The NakbaEcho Dataset: From Oral Testimonies to a Transcribed Arabic History Corpus

Batool Balah[1], Mahmoud Fawzi[2], Houda Elmimouni[3], Walid Magdy[2]
[1]Maqsam, Jordan
[2]The University of Edinburgh, United Kingdom
[3]University of Manitoba, Canada
batoolnajeh@gmail.com, m.f.g.ibrahim@sms.ed.ac.uk
houda.elmimouni@umanitoba.ca, wmagdy@inf.ed.ac.uk

Abstract

We present *NakbaEcho*, a large-scale dataset derived from Palestinian testimonies about the 1948 Nakba. The resource is constructed from transcribing over 2,180 hours of recorded interviews gathered through the Palestine Remembered Oral History index and linked to multiple repositories, including the Palestinian Oral History Archive (POHA) and YouTube-hosted interviews. We harmonize interview-level metadata and generate timestamp-aligned transcripts from the original Arabic recordings using an automatic transcription pipeline configured for Palestinian Arabic. The dataset includes speaker-labeled segments and auxiliary annotations designed to support downstream research in Arabic speech processing, natural language processing, digital humanities, and oral-history analysis. We further provide an empirical evaluation of transcription quality and descriptive analyses of geographic coverage, narrator demographics, and linguistic patterns across the corpus. *NakbaEcho* contributes a structured and extensible computational resource for studying Palestinian oral testimony while expanding the availability of dialectal Arabic materials for speech, text, and social research, supporting reproducibility, benchmarking, and future cross-lingual and multimodal research directions.

Keywords: Oral history, Arabic, Levantine, Palestine, Nakba, automatic transcription, digital humanities

1. Introduction

Oral history has been vital to documenting, preserving, and transmitting Palestinian memory of the 1948 Nakba (Nur, 2008). Testimonies by survivors and their descendants across archives and initiatives document displacement, loss, return, and everyday life before and after 1948. These recordings serve as both historical evidence and linguistic records of spoken Palestinian Arabic across generations.

Despite the richness, Nakba's oral histories remain difficult to process at scale. Much of the material exists only as dispersed audio/video recordings, with limited transcription suitable for NLP or corpus-based analysis(Lamar et al., 2025). Existing computational work has therefore relied either on small translated subsets of archival interviews or on transcripts prepared for specific studies (Ashqar, 2025), rather than on a unified large-scale Arabic speech-and-text resource.

Nakba testimonies are typically produced in spoken Arabic, mostly in Palestinian dialect rather than Modern Standard Arabic (MSA), and they contain historical and geographical references to villages, districts, displacement routes, kinship relations, and political actors. These characteristics make the material particularly valuable for research on dialectal Arabic speech recognition, oral-history processing, narrative analysis, information extraction, and digital humanities. Nevertheless, they also make the data difficult to process using resources designed for edited text or for contemporary broadcast speech (Hamed and Zaidkilani, 2025).

In this paper, we introduce The *NakbaEcho* dataset, a structured textual dataset of interviews transcribed from the *Palestine Remembered*[1] community. The dataset provides useful metadata from the interviews that facilitate direct computational research, including places of origin, recording dates, and significant mentions. It assembles a structured transcription resource with timestamped segments, speaker labels, and auxiliary annotations that enable alignment between text and audio. We show some preliminary interesting patterns that motivate future research directions on the dataset.

The *NakbaEcho* dataset is provided for free to the research community[2]. The current repository release corresponds to the 708 transcripts analyzed in this paper, and additional transcribed hours will be added in future updates as processing continues.We hope it would facilitate producing future studies in the fields of NLP, social science, and political science to analyze the narrative on Nakba through the lens of the people who lived it.

[1]The Website of Palestine Remembered: https://www.palestineremembered.com/index.html
[2]https://github.com/batouln/NakbaEcho

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 1–22
11 May 2026.

2. Related Work

2.1. Historical Context

The Palestinian Nakba of 1948 refers to the mass displacement and dispossession of Palestinians during the war surrounding the end of the British Mandate and the creation of the state of Israel (United Nations Committee on the Exercise of the Inalienable Rights of the Palestinian People, 2025; Encyclopaedia Britannica, 2026; Khalidi, 1992). The term *Nakba*, meaning "catastrophe" in Arabic, refers both to the events of 1948 and to their enduring political, social, and cultural consequences for Palestinians in exile, under occupation, and within present-day Israel (United Nations Committee on the Exercise of the Inalienable Rights of the Palestinian People, 2025; Masalha, 2008; Sa'di and Abu-Lughod, 2007). The rupture of 1948 followed the final years of the British Mandate and the adoption of United Nations General Assembly Resolution 181, which proposed the partition of Palestine into Arab and Jewish states (United Nations General Assembly, 1947; Encyclopaedia Britannica, 2026). During the war, more than 700,000 Palestinians were displaced, and hundreds of villages and urban neighborhoods were destroyed, depopulated, or rendered inaccessible to their inhabitants (United Nations Relief and Works Agency for Palestine Refugees in the Near East, 2026; Khalidi, 1992). *All That Remains* documents the destruction and depopulation of more than 400 Palestinian villages, underscoring the scale of this transformation (Khalidi, 1992). The consequences of 1948 extended far beyond the initial displacement. Palestinian refugees and their descendants became dispersed across the West Bank, Gaza, neighboring Arab countries, and wider diasporas, while the question of return remained central to Palestinian political life and historical memory (Sayigh, 1994; Sa'di and Abu-Lughod, 2007). United Nations General Assembly Resolution 194, adopted in December 1948, established return and compensation as enduring reference points in discussions of Palestinian refugeehood (United Nations General Assembly, 1948; United Nations Relief and Works Agency for Palestine Refugees in the Near East, 2026).

The Nakba, for Palestinians, is not only a historical event but also a framework for understanding ongoing dispossession, exile, and the transmission of collective memory across generations (Masalha, 2008; Sa'di and Abu-Lughod, 2007; Kassem, 2011). Oral history has been central to preserving these experiences where official archives and dominant narratives have marginalized Palestinian perspectives (Sayigh, 1994; Masalha, 2008; Kassem, 2011). Testimony, family narrative, and village-based remembrance remain key to how the Nakba is narrated and sustained socially, including through accounts of life before displacement, expulsion, and meaning of home and return (Sa'di and Abu-Lughod, 2007; Sayigh, 1994; Kassem, 2011). This context situates Nakba's oral recordings as more than speech data:They are records of memory, place, loss, and survival, shaped by the long afterlife of 1948 and the social worlds in which testimony is produced, archived, and transmitted(Sa'di and Abu-Lughod, 2007; Sleiman and Chebaro, 2018; The Nakba Archive, 2002).

2.2. Nakba Oral History Archives

Nakba oral history has been preserved through archival initiatives focused on documentation and access rather than NLP-oriented data release (Nur, 2008). Two important collections are the Nakba Archive (The Nakba Archive, 2002) and the Palestinian Oral History Archive (POHA) (American University of Beirut). The Nakba Archive documents testimonies of Palestinians displaced in 1948 through audio/video interviews, but only a small subset has been transcribed and translated into English, limiting computational use (Lamar et al., 2025). Similarly, POHA preserves over one thousand hours of testimonies with rich metadata, but remains a media archive rather than a text corpus for speech and language research (Sleiman and Chebaro, 2018).

Thus, despite their importance, these archives highlight a persistent challenge: preservation infrastructures do not readily translate into reusable computational corpora. Researchers must often reconstruct transcripts, metadata, and analysis-ready files from heterogeneous sources.

2.3. Computational Work on Nakba

Recent work in Nakba NLP has begun converting parts of existing archives into analyzable resources (Jarrar et al., 2025). Lamar et al. (2025) present a geo-referenced dataset built from translated Nakba Archive interviews, showing how oral-history transcripts can support analysis of place, geography, and displacement. However, it is limited to about thirty English-translated interviews and focuses on manually annotated geographic references. Awad et al. (2025) analyze narrative cohesion in POHA oral histories using derived transcripts, demonstrating the value of large-scale collections for studying collective memory, shared origin, and gendered narratives. However, their contribution remains analytical and does not provide a publicly released, unified Arabic transcript corpus. Other resources draw on literary rather than oral sources. Hamed and Zaidkilani (2025) introduce a topic-classification corpus from eight written Nakba short stories. While useful, it differs from oral testimony in genre, language, and interactional structure, lacking features such as

spontaneous speech, interviewer–interviewee dynamics, and disfluency. Overall, existing work highlights the value of Nakba narratives but also their fragmentation: translated subsets, study-specific transcripts, and small literary corpora support specific tasks, while a reusable large-scale Arabic oral-history corpus remains lacking.

2.4. Arabic speech and dialect resources

The scarcity of large, well-structured Arabic speech resources has been widely noted, especially for dialectal and conversational varieties (Djanibekov et al., 2025). Recent datasets such as Casablanca (Talafha et al., 2024) have helped address this problem by providing dialectal Arabic speech with transcriptions and annotations for downstream tasks. These resources are important because they move beyond Modern Standard Arabic and highlight the need for datasets grounded in naturally occurring spoken language (Keleg et al., 2023). However, existing Arabic speech corpora generally come from conversational media, interviews, online videos, or benchmark-oriented collection settings rather than from historical oral testimony (Talafha et al., 2024). As a result, they do not capture the linguistic and documentary properties that make oral-history interviews distinctive: long-form remembrance, intergenerational narrative, historically specific vocabulary, place-rich discourse, and close coupling between archival metadata and speech content (Sa'di and Abu-Lughod, 2007; Sayigh, 1994; Sleiman and Chebaro, 2018). Nakba testimony therefore occupies a unique space at the intersection of dialectal Arabic speech technology, digital archives, and computational humanities (Jarrar et al., 2025; Lamar et al., 2025; Awad et al., 2025).

3. Data & Methods

3.1. Data Statistics

The *Palestine Remembered* Oral History index contains 1,341 interview entries, where a single entry may include more than one speaker (e.g., a narrator and their spouse). As shown in Table 1, recordings span the period from 1994 to the present, with durations ranging from 10 minutes to over 25 hours, yielding more than 3,000 hours of content.

Of these, **708 interviews** (approximately **2,180 hours**) were successfully transcribed to form the *NakbaEcho* dataset.

Geographic coverage The transcribed interviews cover all 14 historical Palestinian districts. Safed accounts for the largest share (213 interviews, 31.6%), followed by Haifa (124, 18.4%), Ramla (52, 7.7%), and Tiberias (47, 7.0%). Together, Safed and Haifa represent just over half (50.1%) of all transcribed interviews.

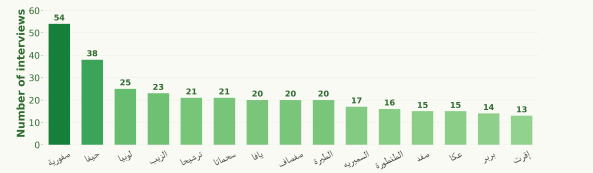

Figure 1: Count of interviews per village.

At the village level, the dataset covers 239 unique villages. Of these, 94 are represented by a single interview, reflecting broad geographic coverage alongside an uneven distribution of documentation across locations.

Narrator demographics The 708 transcribed interviews involve 719 individual narrators, as some recordings include multiple speakers. Gender was automatically detected using Claude Opus from narrator names and manually reviewed for gender-ambiguous Arabic names (e.g., وسام, شحادة, نعمة, جهاد), yielding 536 male narrators (74.5%) and 183 female (25.5%). To validate these labels, we compared them against speaker-level gender annotations independently produced by the Gemini transcription model, which infers narrator gender from acoustic features during diarization. Despite relying on entirely different signals—name morphology versus voice characteristics—the two methods produce near-identical distributions (74.6% male, 25.4% female) and agree on 98.6% of the 579 directly comparable cases (8 mismatches), supporting the reliability of both annotations.

Gender representation differs substantially by source. POHA recordings approach parity (44.1% female), whereas YouTube-sourced interviews are heavily male-dominated (8.7% female), reflecting the different collection contexts of each archive.

Birth years are available for 582 narrators and range from 1897 to 1962, with a median of 1929. The majority (488 of 582, 83.8%) were born between 1920 and 1939, indicating that most narrators were adolescents or young adults at the time of the 1948 Nakba.

3.2. Transcription Pipeline

Here we describe the transcription pipeline used to convert the recordings into text with rich metadata.

Source	Recording period	# interviews	Duration (mins)	Duration (hrs)
POHA	1994–2009	349	25,852	≈431
YouTube	2003–2026	323	102,878	≈1,715
Zochrot	2002–2026	36	2,050	≈34
Total	1994-2026	708	130,779	≈2,180

Table 1: Source-level statistics for the transcribed *NakbaEcho* dataset (708 interviews, ≈2,180 hours).

Start	End	Speaker	Transcript
00:00:22.500	00:00:29.800	SPK_01	السلام عليكم ورحمة الله وبركاته نرحب بكم في حلقة جديدة
00:01:45.000	00:01:49.000	SPK_01	بداية حابين تعرفنا باسمك الكامل الله يخليك
00:01:49.000	00:01:52.000	SPK_02	عبد المجيد محمد أبو سعيد

Figure 2: Examples of transcribed segments and the fields associated with each segment.

Timestamp	Speaker	Gender	Text	Emotion
[00:01:31 → 00:01:37]	SPEAKER_01	Male	عمي تسمح لنا نباشر بأسئلة القسم الأول؟	neutral
[00:12:13 → 00:12:17]	SPEAKER_00	Male	طيب. يا سيدي الله يديم عليك الصحة والعافية إن شاء الله	happy
[01:14:24 → 01:14:35]	SPEAKER_02	Male	سنة الغلاء تروح تجيب بواخر قمح وترميهن بحيفا بالبحر، شايف كيف؟	sad

Table 2: A snapshot of transcribed segments as they appear in the corpus, showing speaker labels, timestamps, and emotion annotations.

3.2.1. Segments

All audio files were automatically transcribed using the Gemini 3 Pro API[3]. The model was configured for dialectal Palestinian Arabic, with explicit constraints to preserve spoken language without normalization into Modern Standard Arabic (MSA).

To minimize hallucination, the transcription prompt enforced: 1) No content invention; unintelligible speech marked as `[inaudible]`; 2) Dialect preserved as spoken; 3) Proper nouns (including village names) retained without correction; and 4)Absolute timestamps required for each segment.

Audio files were processed in chunks to accommodate long interviews. Timestamps were recorded as absolute positions relative to the original audio file. Speaker diarization was performed automatically by the transcription model. Speaker identifiers are consistent within each processed audio chunk. However, speaker identity resolution across chunk boundaries was not performed. Therefore, speaker labels are locally consistent but not globally merged across entire interviews. As shown in Figure 2, each transcript was structured into segments, with each segment showing its `start` timestamp, `end` timestamp, `speaker`, and verbatim `text`. All timestamps follow a strict `HH:MM:SS.mmm` format to enable alignment with the original audio.

[3]Gemini 3 Pro: https://ai.google.dev/gemini-api/docs/gemini-3

3.2.2. Metadata

Speaker metadata: Table 2 shows the information generated for each speaker appearing within a processed chunk. The fields are divided into global ones like `gender` and local ones per segment like `emotion`, which can be *neutral*, *sad*, *angry*, *fear*, *happy*, or *tired*.

These attributes were model-inferred and were not manually validated. They should therefore be interpreted as automatic annotations rather than ground-truth demographic labels.

Literal Mention Flags: To support computational analysis of historical references, the presence of the following three terms was evaluated both on the chunk and the segment level: 1) Nakba, 2) Jews/Israel, and 3) British Mandate.

For each chunk, a boolean flag indicates whether the term is mentioned; if mentioned, a strictly literal summary is generated describing the immediate context of the reference, otherwise, the summary field is left empty. At the segment level, corresponding boolean flags indicate whether a given segment had the term mentioned or not. The annotation strategy follows a literal-first principle: summaries reflect only explicitly stated content without interpretive abstraction.

The terminology reflects the original language used by narrators, including historically situated references such as اليهود; these are preserved verbatim and should be interpreted within their historical

and discursive context rather than as standardized contemporary labels.

3.3. Transcription Quality Evaluation

To assess transcription quality, we conducted a manual evaluation on a stratified sample of 22 clips, each approximately 15 minutes in length (around **5.5 hours** in total). The sample reflects dataset diversity across dialect, recording conditions, generation, and gender.

The sample includes speakers from major Palestinian regions (Galilee, Haifa, Gaza, Yaffa, Ramla–Lyd), with varied recording conditions and generational coverage. **Recording conditions** were also balanced: **9 clips** originate from **POHA** recordings, while the remaining **13 clips** were sourced from YouTube, PalRemembered and Zochrot sources. The narrators span birth years between **1897 and 1937**, covering four generational cohorts (**pre-1915 through 1935–1944**), and represent individuals who directly experienced the events of the 1948 Nakba. The **gender distribution** in the evaluation set (**18 male** and **4 female** speakers) reflects that of the broader corpus.

Transcription quality was measured using Word Error Rate (WER), computed by comparing the automatic transcripts against manually corrected references. The system achieves an **overall micro-averaged WER of 7.00%**, with a **mean per-clip WER of 7.25%** and a **median WER of 4.34%**. Performance is consistent across the sample: **59% of clips fall below 5% WER**, and **86% fall below 10% WER**. While most clips exhibit low error rates, a small number of outliers contribute to a **standard deviation of 9.59%**, with WER ranging from **0.78% to 46.22%**.

Error analysis shows that **deletions (1,018)** constitute the largest proportion of errors, followed by **substitutions (833)** and **insertions (305)**. These results indicate that the transcription system achieves high accuracy overall, with most errors arising from missed tokens rather than incorrect or spurious insertions, which is consistent with the challenges of transcribing dialectal speech under variable acoustic conditions.

4. Ethical Considerations

This dataset is derived from oral testimonies documenting lived experiences of displacement and loss during and after the 1948 Nakba. These materials originate from archival initiatives aimed at preserving historical memory, and this work builds on those efforts by enabling computational access rather than replacing the original archival context.

All transcripts and annotations are generated automatically and may contain errors or simplifications. In particular, model-inferred attributes (e.g., emotion labels and speaker characteristics) are not manually validated and should not be interpreted as ground truth. Analyses based on these annotations should therefore be treated as exploratory.

The dataset preserves the original language used by narrators, including historically situated terminology. These expressions reflect the context of the testimonies and are included as documentary evidence rather than normative labels.

We encourage responsible use of this resource. The dataset is intended for research in language technology and digital humanities, and should be used with attention to historical context and the limitations of automated processing. In particular, references to social, political, or religious groups (e.g., اليهود), "Jews", reflect historically situated usage in the testimonies. These terms are retained to preserve the integrity of the source material and should not be interpreted as generalized or decontextualized labels.

5. NakbaEcho Dataset Analysis

5.1. NakbaEcho Dataset Summary Stats

Table 3 summarises the main statistics of the transcribed *NakbaEcho* dataset. The transcription process enables the extraction of additional descriptive signals from the testimonies, including the number of exchanges per interview, the frequency of named entities (NEs), speaker gender distribution, and inferred emotional attributes.

It is important to emphasize that this analysis is exploratory in nature. The reported statistics rely on automatically generated transcripts and downstream processing pipelines (e.g., NER, speaker attribution, and emotion detection), which are subject to errors and biases. Factors such as transcription inaccuracies, diarization limitations, and model imperfections may affect the reliability of these derived metrics. Therefore, the results should be interpreted as indicative trends rather than precise measurements.

5.2. Emotion, Lexicon, and Gender

Overall Emotion Distribution: Figure 3 presents the distribution of emotion labels assigned to transcript segments across the 708 interviews. The vast majority of segments (988,941; 94.1%) carry a NEUTRAL label, consistent with the long-form, conversational nature of oral-history testimony. Among the 62,219 non-neutral segments (5.9%), SAD constitutes the largest share (33,581; 54.0%), followed by HAPPY (18,839; 30.3%), ANGRY (5,049; 8.1%), and FEAR (3,202; 5.1%). The remaining categories—TIRED, PROUD, SERIOUS, CURIOUS, IN-

Source	Interviews	Exchanges	Tokens	NEs	Male%	Female%
POHA	349	126,506	2,203,952	109,413†	51.5	48.5
YouTube	323	911,144	10,445,044	626,260†	97.7	2.3
Zochrot	36	13,510	201,760	11,959†	54.8	45.2
Total	708	1,051,160	12,850,756	747,632†	91.6	8.4

Table 3: Statistics on the transcribed *NakbaEcho* dataset by source. Gender percentages reflect segment-level speaking time distribution.

QUISITIVE, and NOSTALGIC—together account for 2.3% of non-neutral segments.

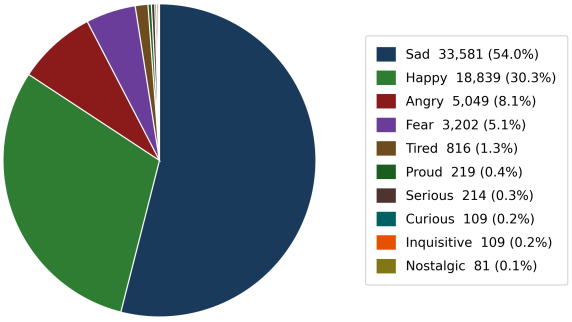

Figure 3: Distribution of non-neutral emotion labels across transcript segments (n = 708 interviews; 62,219 non-neutral segments, 5.9% of total). Neutral segments (94.1%) are excluded to highlight the relative distribution of affective categories.

The predominance of SAD segments reflects the thematic focus of the corpus: testimonies predominantly address displacement, loss, violence, and forced migration during the events of 1948. Because the dataset consists of first-person recollections of these experiences, segments describing expulsion, separation, destruction, and departure are expected to be frequent, resulting in a higher proportion of sadness-associated labels.

Lexical Patterns by Emotion: We examined how emotions are linguistically realized across the corpus. Segments labelled as SAD are dominated by displacement and loss vocabulary: طلعنا (we left), البلد (the town/village), الناس (the people), and اليهود appear prominently, frequently co-occurring in narrative passages describing departure from villages and the disruption of community life. طلعنا in particular functions as a recurring marker of forced departure across both SAD and FEAR segments.

HAPPY segments are characterized by vocabulary associated with social and communal life. High-frequency terms include كنا (we were), عنا (at our place / with us), and العريس (the groom), which occur in passages recounting weddings, family gatherings, and memories of pre-displacement daily life. The formulaic expression شاء (part of *in shā Allāh*, God willing) also appears prominently, reflecting the embedding of historical narration within everyday devotional language.

Segments labelled as ANGRY show higher prominence of collective conflict actors: اليهود, العرب (the Arabs), فلسطين (Palestine), and جيش (army) all rank among the top content words, reflecting descriptions of political confrontation and organised violence. FEAR segments similarly foreground اليهود and علينا (upon us / against us), suggesting narrative passages that describe perceived threat directed at the speaker and their community, alongside قالوا (they said) and قلت (I said), indicating reported speech as a vehicle for conveying danger and urgency.

The distribution of emotion labels and their associated lexicon indicates that emotional expression in the corpus is closely tied to narrative content. Conflict-related events and displacement episodes are predominantly associated with SAD, ANGRY, or FEAR labels, whereas segments recounting domestic life, religious observances, or social celebrations are more frequently associated with HAPPY. This pattern reflects the dual structure of many testimonies: recollections of everyday pre-1948 life alongside accounts of conflict and displacement. The analysis reports surface-level lexical co-occurrence within automatically labelled segments and does not interpret speaker intention beyond the textual evidence. Top content words stratified by emotion label are presented in Appendix E (Table 15).

Lexical Variation by Gender: To examine whether lexical salience differs across speaker demographics, we computed content-word frequency distributions separately for male and female narrators after stop-word removal (male: 8.24M tokens; female: 942K tokens). To test whether observed differences exceed chance variation, we applied log-likelihood ratio (G^2) tests and chi-square tests to all word types appearing at least five times across both subcorpora (110,954 types tested). Frequencies were normalised per 1,000 tokens for reporting purposes, while statistical tests were computed on

(a) Female Narrators

(b) Male Narrators

Figure 4: Word clouds of content terms by gender. Word size is proportional to frequency. Both groups share core narrative anchors (فلسطين, البلد, اليهود), while female speakers show greater prominence of kinship address forms and domestic vocabulary, and male speakers emphasise geographic and spatial terms. Differences are statistically supported via G^2 log-likelihood tests and Bonferroni-corrected chi-square tests (6,161 significant word types; see Table 17).

raw frequency counts. Of the tested types, 23,235 are significant at $p < 0.05$, and 6,161 remain significant after Bonferroni correction ($\alpha = 1.103 \times 10^{-7}$), indicating that the observed differences are statistically robust under the given corpus partition. The full significance table is provided in Appendix F (Table 17).

Both groups share core narrative anchors, including اليهود, البلد, and references to villages and land, indicating strong thematic overlap in how displacement and conflict are narrated. Statistically significant differences emerge in lexical prominence across groups. Male-favoured terms are dominated by spatial and geographic vocabulary: القرية (the village) appears at 2.46 per 1,000 tokens in male speech versus 0.20 in female speech ($G^2 = 3{,}104.8$); منطقة (area/region, $G^2 = 728.2$), متر (metre, $G^2 = 665.3$), and حوالي (approximately, $G^2 = 735.7$) follow a similar pattern, reflecting detailed descriptions of village layout, land boundaries, and geographic orientation. Female-favoured terms are predominantly kinship address forms and domestic vocabulary: يما (mother) appears at 1.08 per 1,000 tokens in female speech versus 0.04 in male speech ($G^2 = 3{,}170.7$); إنتي (you.FEM, $G^2 = 1{,}725.9$), بنت (girl/daughter, $G^2 = 1{,}676.2$), إمي (my mother, $G^2 = 1{,}552.3$), and جوزي (my husband, $G^2 = 1{,}354.5$) follow the same pattern, alongside movement verbs such as طلعنا (we left, $G^2 = 1{,}056.4$), reflecting a narrative register centred on family relationships and lived household experience.

These differences should not be interpreted as categorical distinctions but rather as statistically supported distributional tendencies within a shared narrative framework. We further note that the two gender subcorpora differ substantially in source composition: female segments are concentrated in POHA recordings (44.1% female), while male segments are dominated by YouTube-sourced interviews (97.7% male). As a result, the observed lexical contrasts may reflect not only speaker gender but also differences in recording context, interview structure, collection conditions, and dialectal variation. The reported results should therefore be interpreted as distributional differences between subcorpora rather than purely gender-driven effects.

Appendix F presents the top 15 content words per group (Table 16) and the top Bonferroni-significant words by gender (Table 17).

5.3. Named Entity Analysis

To characterize the informational density of the transcribed corpus, we applied named entity recognition (NER) using the CAMeL-Lab Arabic BERT model (Inoue et al., 2021), fine-tuned on the ANERcorp dataset, to all 708 interviews. The NER component extracts entity mentions and assigns each mention to one of four coarse-grained types: locations (LOC), persons (PERS), organizations (ORG), and a residual MISC category. We report mention counts (how often an entity is referenced) as an indicator of referential density, and unique entity counts (how many distinct entities appear) as an indicator of lexical richness. We note that BERT sub-word tokenisation occasionally produces fragment artefacts; the curated entity lists in Appendix A exclude such fragments.

Overall Entity Yield:

Type	Count	%
LOC	309,655	41.4
PERS	302,167	40.4
MISC	102,972	13.8
ORG	32,838	4.4
Total	**747,632**	

Table 4: Entity mention distribution ($n = 708$).

Named entity recognition was performed on all

708 transcripts, we extracted 747,632 entity mentions. The distribution by type is shown in Table 4. Locations and persons together account for 81.8% of all extracted mentions, consistent with oral-history narratives that foreground spatial anchoring (villages, towns, routes of displacement) and interpersonal reference (family members, community figures, and named individuals).

Top Locations: The most frequently mentioned locations (detailed in Appendix A) reflect both narrators' places of origin and salient nodes in the geography of displacement. Palestine (12,115 mentions), Jaffa (10,252), and Haifa (10,098) lead the ranking, followed by Gaza (4,617) and Jerusalem (4,442), consistent with the prominence of major cities in accounts of 1948 and its aftermath. The appearance of Lydda (3,936) and Ramla (2,885) reflects the well-documented mass expulsion from these twin cities. Several non-Palestinian locations also appear prominently, including Jordan (3,003), Lebanon (2,893), Britain (2,824), and Egypt (2,333), reflecting transnational trajectories of displacement and the roles of regional states and colonial actors referenced in testimony.

Top Persons: The results are dominated by common first names and *mukhtar*/headman references. Muhammad (4,302) is the most frequent person entity, followed by المختار (al-Mukhtār, the village headman; 2,172) and the *kunyā* form Abū Muáammad (1,727). Among historically identifiable figures, results are consistent with the oral register of Palestinian testimony, where social relations and local community networks are foregrounded through first names, kinship-based reference forms, and community titles.

Top Groups & Organizations: Among organisations, UNRWA (وكالة الغوث, *Wakālat al-Ghawth*; 938 mentions) is the most frequent, followed by the Red Cross (الصليب الأحمر, *al-áalīb al-Aámar*; 445) and the United Nations (الأمم المتحدة, *al-Umam al-Muttaáida*; 318). UNRWA also appears under its Arabic acronym الأونروا (144), bringing its combined count to over one thousand. The Arab Bank (البنك العربي, *al-Bank al-Arabī*; 100), the Arab League (الجامعة العربية, *al-Jāmia al-Arabiyya*; 93), and al-Azhar (الأزهر; 83) also appear, reflecting the institutional framework of post-1948 refugee life—including relief provision, education, and international legal and diplomatic structures—as narrated by the interviewees.

Lexical Salience beyond Entity Types:

While the preceding analysis organises entities by semantic class, a complementary perspective is obtained by examining overall lexical frequency independent of entity taxonomy. We aggregated the most frequent substantive terms in the corpus through unigram, bigram, and trigram frequency analysis following stopword and discourse-marker removal. This approach captures recurrent lexical items that structure narration beyond formal entity categories.

Geographical references constitute the most salient lexical stratum. Major urban centres such as حيفا (Haifa) and يافا (Jaffa) occur at high frequency, alongside generic spatial expressions including البلد (the town/village) and القرية (the village), indicating that spatial anchoring is central to testimonial discourse. Markers of social organisation and kinship are likewise prominent: أبو (*Abū*, "father of") and المختار (village headman) reflect the interpersonal and community-based structure of oral narration. Conflict-related vocabulary forms a further major axis of lexical recurrence. Frequently occurring items include اليهود (19,090 occurrences across 649 interviews), جيش الإنقاذ (Arab Salvation Army, 1,399 bigram mentions), and references to British and regional military actors. Religious–cultural formulae, including الله يرحمه and إن شاء الله, demonstrate the embedding of historical narration within everyday devotional language. Temporal markers—most notably references to 1936 and 1948—anchor personal recollection within shared historical chronology. Taken together, the distribution of high-frequency lexical items suggests that the corpus is organised around four interrelated referential domains: locality, social relations, conflict actors, and historical time. These domains emerge inductively from frequency patterns rather than from predefined thematic categories.

Thematic Categories & Illustrative Evidence:

The thematic annotations assigned to transcript segments capture recurrent narrative foci across the corpus. The three most frequent categories—**Nakba**, **Jews/Israel**, and **British Mandate**—are summarised below (detailed examples in Appendix C).

The **Nakba** category comprises segments referring to the events of 1947–1949 and their immediate aftermath, including displacement, village depopulation, military confrontation, and early refugee experience. Nakba-flagged segments total 50,606 across 683 of the 708 processed interviews—a 96.5% prevalence rate that underscores the near-universal presence of displacement narrative across the corpus, regardless of source, recording date, or narrator background. These segments typically combine spatial references (routes of flight, destinations, camps) with descriptions of family separation, material loss, and uncertainty. A recurrent narrative element concerns expectations of return, frequently framed through references to anticipated

Arab military intervention.

The **Jews/Israel** category includes segments in which narrators refer to Jewish armed groups, settlers, or the state of Israel as actors in the described events. The most frequent lexical item within this category is اليهود (*al-Yahūd*), occurring 19,090 times across 649 interviews and representing the most frequent conflict-related term overall (Appendix B). Additional expressions include العصابات الصهيونية (Zionist gangs, 513 mentions) and جيش الإنقاذ (Arab Salvation Army, 1,399 mentions). These forms are reproduced verbatim from the original testimonies and reflect narrators' own historical and linguistic framing of events, where distinctions between religious, ethnic, and national categories may differ from contemporary usage. Jews/Israel-flagged segments total 32,671 across 670 interviews (94.6%), and predominantly describe armed confrontation, detention, killing, or destruction of property, frequently co-occurring with sadness- or seriousness-labelled segments in the emotion annotation layer.

The **British Mandate** category encompasses references to the period of British rule in Palestine (1920–1948). Narrators refer to the British both colloquially (الانجليز) and institutionally (الحكومة البريطانية, الانتداب البريطاني, 1,081 bigram mentions; حكومة الانتداب البريطاني, 326 trigram mentions). These segments total 16,422 across 609 interviews (86.0%) and describe military searches, suppression of revolt, economic regulation, and arms confiscation. The emotional annotation layer indicates frequent co-occurrence with seriousness or anger labels.

Across categories, lexical choices reflect the vernacular Palestinian Arabic register in which the testimonies were delivered. The corpus preserves these terms as documentary evidence of historical memory and narrative framing. Appendix C presents randomly sampled segments illustrating how spatial references, named events, and institutional actors surface in context, and how emotional register varies across thematic domains.

Interview Question Analysis: To characterise interviewer behaviour and narrative scaffolding, we extracted and categorised 250 distinct question forms across the corpus (detailed in Appendix D). A substantial proportion of distinct question forms (70 out of 250) consists of backchannel or filler prompts, reflecting the conversational and oral-history format of the interviews. Substantive informational categories are dominated by questions concerning village geography and layout (44 distinct forms), daily life and economy (32), and name clarification (16), indicating systematic efforts to document spatial organisation, social structure, and personal identity.

We further examined thematic alignment between interviewer questions and corpus-level themes. Of 300,842 question-bearing segments, 3.7% (11,162) are associated with the Nakba theme, 2.2% (6,566) with Jews/Israel, and 1.2% (3,606) with the British Mandate. These proportions reflect the inclusion of YouTube-sourced interviews, which follow less structured thematic protocols than the POHA archival collection. Nevertheless, the relative prominence of Nakba-related question segments confirms that interview protocols are structured to foreground events of 1947–1949 and their aftermath. At the same time, the presence of detailed prompts regarding village layout, agriculture, education, and religious practice demonstrates that interviewers systematically reconstruct pre-displacement social life. The resulting corpus therefore reflects both event-centred historical documentation and ethnographic reconstruction of everyday life.

6. Limitations

While *NakbaEcho* provides a large-scale resource for analyzing Palestinian oral testimony, several limitations remain. Transcription quality is evaluated on a limited sample and should be considered preliminary; performance may vary with audio conditions, dialects, and speakers. Several annotations (e.g., emotion, speaker attributes, and mention flags) are automatically generated and not fully validated, and may reflect model biases. Speaker and gender attribution may contain errors due to diarization and name-based inference, and observed lexical differences may reflect source composition rather than true group effects. Finally, variation in recording conditions and interview styles across sources may influence linguistic patterns. Future work will focus on improving evaluation and annotation reliability.

7. Conclusion & Future Work

In this work, we present a novel transcribed dataset for Nakba narratives that opens up a lot of opportunities for information extraction through text processing and analysis. We show through our sample analyses the presence of potential research directions that can utilize the extracted fields and texts. Future work should leverage this data to inspect the demographics and the circumstances surrounding the event of the Nakba. It should also inspect the similarities and the differences between the patterns associated with the Nakba and other later similar events.

8. Bibliographical References

American University of Beirut. Palestinian oral history archive (poha). https://www.aub.edu.lb/ifi/Pages/poha.aspx. Accessed: 2026-02-27.

Huthaifa I Ashqar. 2025. Sentiment analysis of nakba oral histories: A critical study of large language models. In *Proceedings of the first International Workshop on Nakba Narratives as Language Resources*, pages 30–36.

Ghadir A Awad, Tamara N Rayan, Lavinia Dunagan, and David Gamba. 2025. Collective memory and narrative cohesion: A computational study of palestinian refugee oral histories in lebanon. In *Proceedings of the first International Workshop on Nakba Narratives as Language Resources*, pages 83–102.

Amirbek Djanibekov, Hawau Olamide Toyin, Raghad Alshalan, Abdullah Alatir, and Hanan Aldarmaki. 2025. Dialectal coverage and generalization in arabic speech recognition. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29490–29502.

Encyclopaedia Britannica. 2026. 1948 arab–israeli war. https://www.britannica.com/event/1948-Arab-Israeli-War. Accessed: 2026-02-27.

Osama Hamed and Nadeem Zaidkilani. 2025. Arabic topic classification corpus of the nakba short stories. In *Proceedings of the first International Workshop on Nakba Narratives as Language Resources*, pages 48–55.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in arabic pre-trained language models.

Mustafa Jarrar, Nizar Habash, Mo El-Haj, Amal Haddad Haddad, Zeina Jallad, Camille Mansour, Diana Allan, Paul Rayson, Tymaa Hammouda, and Sanad Malaysha. 2025. Proceedings of the first international workshop on nakba narratives as language resources. In *Proceedings of the first International Workshop on Nakba Narratives as Language Resources*.

Fatma Kassem. 2011. *Palestinian women: Narrative histories and gendered memory*. Bloomsbury Publishing.

Amr Keleg, Sharon Goldwater, and Walid Magdy. 2023. Aldi: Quantifying the arabic level of dialectness of text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10597–10611.

Walid Khalidi. 1992. All that remains. *The Palestinian villages occupied and*.

Annie K Lamar, Rick Castle, Carissa Chappell, Emmanouela Schoinoplokaki, Allene M Seet, Amit Shilo, and Chloe Nahas. 2025. Cognitive geographies of catastrophe narratives: Georeferenced interview transcriptions as language resource for models of forced displacement. In *Proceedings of the first International Workshop on Nakba Narratives as Language Resources*, pages 18–29.

Nur Masalha. 2008. Remembering the palestinian nakba: Commemoration, oral history and narratives of memory. *Holy Land Studies*, 7(2):123–156.

Masalha Nur. 2008. Remembering the palestinian nakba: Commemoration, oral history and narratives of memory. *Holy Land Studies*, 7(2):123–156.

Ahmad H Sa'di and Lila Abu-Lughod. 2007. *Nakba: Palestine, 1948, and the claims of memory*. Columbia University Press.

Rosemary Sayigh. 1994. *Too Many Enemies: The Palestinian Experience in Lebanon*. Zed Books, London.

Hana Sleiman and Kaoukab Chebaro. 2018. Narrating palestine: The palestinian oral history archive project. *Journal of Palestine Studies*, 47(2):63–76.

Bashar Talafha, Karima Kadaoui, Samar Mohamed Magdy, Mariem Habiboullah, Chafei Mohamed Chafei, Ahmed Oumar El-Shangiti, Hiba Zayed, Mohamedou Cheikh Tourad, Rahaf Alhamouri, Rwaa Assi, et al. 2024. Casablanca: Data and models for multidialectal arabic speech recognition. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21745–21758.

The Nakba Archive. 2002. Nakba archive. https://www.nakba-archive.org/. Accessed: 2026-02-27.

United Nations Committee on the Exercise of the Inalienable Rights of the Palestinian People. 2025. About the nakba. https://www.un.org/unispal/about-the-nakba/. Accessed: 2026-02-27.

United Nations General Assembly. 1947. Future government of palestine: United nations general assembly resolution 181 (ii). https://digitallibrary.un.org/record/210008. Adopted 29 November 1947.

United Nations General Assembly. 1948. Palestine—progress report of the united nations mediator: United nations general assembly resolution 194 (iii). https://www.unrwa.org/content/resolution-194. Adopted 11 December 1948.

United Nations Relief and Works Agency for Palestine Refugees in the Near East. 2026. Palestine refugees. https://www.unrwa.org/palestine-refugees. Accessed: 2026-02-27.

A. Appendix: Top Named Entities

This appendix reports the most frequent named entities in the corpus by category.

Table 5: Top named entities by mention frequency across four NER types ($n = 708$ interviews, 747,632 total mentions). Entity text is shown in the original Arabic. Sub-word tokenisation fragment artefacts have been excluded from the ranked lists. MISC and ORG lists are truncated where fragment artefacts begin to dominate the rankings.

LOC — Locations (309,655; 41.4%)				PERS — Persons (302,167; 40.4%)			
#	Entity	Gloss	Freq.	#	Entity	Gloss	Freq.
1	فلسطين	Palestine	12,115	1	محمد	Muáammad	4,302
2	يافا	Jaffa	10,252	2	المختار	the headman	2,172
3	حيفا	Haifa	10,098	3	أبو محمد	Abū Muáammad	1,727
4	غزة	Gaza	4,617	4	أحمد	Aámad	1,461
5	القدس	Jerusalem	4,442	5	موسى	Mūsā	1,406
6	اللد	Lydda	3,936	6	علي	Alī	1,401
7	الخليل	Hebron	3,416	7	صالح	áāliá	1,296
8	عكا	Acre	3,064	8	مختار	headman	1,093
9	الأردن	Jordan	3,003	9	محمود	Maámūd	1,069
10	لبنان	Lebanon	2,893	10	عبد الله	Abd Allāh	1,031
11	الرملة	Ramla	2,885	11	حسن	áasan	1,014
12	بريطانيا	Britain	2,824	12	يوسف	Yūsuf	996
13	الناصرة	Nazareth	2,785	13	حسين	áusayn	945
14	مصر	Egypt	2,333	14	إبراهيم	Ibrāhīm	921
15	بيسان	Baysan	2,183	15	أبو علي	Abū Alī	647

MISC — Miscellaneous (102,972; 13.8%)				ORG — Organisations (32,838; 4.4%)			
#	Entity	Gloss	Freq.	#	Entity	Gloss	Freq.
1	الله	God	21,761	1	وكالة الغوث	UNRWA	938
2	ليرة	lira	3,029	2	الصليب الأحمر	Red Cross	445
3	دينار	dinar	1,360	3	الأمم المتحدة	United Nations	318
4	جنيه	pound	784	4	الأونروا	UNRWA (acronym)	144
5	الإسلام	Islam	506	5	البنك العربي	Arab Bank	100
6	الحرب العالمية الثانية	World War II	487	6	الجامعة العربية	Arab League	93
7	القرآن	the Qurān	299	7	الأزهر	al-Azhar	83

Table 6: Most frequent substantive terms in the corpus (708 interviews). Terms drawn from unigram, bigram, and trigram frequency analysis with stopwords and discourse fillers removed. NER-derived entries (†) are drawn from the full 708-interview NER processing run (747,632 total mentions); frequencies may differ slightly from Table 5 due to sub-word tokenisation and aggregation differences. Entries marked (‡) fall outside the top-200 bigram or top-100 trigram cutoff for the 708-interview corpus; values shown are from the initial subset. *Freq.* = raw mention count; *DF* = document frequency.

Category	Arabic	Transliteration & Gloss	Freq.	DF	Src.
Geography & Displacement	†حيفا	*áayfā* (Haifa)	3,922	—	NER
	†فلسطين	*Filasáīn* (Palestine)	3,724	—	NER
	†غزة	*Ghazza* (Gaza)	1,961	—	NER
	†بيسان	*Bīsān* (Baysan)	1,324	—	NER
	†يافا	*Yāfā* (Jaffa)	1,142	—	NER
	†عكا	*Akkā* (Acre)	1,035	—	NER
	البلد	*al-balad* (the town/village)	37,031	659	1-g
	القرية	*al-qarya* (the village)	20,478	433	1-g
	الأرض	*al-ará* (the land)	7,572	618	1-g
	‡بئر السبع	*Bir al-Sab* (Beersheba)	617	13	2-g
	‡بلد الشيخ	*Balad al-Shaykh*	392	13	2-g
	†دير ياسين	*Dayr Yāsīn* (Deir Yassin)	365	—	NER
	عراق المنشية	*Irāq al-Manshiyya*	559	—	2-g
	‡مدينة حيفا	*madīnat áayfā* (city of Haifa)	269	—	2-g

Table 7: Most frequent substantive terms (continued): People & Social Relations.

Category	Arabic	Transliteration & Gloss	Freq.	DF	Src.
People & Social Relations	أبو	*Abū* (father of — *kunyā*)	43,387	666	1-g
	الناس	*al-nās* (the people)	26,467	674	1-g
	الشيخ	*al-Shaykh* (the sheikh/elder)	10,563	523	1-g
	أهل البلد	*ahl al-balad* (people of the village)	3,419	—	2-g
	‡عبد الله	*Abd Allāh*	759	64	2-g
	أهل القرية	*ahl al-qarya* (village people)	2,104	—	2-g
	دار أبو	*dār Abū* (house/clan of Abū…)	2,655	—	2-g
	أبو محمد	*Abū Muáammad*	1,164	—	2-g
	†المختار	*al-Mukhtār* (village headman)	465	—	NER
	عبد الرحمن	*Abd al-Raámān*	1,252	—	2-g
	‡عز الدين القسام	*Izz al-Dīn al-Qassām*	142	20	3-g

B. Appendix: Most Frequent Substantive Terms

This appendix reports the most frequent substantive terms in the corpus, grouped by thematic category and frequency source.

Table 8: Most frequent substantive terms (continued): Conflict & Military.

Category	Arabic	Transliteration & Gloss	Freq.	DF	Src.
Conflict & Military	اليهود	*al-Yahūd* (the Jews)	19,090	649	1-g
	جيش الإنقاذ	*Jaysh al-Inqādh* (Arab Salvation Army)	1,399	—	2-g
	الانتداب البريطاني	*al-Intidāb al-Brīáānī* (British Mandate)	1,081	—	2-g
	الجيش المصري	*al-Jaysh al-Miárī* (Egyptian Army)	756	—	2-g
	الشعب الفلسطيني	*al-Shab al-Filasáīnī* (Palestinian people)	715	—	2-g
	الجيش البريطاني	*al-Jaysh al-Brīáānī* (British Army)	703	—	2-g
	العصابات الصهيونية	*al-Aáābāt al-áahyūniyya* (Zionist gangs)	513	—	2-g
	‡ثورة الـ ٣٦	*Thawrat al-36* (1936 revolt)	117	20	3-g
	الحرب العالمية الثانية	*al-áarb al-Ālamiyya al-Thāniya* (WWII)	344	—	3-g
	‡مع العصابات الصهيونية	*ma al-Aáābāt al-áahyūniyya* (Zionist gangs)	77	13	3-g

Table 9: Most frequent substantive terms (continued): Institutions & Organizations.

Category	Arabic	Transliteration & Gloss	Freq.	DF	Src.
Institutions & Organizations	المدرسة	*al-madrasa* (the school)	7,428	543	1-g
	†وكالة الغوث	*Wakālat al-Ghawth* (UNRWA)	154	—	NER
	†الصليب الأحمر	*al-áalīb al-Aámar* (Red Cross)	146	—	NER
	†الأمم المتحدة	*al-Umam al-Muttaáida* (United Nations)	102	—	NER
	†البنك العربي	*al-Bank al-Arabī* (Arab Bank)	34	—	NER
	†الجامعة العربية	*al-Jāmia al-Arabiyya* (Arab League)	27	—	NER

Table 10: Most frequent substantive terms (continued): Religious & Cultural Expressions.

Category	Arabic	Transliteration & Gloss	Freq.	DF	Src.
Religious & Cultural Expressions	‡الله يرحمه	*Allāh yiráamu* (God rest his soul)	854	86	2-g
	‡إن شاء الله	*in shā Allāh* (God willing)	779	126	3-g
	الحمد لله	*al-áamdu li-llāh* (praise be to God)	1,723	—	2-g
	سيدنا محمد	*Sayyidnā Muáammad* (our Prophet)	1,150	—	2-g
	‡الله سبحانه وتعالى	*Allāh subáānahu wa-taālā* (God Almighty)	89	16	3-g

Table 11: Most frequent substantive terms (continued): Temporal & Historical.

Category	Arabic	Transliteration & Gloss	Freq.	DF	Src.
Temporal & Historical	‡الـ ٣٦	*al-36* (the year 1936)	324	32	2-g
	ثاني يوم	*thānī yōm* (the next day)	1,075	—	2-g
	الثمانية وأربعين	*al-thamāniya wa-arbaīn* (1948)	721	—	2-g
	الستة وثلاثين	*al-sitta wa-thalāthīn* (1936)	543	—	2-g
	‡الـ ٤٨	*al-48* (the year 1948)	271	—	2-g
	ذلك الوقت	*dhālika al-waqt* (that time)	684	—	2-g
	‡ثورة الستة وثلاثين	*thawrat al-sitta wa-thalāthīn* (1936 revolt)	93	15	3-g

C. Appendix: Thematic Examples

This appendix presents illustrative transcript segments for the three main thematic categories discussed in the paper.

Table 12: Illustrative transcript segments for each thematic category, drawn from randomly sampled and emotionally tagged examples. Text is presented in the original Palestinian Arabic as transcribed. Translations are provided by the authors.

Theme	Emotion	Original (Arabic)	Translation (English)
Nakba	Sad	محدش قال لنا اهربوا هي العالم تشردت لحالها.	Nobody told us to flee… People were displaced on their own.
Nakba	Sad	القتل والدمار والتهجير والذبح أمام أعيننا بالعجزة والأطفال. لم يأخذوا فينا إلا ولا رحمة… وكنا نعتقد أن الدول العربية والعرب سوف يقومون بهذا الجهد ويحررونا ونعود بعد قليل.	Killing, destruction, displacement, and slaughter before our eyes—the elderly and the children. They showed us no mercy… We believed the Arab states would make the effort, liberate us, and we would return soon.
Nakba	Fear	حدا بيقدر يقعد؟ هم كانوا يلفلفوا حالهم بدهم، محضرين حالهم بيقولوا بكرا بيلحقنا الضرب.	Could anyone stay? They were gathering themselves, preparing, saying tomorrow the shelling will reach us.
Jews/Israel	Serious	احنا بالاخير لانه في ناس تتعططوا… من دير ياسين لعندنا… عيبعثولنا انه سقطت البلد الفلانية… اليهود طبوا على دير ياسين ودمروا البيوت على اهاليها.	In the end, people came to us in distress… from Deir Yassin… They would send word that such-and-such village had fallen… The Jews attacked Deir Yassin and destroyed the houses on top of their inhabitants.
Jews/Israel	Sad	يومن صار الاجتياح إجا، اتوفى… وين كان؟ كان كمشينه اليهود في… الهراوي وفي النبي يوشع.	The day the invasion came, he died… Where was he? The Jews had detained him in… al-Hirawi and in Nabi Yusha.
Jews/Israel	Sad	مسكوهم اليهود.	The Jews caught them.
British Mandate	Serious	ويفتشوا على أشخاص معينين، على ثوار، على أسلحة… بعدين كانوا يقولوا إنه اللي كان يمسك عنده مجرد فشكة كانوا يقتلوه.	They would search for specific people, for revolutionaries, for weapons… They said whoever was caught with even a single cartridge would be killed.
British Mandate	Serious	فرنسا بتحب تزيح بريطانيا من المنطقة، وبريطانيا بتحب تزيح فرنسا من المنطقة.	France wanted to push Britain out of the region, and Britain wanted to push France out.
British Mandate	Angry	الحكومة البريطانية يا سيدي… هي اللي غلت الحبوب علينا… بقى يجي واحد اسمه مخمن من طرف الدولة البريطانية… يجي على البلد، يجي على المختار.	The British government, sir… they are the ones who raised the price of grain on us… An assessor would come on behalf of the British state… come to the village, come to the *mukhtar*.

D. Appendix: Interview Question Analysis

Table 13: Distribution of interviewer question categories (250 distinct question forms). *Distinct* = unique question forms; *Mentions* = total occurrences across interviews. This categorisation was derived through manual annotation of an initial interview subset and has not been re-applied to the full 708-interview corpus; counts reflect the initial analysis.

Category	Distinct	Mentions
Backchannel / filler	70	319
Village geography & layout	44	89
Other	43	91
Daily life & economy	32	66
Name clarification	16	59
Identity & consent	12	32
Social customs & religion	11	22
Nakba & historical events	9	18
Education	7	14
Agriculture & land	6	12

Table 14: Thematic coverage of interviewer questions across the 708-interview corpus. Each row reports the number of question-bearing segments associated with a thematic tag out of 300,842 total question segments. Absolute counts for Nakba remain stable relative to the initial subset despite the corpus more than doubling; the drop in percentage reflects the large volume of YouTube-sourced interviews, which contain more question segments overall but proportionally fewer thematic flags than POHA archival interviews.

Theme	Question segments	%
Nakba	11,162	3.7
Jews / Israel	6,566	2.2
British Mandate	3,606	1.2

E. Appendix: Emotion and Lexical Data

Table 15: Top 12 content words by emotion category (stopwords, punctuation, dialectal prepositions, and discourse markers removed), based on 708 interviews. Segment counts in parentheses refer to non-neutral segments only. †شاء is a fragment of the formulaic expression *in shā' Allāh* (God willing).

Sad (33,581)				Happy (18,839)			
#	Word	Gloss	F	#	Word	Gloss	F
1	واحد	one/someone	2,988	1	واحد	one/someone	1,291
2	الناس	the people	2,270	2	كنا	we were	1,169
3	البلد	the town/village	2,071	3	أبو	father of (kunyā)	1,036
4	حدا	someone (dial.)	1,546	4	لي	to me / for me	1,002
5	لي	to me / for me	1,440	5	كنت	I was	780
6	يوم	day	1,423	6	يوم	day	777
7	هناك	there	1,388	7	الناس	the people	713
8	أبو	father of (kunyā)	1,386	8	نعم	yes	669
9	كلها	all of it	1,276	9	البلد	the town/village	668
10	إلا	except / only	1,275	10	عنا	with us	649
11	اليهود	the Jews	1,256	11	شاء	God wills†	617
12	طلعنا	we left	1,210	12	العريس	the groom	590

Angry (5,049)				Fear (3,202)			
#	Word	Gloss	F	#	Word	Gloss	F
1	قلت	I said	535	1	اليهود	the Jews	306
2	واحد	one/someone	505	2	قلت	I said	296
3	اليهود	the Jews	500	3	الناس	the people	272
4	لي	to me / for me	359	4	البلد	the town/village	270
5	إذا	if / when	308	5	واحد	one/someone	253
6	فلسطين	Palestine	293	6	علينا	upon us	208
7	أبو	father of (kunyā)	272	7	قالوا	they said	199
8	الناس	the people	264	8	إلا	except / only	185
9	إلا	except / only	242	9	لي	to me / for me	171
10	عمي	my uncle	231	10	حدا	someone (dial.)	165
11	البلد	the town/village	230	11	عمي	my uncle	132
12	العرب	the Arabs	219	12	كنا	we were	125

F. Appendix: Lexical Variation by Speaker Gender

Table 16: Top 15 content words by speaker gender after filtering, based on segment-level annotations across 708 interviews (male: 8.24M tokens; female: 942K tokens).

Male (8.24M tokens)				Female (942K tokens)			
#	Word	Gloss	Freq.	#	Word	Gloss	Freq.
1	أبو	father of (kunyā)	42,367	1	الله	God	4,813
2	البلد	the town/village	34,357	2	حدا	someone (dial.)	3,191
3	مثلاً	for example	32,146	3	هون	here (dial.)	2,823
4	الله	God	31,834	4	بدي	I want (dial.)	2,745
5	اسمه	his name is	26,689	5	البلد	the town/village	2,675
6	عمي	my uncle	26,056	6	أبو	father of (kunyā)	2,649
7	الناس	the people	24,493	7	صار	became / happened	2,499
8	القرية	the village	20,293	8	عم	progressive marker	2,293
9	إشي	thing (dial.)	18,923	9	عنا	with us / at ours	2,258
10	بيت	house	18,856	10	قالت	she said	2,216
11	دار	house / clan	18,495	11	كنا	we were	2,207
12	مين	who	18,298	12	اليهود	the Jews	2,201
13	اليهود	the Jews	16,891	13	مين	who	2,167
14	هذول	these (dial.)	16,374	14	إشي	thing (dial.)	2,163
15	طبعاً	of course	16,374	15	كمان	also / too	2,071

Table 17: Most significantly differentiated content words by speaker gender, ranked by log-likelihood ratio (G^2). All listed words are Bonferroni-significant ($\alpha = 1.103 \times 10^{-7}$; all $p < 10^{-10}$). Frequencies normalised per 1,000 tokens. †برضه is a dialectal adverb meaning "also/too" (Levantine/Egyptian variety), significantly more frequent in male speech.

	Male-favoured						Female-favoured				
#	Word	Gloss	M/1k	F/1k	G^2	#	Word	Gloss	M/1k	F/1k	G^2
1	القرية	the village	2.46	0.20	3104.8	1	يما	(O) mother	0.04	1.08	3170.7
2	هذي	this (dem.)	1.85	0.31	1725.9	2	كتير	a lot (dial.)	0.31	1.53	1900.8
3	أبو	father of (*kunyā*)	4.97	2.55	1245.3	3	قام	(s)he got up	0.15	1.07	1742.7
4	†برضه	also/too (dial.)	1.11	0.19	1040.7	4	إنتي	you (fem.)	0.10	0.89	1725.9
5	اسمه	his name is	3.20	1.55	908.3	5	بنت	girl/daughter	0.37	1.56	1676.2
6	عبد	(name part)	1.74	0.63	814.8	6	ويا	and with her	0.06	0.71	1604.7
7	كيلو	kilo/km	0.89	0.16	813.3	7	إيه	yes (dial.)	0.65	2.07	1564.3
8	موجود	present/existing	0.99	0.23	739.5	8	إمي	my mother	0.07	0.73	1552.3
9	حوالي	approximately	0.84	0.16	735.7	9	متل	like (dial.)	0.24	1.15	1386.2
10	قرية	village (var.)	0.92	0.20	728.3	10	جوزي	my husband	0.002	0.35	1354.5
11	منطقة	area/region	0.90	0.19	728.2	11	حرام	shame/forbidden	0.04	0.52	1170.2
12	شايف	(I) see/notice	0.73	0.12	680.3	12	طلعنا	we left	0.43	1.38	1056.4
13	متر	metre	0.62	0.08	665.3	13	بالبيت	in the house	0.06	0.52	1001.3
14	أما	as for	1.15	0.36	648.8	14	قامت	she got up	0.07	0.55	992.7
15	القرى	the villages	0.95	0.25	643.6	15	حبيبتي	my dear	0.004	0.28	989.5

G. Appendix: Transcription Prompt and Output Schema

This appendix provides the exact prompt and structured output schema used in the transcription pipeline. The prompt was designed to enforce high-fidelity transcription of Palestinian Arabic oral testimonies, with strict constraints on hallucination, dialect preservation, timestamp formatting, and structured metadata extraction.

G.1. Transcription Prompt

```
You are transcribing low-quality archival audio interview in Palestinian Arabic
↪  dialect (elderly speakers).

Output MUST be valid JSON matching the provided schema.

GLOBAL RULES (CRITICAL)
- Do NOT invent content. If unclear, write [inaudible]. Do NOT guess.
- Keep dialect as spoken (do not convert to MSA).
- Some tokens may be Palestinian place/village names (proper nouns). Do NOT correct
↪  them into common words.
  If you strongly suspect a token is a place name but youre not fully sure, keep the
  ↪  token and add: [PLACE?]
- Provide speaker diarization using SPEAKER_01, SPEAKER_02... and keep IDs
↪  consistent WITHIN THIS CHUNK.
- Provide segment timestamps as ABSOLUTE timestamps in the ORIGINAL FILE.

TIMESTAMP FORMAT (ABSOLUTELY REQUIRED)
- Every timestamp MUST be exactly: HH:MM:SS.mmm
- Always include HOURS, MINUTES, SECONDS, and MILLISECONDS.
- Use a DOT for milliseconds (.) not a comma.
- If you are unsure about milliseconds, still output .000 (never omit it).

SPEAKER METADATA (REQUIRED)
For each speaker in this chunk, fill:
- gender: male|female|unknown
- dominant_emotion: neutral|sad|angry|fear|happy|tired|unknown
- voice_signature: very short literal signature to help match across chunks (e.g.,
↪  "elderly female, soft voice, slow pace").

MENTION FIELDS (REQUIRED per chunk)
1) Nakba mention:
- Determine whether the Nakba is mentioned or referenced in this chunk.
- If mentioned: write nakba_literal_summary strictly literally (what is said,
↪  immediate context).
- If not: nakba_literal_summary must be empty string.

2) Jews/Israel mention:
- Detect whether Jews () or Israel () are mentioned or referenced in this chunk.
- If mentioned: write jews_israel_literal_summary strictly literally.
- If not: jews_israel_literal_summary must be empty string.

3) British Mandate mention:
- british_mandate_mentioned=true if explicitly:     / .
- If true: british_mandate_summary literal. If false: empty string.

SEGMENT-LEVEL FLAGS (OPTIONAL)
Inside each segment:
- set nakba_mentioned / jews_israel_mentioned / british_mandate_mentioned true if
↪  that segment contains the mention.
```

G.2. Output Schema (Simplified)

```
{
  "chunk_start": "HH:MM:SS.mmm",
  "chunk_end": "HH:MM:SS.mmm",

  "nakba_mentioned": true/false,
  "nakba_literal_summary": "...",

  "jews_israel_mentioned": true/false,
  "jews_israel_literal_summary": "...",

  "british_mandate_mentioned": true/false,
  "british_mandate_summary": "...",

  "speakers": [
    {
      "speaker": "SPEAKER_01",
      "gender": "male|female",
      "dominant_emotion": "...",
      "voice_signature": "..."
    }
  ],

  "segments": [
    {
      "start": "HH:MM:SS.mmm",
      "end": "HH:MM:SS.mmm",
      "speaker": "SPEAKER_01",
      "text": "...",
      "uncertain": true/false,
      "emotion": "...",
      "nakba_mentioned": true/false,
      "jews_israel_mentioned": true/false,
      "british_mandate_mentioned": true/false
    }
  ]
}
```

G.3. Output Schema (Simplified)

```
{
  "chunk_start": "HH:MM:SS.mmm",
  "chunk_end": "HH:MM:SS.mmm",

  "nakba_mentioned": true/false,
  "nakba_literal_summary": "...",

  "jews_israel_mentioned": true/false,
  "jews_israel_literal_summary": "...",

  "british_mandate_mentioned": true/false,
  "british_mandate_summary": "...",

  "speakers": [
    {
      "speaker": "SPEAKER_01",
      "gender": "male|female",
      "dominant_emotion": "...",
      "voice_signature": "..."
    }
  ],

  "segments": [
    {
      "start": "HH:MM:SS.mmm",
      "end": "HH:MM:SS.mmm",
      "speaker": "SPEAKER_01",
      "text": "...",
      "uncertain": true/false,
      "emotion": "...",
      "nakba_mentioned": true/false,
      "jews_israel_mentioned": true/false,
      "british_mandate_mentioned": true/false
    }
  ]
}
```