Munazarat 1.0: A Corpus of Arabic Competitive Debates

Mohammad Majed Khader[1], **Abdul Gabbar Al-Sharafi**[2],
Mohamad Hamza Al-Sioufy[3], **Wajdi Zaghouani**[4], **Ali Al-Zawqari**[5]
[1]QatarDebate Center, [2]Sultan Qaboos University,
[3]Georgetwon University in Qatar, [4]Hamad Bin Khalifa University,
[5]Department of Fundamental Electricity and Instrumentation, Vrije Universiteit Brussel
[1]mkhader@qatardebate.org, [2]alsharaf@squ.edu.om, [3]ma2052@georgetown.edu
[4]wzaghouani@hbku.edu.qa, [5]aalzawqa@vub.be

Abstract

This paper introduces the Corpus of Arabic Competitive Debates, Munazarat. Despite the significance of competitive debating in fostering critical thinking and promoting dialogue, researchers in the fields of Arabic Natural Language Processing (NLP), linguistics, argumentation studies, and education have limited access to datasets on competitive debating. At this stage of the study, we introduce Munazarat 1.0, which combines transcribed recordings of approximately 50 hours from 73 debates at QatarDebate-recognized tournaments, all available on YouTube. Munazarat is a novel specialized Arabic speech corpus, predominantly in Modern Standard Arabic (MSA), covering diverse debating topics and accompanied by metadata for each debate. The transcription of debates was performed using Fenek, a speech-to-text Kanari AI tool, and reviewed by three native Arabic speakers to enhance quality. The Munazarat 1.0 dataset can serve as a valuable resource for training Arabic NLP tools, developing argumentation mining machines, and analyzing Arabic argumentation and rhetoric styles.

Keywords: Arabic Speech Corpus, Modern Standard Arabic, Debates

1. Introduction

Arabic is the sixth most spoken language in the world. As a Semitic language, Arabic distinguishes itself from the Indo-European linguistic family in several dimensions: phonetically, morphologically, syntactically, and semantically. Thus, the development and research of Arabic NLP applications face various challenges based on the language's linguistic structure (Shaalan et al., 2019). Furthermore, an additional challenge is that Arabic exists today in three forms: (i) Classical Arabic, (ii) Modern Standard Arabic (MSA), and (iii) Dialectical Arabic, which varies significantly based on geographical regions. The Arabic language is suffering from a scarcity of available open datasets compared to English and other languages like Chinese, German, and French. In Papers With Code (pap), a repository showed results of open text datasets in March 2024: 1446 for English, 205 for Chinese, 126 for German, and only 54 for Arabic. While Hugging Face repository (hug) showed results of only 446 Arabic datasets out of 126,088 open text datasets in comparison to 8,826 for English, 1005 for Chinese, and 667 for German.

Competitive debating, an intellectually rigorous oral argumentative discourse activity governed by specific rules and regulations, typically takes place in the context of large tournaments. Thousands of university and school students from different geographical regions around the world participate in local and international Arabic debating tournaments. For Arabic debating, QatarDebate Center (www.qatardebate.org) is considered the leading debate institution, organizing major international Arabic debating tournaments and publishing the recordings of debates on YouTube. QatarDebate's 3 vs 3 debate format, as shown in Figure 1, a modified format of the World Schools Debating Championship (www.wsdcdebating.org), is dominant in Arabic competitive debating activities. A motion is presented for every debate in this format, and two opposing teams compete against one another. Every team consists of three speakers, and each one is allowed to talk for a total of 6-7 minutes, beginning with the first proposition speaker, followed by the first opposition speaker, and so on till the last opposition speaker. Then, each team delivers a three-minute technical speech called the "Reply Speech" that does not include any new argument. Due to the competitive nature of these debates, an adjudication panel of an odd number of judges votes for the most persuasive team to win and assign individual speakers' scores. The effectiveness of the offered argumentation and refutation is the primary criterion for judging debate presentations. This type of debate is very structured and follows specific rules and regulations that govern the flow of the debate and its evaluation.

The significance of creating an Arabic debate corpus comes from the fact that debates are rich in argumentative and sentimental speeches that can help study Arabic argumentation and rhetoric styles. It can also be used to study various linguistic features of the spoken MSA among native and non-native debaters. In addition, it provides raw data for developing Arabic NLP tools for argumen-

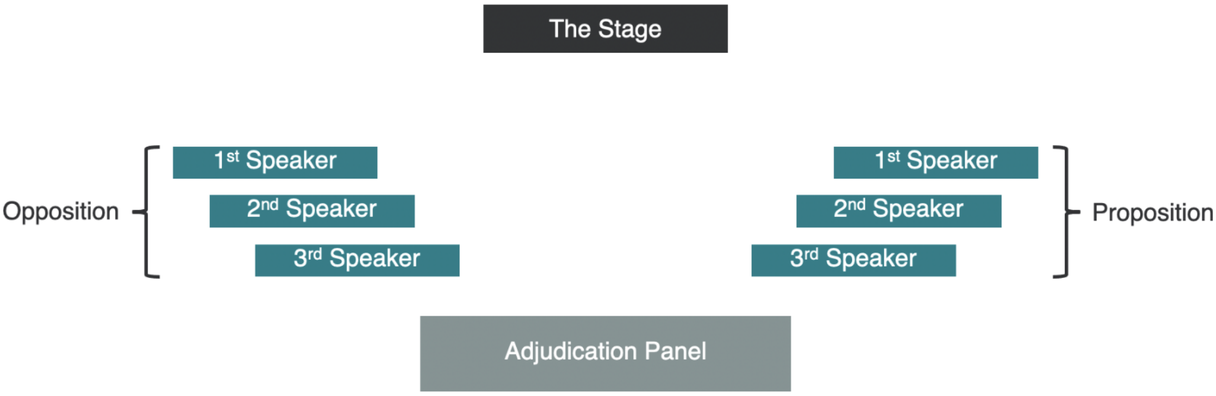


Figure 1: Illustration of 3 vs 3 Debate Format

tation mining, speech recognition, etc. Unlike other datasets, Munazarat 1.0 stands today as the specialized corpus of Arabic competitive debate and the largest corpus of argumentative Arabic content.

2. Related Work

Dataset availability is an essential key to developing NLP applications. However, the cost of acquiring corpora represents a major challenge, especially in Arabic NLP with all of its variations (Ahmed et al., 2022c; Zaghouani, 2014). After a survey of available Arabic resources today (Al-sulaiti and Atwell, 2006; El-Khair, 2016; Al-Twairesh et al., 2018; Graja et al., 2010; Almeman et al., 2013; Alrabiah et al., 2013; Ahmed et al., 2022a; Mubarak et al., 2021; Khader, 2020; Al-Fetyani et al., 2023; Bouamor et al., 2018), and despite the recent efforts in the field of Arabic NLP (Darwish et al., 2021), the available specialized Arabic corpora remain in shortage. Datasets of relevance to our study manifest as either Arabic speech corpora or compilations encompassing discourse of a debating or argumentative nature.

The development of the Arabic PropBank has been instrumental in the semantic analysis of Arabic texts. These efforts have laid the groundwork for parsing argument structures in sentences (Palmer et al., 2008; Diab et al., 2008; Zaghouani et al., 2010) while Error annotation is essential for the accuracy and reliability of language resources. Studies focusing on large-scale Arabic error annotation and non-native text correction have significantly contributed to the field (Zaghouani et al., 2014) and (Zaghouani et al., 2015). Furthermore, Dialectal variation in Arabic poses unique challenges for argumentation analysis. The MADAR project and the DIACT corpus have addressed this by focusing on dialect-specific expressions and the use of rhetorical devices such as irony (Bouamor et al., 2018; Abbes et al., 2020).

By situating our work alongside these significant contributions, we aim to address the gap in resources specific to argumentation within the Arabic language, building on the robust foundations laid by these earlier works. Each cited resource provides a unique perspective on the intricacies of argumentative discourse, from structural annotations to the subtleties of linguistic diversity.

2.1. Speech Corpora

Lately, two Arabic speech corpora were introduced: the Massive Arabic Speech Corpus (MASC) (Al-Fetyani et al., 2023), which contains 1,000 hours from over 700 YouTube channels, and QCRI Aljazeera Speech Resource (QASR) (Mubarak et al., 2021) which is the largest Arabic speech corpus to date and consists of 2,000 hours from Aljazeera TV channel shows. Recently, a digital corpus of the Australian Parliamentary Debates was published (Katz and Alexander, 2023) following the lead of the Canadian Parliamentary Debates (Beelen et al., 2017). Those two studies show the recent interest in collecting and publishing specialized debate corpora, namely political debates. The availability of English debate corpora highlights the gap for an equivalent Arabic debate collection we seek to address in providing a source for Arabic competitive debates.

2.2. Debate & Argumentation Corpora

Many corpora were found to be interested in studying debates and argumentation models in English (Hautli-Janisz et al., 2022; Serban et al., 2015; Fisas et al., 2016; Peldszus and Stede, 2015). Several studies have compiled corpora to advance research in argument mining and related tasks. Walker et al. (2012) introduced a corpus of English language debates annotated with argumentative discourse units to facilitate computational argumentation research. Zhang et al. (2021) presented a corpus of Wikipedia talk page conversations annotated for conversational failure, enabling the study of breakdowns in cooperative discussion. Lawrence and Reed (2020) surveyed datasets for argument mining, reviewing annotation approaches across key tasks.

Other efforts have focused on particular argumentation genres and languages. Al Khatib et al. (2018) annotated German Wikipedia articles with argument strategies, like evidence types, to analyze deliberative argumentation. Bar-Haim et al. (2006) overviewed textual entailment challenges involving argumentation data. Orăsan and Evans (2007) developed a corpus of noun phrase animacy annotations to assist anaphora resolution with potential dialogue applications. Some datasets have annotated the persuasiveness of arguments. Habernal and Gurevych (2016) presented a corpus of web argument pairs annotated for comparative convincingness to predict persuasiveness. Hidey and McKeown (2018) annotated student essays for argument persuasiveness and sequencing.

Other studies have advanced annotation methodologies. For instance, Musi et al. (2018) performed an annotation study of argument schemes like expert opinion to provide guidelines. On the other hand, Aharoni et al. (2014) annotated claims and evidence in controversial topics for automatic detection. There are also argument-mining efforts in other languages like Italian (Durmus et al., 2021) and argument relation annotations from multilingual social media like X (Twitter previously) (Bosc et al., 2016).

For the purpose of this study, the most notable previous work is QT30 corpus (Hautli-Janisz et al., 2022), which contains public debates from the BBC's show 'Question Time'. However, it is limited to only 30 episodes and focuses solely on political debates. Yet, to the best of our knowledge, there is no work focusing on building a corpus in Arabic for argumentation or debating, except for two recent projects. The first one is a project of (Khader, 2020), which introduced a small corpus containing only 12 debate recordings. The other one is the Qatari Corpus of Argumentative Writing (QCAW) (Ahmed et al., 2024; Zaghouani et al., 2024; Ahmed et al., 2022a), which targets bilingual (Arabic/English) argumentative texts by students, providing a novel resource for cross-linguistic argumentation studies with 195 texts in Arabic and 195 texts in English. This corpus facilitates a deeper understanding of argumentative structures within an educational context, contributing to the field of discourse analysis. Yet, QCAW does not incorporate any spoken argumentative content.

The availability of rich resources for argumentative and persuasive Arabic speech is nonexistent. Yet, argument mining from spoken content could enable studies on rhetoric, reasoning, and dialectics across the language's breadth. Competitive debating generates valuable linguistic data - structured speeches rich in argumentation, sentiment, and diverse vocabulary spanning different topics. Debates capture authentic goal-oriented argumentation between experts, unlike other dialogues (Serban et al., 2015). The lack of argumentative and conversational Arabic speech data poses challenges for speech recognition, dialect studies, and MSA research (Al-Fetyani et al., 2023). Applications like argument mining also require substantial training corpora (Lawrence and Reed, 2019).

Munazarat 1.0 data can facilitate Arabic research on linguistics, reasoning, debating, and NLP applications through this resource. Our work addresses the key limitations of scarce available Arabic corpora compared to other languages, very minimal argumentative or conversational Arabic data, lack of large-scale Arabic speech resources for training models, and the absence of a dedicated corpus for the rich Arabic debating domain.

3. Methods

3.1. Debate Collection

Munazarat 1.0 consists of approximately 50 hours of transcribed Arabic competitive debates that QatarDebate Center hosted in several tournaments. The corpus is created using 73 debates prerecorded and already published online by the host, as we collected them from YouTube without disclosing any extra private information about the debaters. The collected debates comprise a combination of university and school debates held between 2013 and 2023. The corpus will be expanded, with an expected goal of reaching 120 debates by the end of 2024.

3.2. Transcription and Human Review

All debates were transcribed using Fenek, a multilingual Arabic/English speech-to-text tool from Kanari AI (www.kanari.ai) (Khurana and Ali, 2016). After that, all debate transcripts were cleaned, briefly annotated, and reviewed in three stages, as described below, to ensure the transcription quality. We also published more details on the human reviewing guidelines that we used in this project with the dataset for public usage. It is important to mention that 10 transcripts were taken from publicly available previous work of (Khader, 2020) and those transcripts went only through stages two and three of the human review.

- **Stage One:** During the first review, the reviewer listens to the debate from the YouTube link while reading the transcription in order to eliminate any mistakes made by the Artificial Intelligence (AI) tool in the transcription. We identified four types of transcription mistakes for the reviewer to correct: added words (highlighted in red), missed words (highlighted in yellow), spelling mistakes (highlighted in green), and language detection mistakes (highlighted in blue). The reviewer also deletes any side talks that happen in the recording that are not part of the six essential debate speeches, such as deleting the chair's organizational remark. We also decided to remove the "Reply Speeches" from the script since they are not essential to the debate and are not standard in all debating tournaments. The reviewer cleans any repetitions during the speeches if they were caused by unintentional stammering. By the end of this stage, the reviewer produces a clean file ready for the next reviewer with three marking steps: (i) making a brief annotation by marking the beginning of each debater's speech by (#) symbol and indicating the speaker's order and position; (ii) mentioning the gender of the speaker at the beginning of each speech as illustrated in Figure 2; (iii) marking any Point of Information (POI) from the opponent team as shown in Figure 3.

- **Stage Two:** In the second review, another reviewer reviews the script. However, this time, the reviewer only reads the text and does not listen to the debate. This stage was meant to account for any typos, grammatical and spelling mistakes, etc., that the first reviewer did not catch. In rare cases, the second reviewer revisits the debate video to cross-check the transcript.

- **Stage Three:** This stage is for quality control, where the third reviewer eliminates any mistakes that were left by the previous two reviewers and provides feedback for them during the periodic reviewers' meetings. In addition, the third reviewer tries to organize the transcript in the form of paragraphs to produce a better readable file.

4. Data Records & Analysis

Munazarat 1.0 is a unique resource for researchers interested in various aspects of Arabic competitive debating, Arabic linguistics studies, argumentation studies, education, and Arabic NLP. Munazarat 1.0 is available for public download as a ZIP file containing 73 debate files in TXT format to facilitate its effective use. Researchers can access and download this dataset via an open access github[1]. Each file is named descriptively, incorporating information about the debate, including the serial number, tournament, year, gender, and whether the speakers are native or non-native Arabic speakers. For example, 028-IUDC-2017-MFMFMF-AA represents a debate with serial number 028, from the International Universities Debating Championship (IUDC), featuring three male speakers in the proposition team and three female speakers in the opposition team, all of whom are Arabic native speakers. Each TXT file includes basic annotations that indicate the order and the gender of the speaker as well as any POI from the opponent speakers.

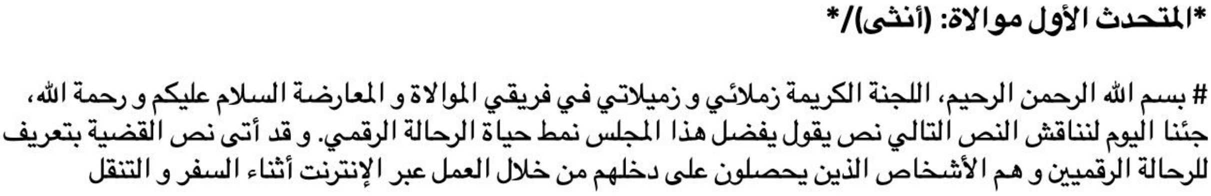
The corpus represents a diverse collection in several aspects, as shown in Table 1. The demographic representation in this corpus is rich. A list of 27 countries in the corpus is shown in Table 2, and the higher occurrences are relevant to the country's history of participation in the Arabic debate activity. The corpus is inclusive of 51 debates between native Arabic speakers, 22 debates between native and non-native speakers, 51 university-level debates from international tournaments, 11 school debates from international tournaments, and 11 school debates from Qatar.

Munazarat 1.0 also displays a well-balanced male-to-female ratio of (M:223, F:215) since some studies pointed out the differences in speech patterns among genders in English debates (Shaw, 2000; Hargrave and Langengen, 2021), which needs to be examined against an Arabic dataset. The debate motions are diverse and wide-ranging, from politics and philosophy to sports. Table 3 provides the overall topic distribution of the debates. For each debate, we have the following: a video recording with a YouTube link, a transcribed text (TXT) file of the debate's script, and some metadata, explained later in the Data Records section.

Table 1: Diversity Representation

Category	Count
Tournament Level - Debate Count	
University Level	51
School Level	22
Geographic Representation - Debate Count	
Local (Qatar)	11
International	62
Language Proficiency - Debate Count	
All Native Arab Speakers	51
Natives and Non-Natives Speakers	22
Gender Representation - Speakers Count	
Male Debaters	223
Female Debaters	215

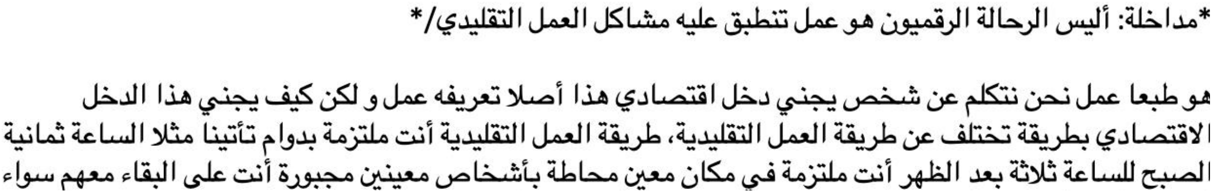
[1] https://github.com/moh72y/Munazarat1.0/



المتحدث الأول موالاة: (أنثى)/

بسم الله الرحمن الرحيم، اللجنة الكريمة زملائي و زميلاتي في فريقي الموالاة و المعارضة السلام عليكم و رحمة الله،
جئنا اليوم لنناقش النص التالي نص يقول يفضل هذا المجلس نمط حياة الرحالة الرقمي. و قد أتى نص القضية بتعريف
للرحالة الرقميين و هم الأشخاص الذين يحصلون على دخلهم من خلال العمل عبر الإنترنت أثناء السفر و التنقل

Figure 2: Beginning of Speech Annotation: Debater’s Role, Gender, and # Symbol.
English Translation: *First speaker Poposition (Female) In the name of God, the Most Gracious, the Most Merciful, the honorable committee, my colleagues in the proposition and opposition teams, may God’s peace and mercy be upon you. We have come today to discuss the following motion, the motion says that, This house prefers the digital nomad’s lifestyle. The motion came with a definition of digital nomads, who are the people who obtain their income through working online while traveling*



مداخلة: أليس الرحالة الرقميون هو عمل تنطبق عليه مشاكل العمل التقليدي/

هو طبعا عمل نحن نتكلم عن شخص يجني دخل اقتصادي هذا أصلا تعريفه عمل و لكن كيف يجني هذا الدخل
الاقتصادي بطريقة تختلف عن طريقة العمل التقليدية، طريقة العمل التقليدية أنت ملتزمة بدوام تأتينا مثلا الساعة ثمانية
الصبح للساعة ثلاثة بعد الظهر أنت ملتزمة في مكان معين محاطة بأشخاص معينين مجبورة أنت على البقاء معهم سواء

Figure 3: Point of Information (POI) Annotation.
English Translation: *Point of Information: Isn’t digital nomads a job to which the problems of traditional work apply/ It is, of course, work. We are talking about a person who earns economic income. This is basically the definition of work, but how does he earn this economic income in a way that differs from the traditional method of work. The traditional method of work, You are committed to a shift. You come, for example, at eight in the morning until three in the afternoon. You are committed to a specific place surrounded by specific people that you are forced to stay with them.*

Table 2: Country Distribution

Country	No. of Teams	University Level	School Level
Qatar	35	16	19
Jordan	13	12	1
Sudan	12	12	0
Oman	12	11	1
Tunisia	10	10	0
Malaysia	9	5	4
Kuwait	8	7	1
Palestine	6	3	3
Libya	6	6	0
Lebanon	6	3	3
Türkiye	5	2	3
USA	5	5	0
Indonesia	3	2	1
Syria	3	0	3
Algeria	1	1	0
Iraq	1	1	0
Somalia	1	1	0
Bahrain	1	1	0
Norway	1	1	0
Canada	1	1	0
Poland	1	1	0
Morocco	1	0	1
Pakistan	1	0	1
Singapore	1	0	1
Yemen	1	0	1
Côte d’Ivoire	1	1	0
Australia	1	0	1
Total	146	102	44

Table 3: Topic Distribution

Topic	No. of Debates
Politics	16
Ethics/Philosophy	16
Human Rights	10
Media	6
Education	5
Technology	5
Culture	3
Environment	3
Law	3
Sports	3
Economy	2
Lifestyle	1
Total	73

5. Technical Validation

5.1. AI Transcription Accuracy Report

This human validation process was done fully on 63 newly transcribed debates out of 73, and partially on the other 10 transcripts that were taken from previous work by (Khader, 2020) as the transcripts were available for public use online. During the first human review stage mentioned above, while listening to and editing the debates, the reviewer identified transcription mistakes in four categories. The red category is used to highlight any additional words that the tool added but were not originally spoken by the speaker during their speech. The yellow category is used to highlight any words that were added by the reviewer and were missed by the tool. The green category is used with the words caught wrongly by the tool and thus modified by the reviewer. Finally, language detection mistakes in the blue category to highlight words that were in a different language, as the debates were conducted originally in Arabic, but some terminologies in English might appear and were written in a wrong way by the tool. Figure 4 and Figure 5 show a sample of the color coding process. After that, the reviewer stores the data from each debate regarding the number of mistakes in each category and the total number of mistakes in the whole debate. Table 4 demonstrates the Mean and Median of mistake count per debate for each category reported by the human reviewer.

Following the transcription of the 63 debates using the speech-to-text tool from Kanari AI, we report an average accuracy rate of 96% per debate. Approximately 40% of the tool's mistakes fall under the 'Missed Word' category, which we attribute to microphone quality and the fast speaking pace of some debaters. Conversely, the tool effectively detected language switches when debaters used English for certain terminologies.

Table 4: Mean (M) and Median (Mdn) Transcription Accuracy Report

Category	M per Debate	Mdn per Debate
Word Count	4546	4458
Added Words	49	37
Missed Words	76	28
Spelling Mistake	58	45
Language Detection Mistake	1	0
Total Mistakes	185	140
Accuracy Rate	96%	97%

5.2. Keyword Analysis

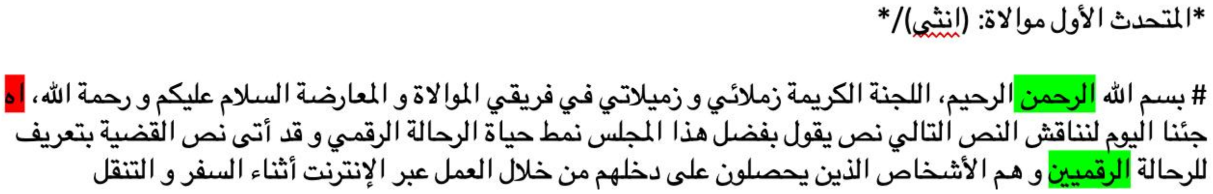
Keyword analysis is a vital aspect of corpus studies, helping unveil a corpus's underlying themes and domain. In exploring Munazarat 1.0, a diverse debate corpus, we employed AntConc software (Anthony, 2023) to conduct a comprehensive keyword analysis. The keyness function in AntConc generates the keyword list of the studied corpus compared to a reference, usually a much larger and generic one. These keywords are not merely the most frequent words in the corpus; they represent statistically significant words that shed light on the corpus's domain. This function helps filter out stopwords, insignificant words, and letters, allowing us to recognize the corpus's domain and key themes. For this analysis, we utilized QASR (Mubarak et al., 2021), one of the largest available Arabic speech corpora, as our reference. In Table 5, we present the keywords from various categories within Munazarat 1.0, both in comparison to the corpus itself and against QASR.

A preliminary review of the keyword list from a complete or partial corpus analysis, in comparison to QASR, reveals the distinctive nature of Munazarat 1.0 as a debate corpus, with terms like "proposition", "team", "speaker". and "this house" stand out. However, it is important to note that the initial analysis of keywords within specific portions, category-based, of the corpus against Munazarat 1.0 primarily reflects the debated topics within that portion rather than providing insights regarding the characteristics of the studied category. Still, a dedicated study among various categories in Munazrart 1.0 might reveal some linguistic styles that can be associated with non-native debaters, school debaters, Qatari debaters, etc.

Table 5: Top Five Keywords per Category

Category	Against Munazarat 1.0	Against QASR
Native Speakers - University Level	العمال	الموالاة
	Labours	Proposition
	الحرب	فريق
	War	Team
	سادتي	المتحدث
	Gentlemen	Speaker
	الأندية	سادتي
	Clubs	Gentlemen
	الحقوق	سوف
	Rights	Will
General - Schools Level	تحليل	الموالاة
	Analysis	Proposition
	الطلاب	المتحدث
	Students	Speaker
	العربية	أيها
	Arabic	Hey
	الجمهور	فريق
	Audience	Team
	الأيام	المتحدثة
	Days	Speaker - female
Native Students from Qatari Schools	تحليل	الموالاة
	Analysis	Proposition
	الأيام	المتحدث
	Days	Speaker
	التواصل	تحليل
	Communication	Analysis
	يوم	المتحدثة
	Day	Speaker - female
	مواقع	بركاته
	Sites	His Blessings

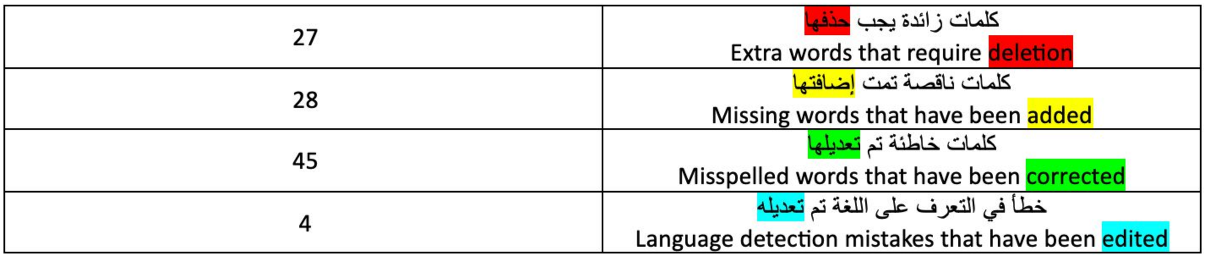
Table 6 demonstrates a sample from the keyword analysis per theme. Debates were selected from three themes: Politics, Ethics/Philosophy, and Technology. The results show that generic debate terms appeared, as expected, against QASR (Mubarak et al., 2021) for both politics and ethics. However, theme-specific words related to AI most notably appeared for the technology theme, telling us that QASR is most probably poor for AI terms despite its length and diversity. On the other hand, the theme-based keyword analysis reflected the themes when run against Munazarat 1.0. The words "Intelligence" and "Artificial" were the most highlighted keywords which reflects the fact that four debate transcripts out of five in the technology



المتحدث الأول موالاة: (انثي)/

بسم الله الرحمن الرحيم، اللجنة الكريمة زملائي و زميلاتي في فريقي الموالاة و المعارضة السلام عليكم و رحمة الله، اه
جئنا اليوم لنناقش النص التالي نص يقول بفضل هذا المجلس نمط حياة الرحالة الرقمي و قد أتى نص القضية بتعريف
للرحالة الرقمين و هم الأشخاص الذين يحصلون على دخلهم من خلال العمل عبر الإنترنت أثناء السفر و التنقل

Figure 4: Sample of Color Coding Transcription Mistakes
English Translation: *First speaker poposition (Female) In the name of God, the Most Gracious, the Most Merciful, the honorable committee, my colleagues in the proposition and opposition teams, may God's peace and mercy be upon you. We have come today to discuss the following motion, the motion says that, This house prefers the digital nomad's lifestyle. The motion of the issue came with a definition of digital nomads, who are the people who obtain their income through working online while traveling*



27	كلمات زائدة يجب حذفها Extra words that require deletion
28	كلمات ناقصة تمت إضافتها Missing words that have been added
45	كلمات خاطئة تم تعديلها Misspelled words that have been corrected
4	خطأ في التعرف على اللغة تم تعديله Language detection mistakes that have been edited

Figure 5: Sample of the Mistakes Table in the First Human Review Process

category are debates about AI.

6. Usage Notes

Along with the debate transcript files, we offer a detailed Excel sheet that provides metadata for each debate. This metadata includes information such as the tournament, university or school level, debate motion, proposition and opposition teams, the number of male and female debaters, word count, YouTube link, and the debate topic genre (e.g., Politics, Economy, Human Rights, Law, etc.). Researchers can use this metadata for various analytical purposes and to filter debates based on specific criteria.

The dataset provided in this study is the largest available Arabic argumentative transcribed text to date, which makes it suitable for several applications including but not limited to the three following suggestions: (i) using UBIAI (www.ubiai.tools) text annotation online software to annotate the speeches' argument scheme since it is compatible with the Arabic text; (ii) applying sentimental analysis on the corpus using tools such as Repustate (www.repustate.com); and (iii) running more linguistic analysis through AntConc (www.laurenceanthony.net/software/antconc/).

The dataset is currently provided in a separate TXT file for each debate. However, it can be easily converted to other formats as per the researchers' requirements. It can also be easily segmented into separate files per speech for extra gender or demographical-based analysis. To facilitate the segmentation process, the beginning of each speech is marked by a (#) symbol.

While Munazarat 1.0 serves as a substantial raw corpus, it currently lacks standard splits into training and test sets to enable benchmarking of AI models. Creating such splits by partitioning the data while maintaining balance across dialects, speaker demographics, topics, and other variables is an important area we aim to pursue in future work. We plan to take measures to avoid speaker overlap between the splits. The speaker metadata captured in our annotations will assist in creating speaker-independent partitions. Providing standardized training and test splits will allow Munazarat 1.0 to serve as a rigorous benchmark dataset for developing and evaluating Arabic argument mining and related NLP models. We will make the splits available along with the corpus.

7. Limitations

Munazarat 1.0 has some limitations to highlight. In the current version, only competitive debating content is included. Adding other genres, like talk shows, could improve the diversity of the dataset. Moreover, the metadata currently captures basic attributes. More fine-grained speakers and sociolinguistic metadata could enable deeper analyses. The semi-automated transcription allows some errors; therefore, periodic human checks on newer data may help enhance quality. Finally, the release

Table 6: Top Five Keywords per Selected Themes

Theme	Against Munazarat 1.0	Against QASR
Politics	الدول Countries	الموالاة Proposition
	المساعدات Aids	الدول Countries
	الصين China	سادتي Gentlemen
	التدخل Intervention	فريق Team
	روسيا Russia	المتحدث Speaker
Ethics & Philosophy	الطبيب Physician	الموالاة Proposition
	الرقابة Surveillance	أيها Hey
	الأيام Days	فريق Team
	الآثار Monuments	المتحدث Speaker
	الإسلام Islam	السادة Gentlemen
Technology	الذكاء Intelligence	الذكاء Intelligence
	الصناعي Artificial	الصناعي Artificial
	الاصطناعي Artificial	الاصطناعي Artificial
	التقدم Progress	المولاة Proposition
	الإنسان Human	سادتي Gentlemen

rights limit sharing some video recordings publicly, and getting broader rights could increase accessibility. Addressing these limitations through corpus expansion, increased metadata, transcription quality checks, and enhanced accessibility can make Munazarat 1.0 an even more impactful community resource. We aim to pursue these improvements in ongoing and future work.

8. Conclusion

We have introduced Munazarat 1.0, the first large-scale corpus of transcribed Arabic competitive debates. Spanning 50 hours of content across 73 university and school-level debates, Munazarat 1.0 represents a valuable linguistic resource for Arabic NLP and related fields. We described the rigorous process of collecting high-quality video recordings and machine transcribing debates using speech recognition, followed by extensive human reviews.

With the provided metadata, including speaker demographics and debate topics, Munazarat 1.0 enables multifaceted analyses of argumentation, rhetoric, dialectal variations, and other phenomena in Arabic debates. Our validation demonstrates the accuracy of the AI-generated transcripts. Keyword analyses reveal the corpus's core themes like argumentation and specific debate motions. Munazarat 1.0 provides Arabic researchers with a substantial dataset to train computational models and drive advancements for impactful applications in education, linguistics, and reasoning analysis. Currently, two works in the literature are introduced to take advantage of Munazarart 1.0, namely in (Al-Sharafi et al., accepted 2024; Al-Zawqari et al., accepted 2024). The first one is developing an annotation model for argumentation in competitive debates, and the second is focusing on the classification of persuasion modes in Arabic debates according to Aristotle's rhetoric.

9. Ethical Statement

In compiling and releasing Munazarat 1.0, rigorous procedures were followed to protect user privacy and obtain consent. The included debates were exclusively sourced from publicly accessible YouTube videos released by participating institutions with debaters' consent. The corpus does not reveal any extra personal data that was not already published publicly. The textual transcripts contain no direct user IDs or handles. Furthermore, the educational institutions that originally published the footage were contacted regarding the research use of this content. Only recordings that we received consent to share in Munazarat 1.0 were included. Those rigorous procedures ensure that, while maximizing the data's research utility, we maintain participant privacy and ethics in compiling and releasing this corpus.

10. Acknowledgements

This work was made possible by two QD Fellowship awards [QDRF-2022-01-003] and [QDRF-2022-01-005] from QatarDebate Center. We would like also to thank the group of native Arab students who contributed to this project by carrying on the task of reviewing text and human validation: Besher Al-Sioufy, Hilmi AbuAlyyan, Moaz Jemmieh, Abdullah Al-Shaar, Jenen Al-Hanai, and Abdullah Al-Kubaisi.

11. Bibliographical References

Datasets | hugging face the ai community building the future. https://huggingface.co/datasets?sort=trending. Accessed: 2024-03-30.

The latest in machine learning | papers with code. https://paperswithcode.com/. Accessed: 2024-03-30.

Ines Abbes, Wajdi Zaghouani, Omaima El-Hardlo, and Faten Ashour. 2020. Daict: A dialectal arabic irony corpus extracted from twitter. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6265–6271.

Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. 2014. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *Proceedings of the first workshop on argumentation mining*, pages 64–68.

Abdelhamid Ahmed, Debra Myhill, Esmaeel Abdollahzadeh, Lee McCallum, Wajdi Zaghouani, Lameya Rezk, Anissa Jrad, and Xiao Zhang. 2022a. Qatari corpus of argumentative writing.

Abdelhamid M Ahmed, Xiao Zhang, Lameya M Rezk, and Wajdi Zaghouani. 2024. Building an annotated l1 arabic/l2 english bilingual writer corpus: The qatari corpus of argumentative writing (qcaw). *Corpus-based Studies across Humanities*.

Arfan Ahmed, Nashva Ali, Mahmood Alzubaidi, Wajdi Zaghouani, Alaa Abd-alrazaq, and Mowafa Househ. 2022b. Arabic chatbot technologies: A scoping review. *Computer Methods and Programs in Biomedicine Update*, 2(100057).

Arfan Ahmed, Nashva Ali, Mahmood Alzubaidi, Wajdi Zaghouani, Alaa A Abd-alrazaq, and Mowafa Househ. 2022c. Freely available arabic corpora: A scoping review. *Computer Methods and Programs in Biomedicine Update*, 2(100049).

Mohammad Al-Fetyani, Muhammad Al-Barham, Gheith Abandah, Adham Alsharkawi, and Maha Dawas. 2023. Masc: Massive arabic speech corpus. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 1006–1013. IEEE.

Khalid Al Khatib, Henning Wachsmuth, Kevin Lang, Jakob Herpel, Matthias Hagen, and Benno Stein. 2018. Modeling deliberative argumentation strategies on wikipedia. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2545–2555.

Abdul Gabbar Al-Sharafi, Mohammad Majed Khader, Mohamad Hamza Al-Sioufy, Wajdi Zaghouani, and Ali Al-Zawqari. accepted 2024. A hybrid annotation model for arabic argumentative debate corpus. In *The Eighth International Conference on Arabic Language Processing, ICALP 2023, Rabat, Morocco, April 19–20, 2024*. Springer.

Latifa Al-sulaiti and Eric Atwell. 2006. The design of a corpus of contemporary arabic. *International Journal of Corpus Linguistics*, 11:135–171.

Nora Al-Twairesh, Rawan Al-Matham, Nora Madi, Nada Almugren, Al-Hanouf Al-Aljmi, Shahad Alshalan, Raghad Alshalan, Nafla Alrumayyan, Shams Al-Manea, Sumayah Bawazeer, et al. 2018. Suar: Towards building a corpus for the saudi dialect. *Procedia computer science*, 142:72–82.

Ali Al-Zawqari, Abdul Gabbar Al-Sharafi, Mohamed Ahmed, Mohammad Majed Khader, and Gerd Vandersteen. accepted 2024. Classifying persuasion modes in arabic debates: A preliminary language model-based analysis. In *The Eighth International Conference on Arabic Language Processing, ICALP 2023, Rabat, Morocco, April 19–20, 2024*. Springer.

Khalid Almeman, Mark Lee, and Ali Abdulrahman Almiman. 2013. Multi dialect arabic speech parallel corpora. In *2013 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA)*, pages 1–6. IEEE.

Maha Alrabiah, A Al-Salman, and ES Atwell. 2013. The design and construction of the 50 million words ksucca. In *Proceedings of WACL'2 Second Workshop on Arabic Corpus Linguistics*, pages 5–8. The University of Leeds.

L Anthony. 2023. Antcconc. https://www.laurenceanthony.net/software/antconc/.

Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, and Danilo Giampiccolo. 2006. The second pascal recognizing textual entailment challenge. In *Proceedings of the second PASCAL challenges workshop on recognizing textual entailment*, volume 7, pages 785–794.

Kaspar Beelen, Timothy Alberdingk Thijm, Christopher Cochrane, Kees Halvemaan, Graeme Hirst, Michael Kimmins, Sander Lijbrink, Maarten Marx, Nona Naderi, Ludovic Rheault, Roman Polyanovsky, and Tanya Whyte. 2017. Digitization of the canadian parliamentary debates. *Canadian Journal of Political Science / Revue canadienne de science politique*, 50(3):pp. 849–864.

Tom Bosc, Elena Cabrio, and Serena Villata. 2016. Dart: A dataset of arguments and their relations on twitter. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1258–1263.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, et al. 2018. The madar arabic dialect corpus and lexicon. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Kareem Darwish, Nizar Habash, Mourad Abbas, Hend Al-Khalifa, Huseein T Al-Natsheh, Houda Bouamor, Karim Bouzoubaa, Violetta Cavalli-Sforza, Samhaa R El-Beltagy, Wassim El-Hajj, et al. 2021. A panoramic survey of natural language processing in the arab world. *Communications of the ACM*, 64(4):72–81.

Mona Diab, Aous Mansouri, Martha Palmer, Olga Babko-Malaya, Wajdi Zaghouani, Ann Bies, and Mohammed Maamouri. 2008. A pilot arabic propbank. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*.

Esin Durmus, Marco Lippi, and Paolo Torroni. 2021. Argumentation mining on news editorials and blog posts in italian. In *Proceedings of the Eighth Italian Conference on Computational Linguistics (CLiC-it 2021)*, pages 1–6.

Ibrahim Abu El-Khair. 2016. 1.5 billion words arabic corpus. *arXiv preprint arXiv:1611.04033*.

Beatriz Fisas, Francesco Ronzano, and Horacio Saggion. 2016. A multi-layered annotated corpus of scientific papers. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3081–3088, Portorož, Slovenia. European Language Resources Association (ELRA).

Marwa Graja, Maher Jaoua, and L Hadrich Belguith. 2010. Lexical study of a spoken dialogue corpus in tunisian dialect. In *The international arab conference on information technology (acit), benghazi–libya*.

Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional lstm. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599.

Lotte Hargrave and Tone Langengen. 2021. The gendered debate: Do men and women communicate differently in the house of commons? *Politics &amp; Gender*, 17(4):580–606.

Annette Hautli-Janisz, Zlata Kikteva, Wassiliki Siskou, Kamila Gorska, Ray Becker, and Chris Reed. 2022. QT30: A corpus of argument and conflict in broadcast debate. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3291–3300, Marseille, France. European Language Resources Association.

Christopher Hidey and Kathy McKeown. 2018. Persuasive influence detection: The role of argument sequencing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Lindsay Katz and Rohan Alexander. 2023. Digitization of the australian parliamentary debates, 1998-2022. *Scientific Data*, 10.

Mohammad Majed Khader. 2020. A digital study on public speaking: Nlp arguments analysis of the first corpus of arabic debates. Master's thesis, Hamad Bin Khalifa University (Qatar).

Sameer Khurana and Ahmed Ali. 2016. Qcri advanced transcription system (qats) for the arabic multi-dialect broadcast media recognition: Mgb-2 challenge. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 292–298. IEEE.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Hamdy Mubarak, Amir Hussein, Shammur Absar Chowdhury, and Ahmed Ali. 2021. Qasr: Qcri aljazeera speech resource–a large scale annotated arabic speech corpus. *arXiv preprint arXiv:2106.13000*.

Elena Musi, Debanjan Ghosh, and Smaranda Muresan. 2018. Towards feasible guidelines for the annotation of argument schemes. In *Proceedings of the First Workshop on Argument Mining*, pages 154–163.

Constantin Orăsan and Richard Evans. 2007. Np animacy identification for anaphora resolution. *Journal of Artificial Intelligence Research*, 29:79–103.

Martha Palmer, Olga Babko-Malaya, Ann Bies, Mona T Diab, Mohamed Maamouri, Aous Mansouri, and Wajdi Zaghouani. 2008. A pilot arabic propbank. In *International Conference on Language Resources and Evaluation*.

Andreas Peldszus and Manfred Stede. 2015. An annotated corpus of argumentative microtexts. In *Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation, Lisbon*, volume 2, pages 801–815.

Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. 2015. A survey of available corpora for building data-driven dialogue systems. *arXiv preprint arXiv:1512.05742*.

Khaled Shaalan, Sanjeera Siddiqui, Manar Alkhatib, and Azza Abdel Monem. 2019. Challenges in arabic natural language processing. In *Computational linguistics, speech and image processing for arabic language*, pages 59–83. World Scientific.

Sylvia Shaw. 2000. Language, gender and floor apportionment in political debates. *Discourse & society*, 11(3):401–418.

Jacky Visser, Bartosz Konat, Rory Duthie, Marcin Koszowy, Katarzyna Budzynska, and Chris Reed. 2020. Argumentation in the 2016 us presidential elections: Annotated corpora of television debates and social media reaction. *Language Resources and Evaluation*, 54(1):123–154.

Marilyn A Walker, Jean E Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, volume 12, pages 812–817. Istanbul, Turkey.

Wajdi Zaghouani. 2014. Critical survey of the freely available arabic corpora. *OSACT, Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*.

Wajdi Zaghouani, Abdelhamid Ahmed, Xiao Zhang, and Lameya Rezk. 2024. Qcaw 1.0: Building a qatari corpus of student argumentative writing. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*.

Wajdi Zaghouani, Mona Diab, Aous Mansouri, Sameer Pradhan, and Martha Palmer. 2010. The revised arabic propbank. In *Proceedings of the fourth linguistic annotation workshop*, pages 222–226.

Wajdi Zaghouani, Nizar Habash, Houda Bouamor, Alla Rozovskaya, Behrang Mohit, Abeer Heider, and Kemal Oflazer. 2015. Correction annotation for non-native arabic texts: Guidelines and corpus. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 129–139.

Wajdi Zaghouani, Behrang Mohit, Nizar Habash, Ossama Obeid, Nadi Tomeh, Alla Rozovskaya, Noura Farra, Sarah Alkuhlani, and Kemal Oflazer. 2014. Large scale arabic error annotation: Guidelines and framework. In *Proceedings of the Ninth Language Resources and Evaluation Conference*.

Amy Zhang, Cristian Danescu-Niculescu-Mizil, Jure Lee, Jilin Chen, Tianze Hua, and Dario Taraborelli. 2021. Conversations gone awry: Detecting early signs of conversational failure. *arXiv preprint arXiv:2101.06814*.