-to_End based on XLS-R wav2vec2.0 | 35.62 | 31.23 | 24.64 | 20.9 |

Table 8: E2E results (%)

transcription. The ASR used in this SLU pipeline approach is the one described in the previous section.

Speech act classification and slot filling are jointly trained using a multi-task deep neural network architecture. This join learning allows to benefit from the correlation between the two tasks. To encode our input tokens, we use a pre-trained language model, trained with a multilingual dataset (mBert[5]) (Devlin et al., 2018). Our architecture consists of a bi-LSTM layer on top of which we train a bi-LSTM-CRF for slot tagging and a bi-LSTM for the speech act classification. We use 200 hidden units for all three bi-LSTM layers in our architecture. Parameter optimization is performed with Adam optimizer with a learning rate of 0.0005. The maximum number of epochs was chosen to be 50.

The results are provided in table 7. We can notice the impact of ASR errors on the model performance on the slot filling task, which is limited for speech act classification.

**End-to-End SLU Task** Unlike the pipeline approach, here we train one system to resolve the SLU tasks given the speech segments. That is, no longer rely on an ASR module's transcription outputs. To do so, we added a linear layer on top of pretrained XLS-R 128 wav2vec2.0 model (Baevski et al., 2020) and fine-tuned it using the TARIC-SLU training set with a character-level CTC loss function (Graves et al., 2006). We formulate the End-to-End SLU task as character level prediction where slots are delimited by tag-specific special characters as in (Yadav et al., 2020; Ghannay et al., 2018). We also added the speech act token to the reference annotation as the first token of each sequence of words. In this way, the end-to-end learns to both classify the utterances in terms of speech act, and recognize slot/value pairs present in the speech segment. As input, the neural network receives a wav audio file, and the output is a transcription enriched with semantic labels and speech acts. After processing through the softmax layer (which have the size of 104[6]), the outputs are generated by a simple greedy decoder. Similar to the ASR training, we employed a batch size of 6 samples, distributed across 4 NVIDIA V100 32GB GPU cards. We utilized two optimizers: Adadelta for updating the additional layers' weights and Adam for fine-tuning the self-supervised learning (SSL) model. The initial learning rate for Adadelta was $1.0$, while for Adam, this value was $0.0001$. The maximum number of epochs was chosen to be 100.

End-to-end results are reported in table 8. The Concept Error rate is 35.62% for the development set and 31.23% for the test set. The Error rate for speech classification is 24.64% for the development set and 20.9% for the test set. This study concludes that the two architectures (pipeline vs. end-to-end) remain valid and competitive using TARIC-SLU dataset.

## 6. Conclusion

This paper introduces the TARIC-SLU corpus, a valuable resource for the Tunisian dialect. It also presents baseline results for automatic speech recognition, speech act recognition, and slot-filling tasks. The release of this corpus, recipes and the baseline results provide a foundation for further research and development in the field of spoken language understanding in Tunisian dialect. We believe that this resource, along with the baseline results provided in this paper, will foster collaborations and innovations that contribute to the broader goal of improving human-computer interaction in the Tunisian dialect and other low-resourced languages.

[5]trained on the top 104 languages with the largest Wikipedia using a masked language modeling objective

[6]40 characters that cover the alphabet of TARIC dataset, 60 characters for slots, one character for closing slots, three characters for speech acts

## 7. Acknowledgrments

This work was performed using HPC resources from GENCI-IDRIS (grant AD011012551R2) and received funding from the EU H2020 SELMA (grant No 957017) and ESPERANTO research and innovation program under the Marie Skłodowska-Curie (grant No 101007666).

## 8. Bibliographical References

Ahmed Amine Ben Abdallah, Ata Kabboudi, Amir Kanoun, and Salah Zaiem. 2023. Leveraging data collection and unsupervised learning for code-switched tunisian arabic automatic speech recognition. *arXiv preprint arXiv:2309.11327*.

Deeksha Adiani, Kelley Colopietro, Joshua Wade, Miroslava Migovich, Timothy J Vogus, and Nilanjan Sarkar. 2023. Dialogue act classification via transfer learning for automated labeling of interviewee responses in virtual reality job interview training platforms for autistic individuals. *Signals*, 4(2):359–380.

N Alcácer, JM Benedı, F Blat, R Granell, CD Martınez, and F Torres. 2005. Acquisition and labelling of a spontaneous speech dialogue corpus. In *Proceeding of 10th International Conference on Speech and Computer (SPECOM). Patras, Greece*, pages 583–586.

Siddhant Arora, Hayato Futami, Emiru Tsunoo, Brian Yan, and Shinji Watanabe. 2023. Joint modelling of spoken language understanding tasks with integrated dialog history. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Emanuele Bastianelli, Andrea Vanzo, Pawel Swietojanski, and Verena Rieser. 2020. Slurp: A spoken language understanding resource package. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7252–7262.

Frédéric Béchet, Allen L Gorin, Jeremy H Wright, and Dilek Hakkani Tür. 2004. Detecting and extracting named entities from spontaneous speech in a mixed-initiative spoken dialogue context: How may i help you? sm, tm. *Speech Communication*, 42(2):207–225.

José-Miguel Benedı, Eduardo Lleida, Amparo Varona, Marıa-José Castro, Isabel Galiano, Raquel Justo, I López, and Antonio Miguel. 2006. Design and acquisition of a telephone spontaneous speech dialogue corpus in spanish: Dihana. In *Fifth International Conference on Language Resources and Evaluation (LREC)*, pages 1636–1639.

Hélène Bonneau-Maynard, Christelle Ayache, Frédéric Bechet, Alexandre Denis, Anne Kuhn, Fabrice Lefèvre, Djamel Mostefa, Matthieu Quignard, Sophie Rosset, Christophe Servan, et al. 2006. Results of the french evalda-media evaluation campaign for literal understanding. In *The fifth international conference on Language Resources and Evaluation (LREC 2006)*.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. Multiwoz–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278*.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*.

Samira Ben Dbabis, Fatma Mallek, Hatem Ghorbel, and Lamia Belguith. 2012. Dialogue acts annotation scheme within arabic discussions. In *Proceedings of SemDial 2012 (SeineDial): The 16th Workshop on the Semantics and Pragmatics of Dialogue*, page 88.

Ahana Deb, Sayan Nag, Ayan Mahapatra, Soumitri Chattopadhyay, Aritra Marik, Pijush Kanti Gayen, Shankha Sanyal, Archi Banerjee, and Samir Karmakar. 2023. Beats: Bengali speech acts recognition using multimodal attention fusion. *arXiv preprint arXiv:2306.02680*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Marco Dinarelli, Nikita Kapoor, Bassam Jabaian, and Laurent Besacier. 2020. A Data-Efficient End-to-End Spoken Language Understanding Architecture. In *International Confer-*

*ence on Acoustics, Speech, and Signal Processing (ICASSP)*, Barcellone, Spain.

AbdelRahim A. Elmadany, Sherif M. Abdou, and Mervat Gheith. 2015. Jana: An arabic human-human dialogues corpus. In *2015 IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS)*, pages 347–352.

Sahar Ghannay, Antoine Caubrière, Yannick Estève, Antoine Laurent, and Emmanuel Morin. 2018. End-to-end named entity extraction from speech. *SLT 2018*.

Sahar Ghannay, Antoine Caubrière, Salima Mdhaffar, Gaëlle Laperrière, Bassam Jabaian, and Yannick Estève. 2021. Where are we in semantic concept extraction for spoken language understanding? In *SPECOM 2021 23rd International Conference on Speech and Computer*.

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

Marwa Graja, Maher Jaoua, and Lamia Hadrich Belguith. 2013. Discriminative framework for spoken tunisian dialect understanding. In *Statistical Language and Speech Processing: First International Conference, SLSP 2013, Tarragona, Spain, July 29-31, 2013. Proceedings 1*, pages 102–110. Springer.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Imane Guellil, Houda Saâdane, Faical Azouaou, Billel Gueni, and Damien Nouvel. 2021. Arabic natural language processing: An overview. *Journal of King Saud University-Computer and Information Sciences*, 33(5):497–507.

Nizar Habash, Mona T Diab, and Owen Rambow. 2012. Conventional orthography for dialectal arabic. In *LREC*, pages 711–718.

Parisa Haghani, Arun Narayanan, Michiel Bacchiani, Galen Chuang, Neeraj Gaur, Pedro Moreno, Rohit Prabhavalkar, Zhongdi Qu, and Austin Waters. 2018. From audio to semantics: Approaches to end-to-end spoken language understanding. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 720–726. IEEE.

Charles T Hemphill, John J Godfrey, and George R Doddington. 1990. The atis spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*.

Gaëlle Laperrière, Valentin Pelloin, Antoine Caubrière, Salima Mdhaffar, Nathalie Camelin, Sahar Ghannay, Bassam Jabaian, and Yannick Estève. 2022. The spoken language understanding media benchmark dataset in the era of deep learning: data updates, training and evaluation tools. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1595–1602.

Chahira Lhioui, Anis Zouaghi, and Mounir Zrigui. 2013. A combined method based on stochastic and linguistic paradigm for the understanding of arabic spontaneous utterances. In *Computational Linguistics and Intelligent Text Processing: 14th International Conference, CICLing 2013, Samos, Greece, March 24-30, 2013, Proceedings, Part II 14*, pages 549–558. Springer.

Chia-Hsuan Li, Szu-Lin Wu, Chi-Liang Liu, and Hung-yi Lee. 2018. Spoken squad: A study of mitigating the impact of speech recognition errors on listening comprehension. *arXiv preprint arXiv:1804.00320*.

Mohamed Lichouri, Khaled Lounnas, and Adil Bakri. 2023. Toward building another arabic voice command dataset for multiple speech processing tasks. In *2023 International Conference on Advances in Electronics, Control and Communication Systems (ICAECCS)*, pages 1–5. IEEE.

Mohamed Lichouri, Khaled Lounnas, Rachid Djeradi, and Amar Djeradi. 2022. Performance of end-to-end vs pipeline spoken language understanding models on multilingual synthetic voice. In *2022 International Conference on Advanced Aspects of Software Engineering (ICAASE)*, pages 1–6. IEEE.

Bing Liu and Ian Lane. 2016. Attention-Based Recurrent Neural Network Models for Joint Intent Detection and Slot Filling. In *Proc. Interspeech 2016*, pages 685–689.

Loren Lugosch, Piyush Papreja, Mirco Ravanelli, Abdelwahab Heba, and Titouan Parcollet. 2021. Timers and such: A practical benchmark for spoken language understanding with numbers. *arXiv preprint arXiv:2104.01604*.

Loren Lugosch, Mirco Ravanelli, Patrick Ignoto, Vikrant Singh Tomar, and Yoshua Bengio. 2019. Speech Model Pre-Training for End-to-End Spoken Language Understanding. In *Proc. Interspeech 2019*, pages 814–818.

Abir Masmoudi, Mariem Ellouze Khmekhem, Yannick Esteve, Lamia Hadrich Belguith, and Nizar Habash. 2014. A corpus and phonetic dictionary for tunisian arabic speech recognition. In *LREC*, pages 306–310.

Salima Mdhaffar, Jarod Duret, Titouan Parcollet, and Yannick Estève. 2022a. End-to-end model for named entity recognition from speech without paired training data. In *Interspeech 2022*.

Salima Mdhaffar, Valentin Pelloin, Antoine Caubrière, Gaëlle Laperrière, Sahar Ghannay, Bassam Jabaian, Nathalie Camelin, and Yannick Estève. 2022b. Impact analysis of the use of speech and language models pretrained by self-supersivion for spoken language understanding. In *LREC 2022*.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. *Interspeech 2019*.

Lance A Ramshaw and Mitchell P Marcus. 1999. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, pages 157–176. Springer.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. 2021. Speechbrain: A general-purpose speech toolkit.

John R. Searle. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.

Paden Tomasello, Akshat Shrivastava, Daniel Lazar, Po-Chun Hsu, Duc Le, Adithya Sagar, Ali Elkahky, Jade Copet, Wei-Ning Hsu, Yossi Adi, et al. 2023. Stop: A dataset for spoken task oriented semantic parsing. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 991–998. IEEE.

Natalia Tomashenko, Christian Raymond, Antoine Caubrière, Renato De Mori, and Yannick Estève. 2020. Dialogue History Integration into End-to-End Signal-to-Concept Spoken Language Understanding Systems. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 5, Barcelona, Spain.

Roger Trias-Sanz and José B Marino. 2002. Basurde [lite], a machine-driven dialogue system for accessing railway timetable information. In *INTERSPEECH*.

Gokhan Tur and Renato De Mori. 2011. Spoken language understanding: Systems for extracting semantic information from speech.

Ye-Yi Wang, Li Deng, and Alex Acero. 2005. Spoken language understanding. *IEEE Signal Processing Magazine*, 22(5):16–31.

Hemant Yadav, Sreyan Ghosh, Yi Yu, and Rajiv Ratn Shah. 2020. End-to-end named entity recognition from english speech. *Interspeech*.

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

Xiaodong Zhang and Houfeng Wang. 2016. A joint model of intent determination and slot filling for spoken language understanding. In *IJCAI*, volume 16, pages 2993–2999.