A Word Analysis for LahjaTube corpus

This appendix provides a detailed analysis of word frequency distributions in the LahjaTube corpus, highlighting the most common words unique to each dialect (EGY, GLF, LEV, MGR) and MSA, as well as the overlapping vocabulary between them.

DA	Count	MSA	Count	Overlap	DA Count	MSA Count
ده	2308	إلى	743	في	3250	3240
اللي	2259	أيضا	379	من	1806	2483
كده	993	ماذا	282	ما	1452	997
مش	851	الآن	274	يا	1039	934
بس	497	وهذا	272	هو	698	921
اي	434	كيف	256	يعني	561	361
عشان	324	هكذا	247	كل	534	525
عايز	318	عندما	243	كان	512	544
طب	311	حتى	216	عن	456	504
كمان	285	نحن	213	لا	284	1144

Table 7: Top words in EGY only, MSA only, and overlapped words between both EGY and MSA

Figure 3: The word cloud represents the most frequent words in EGY that do not appear in MSA

DA	Count	MSA	Count	Overlap	DA Count	MSA Count
اللي	1282	أن	1988	ما	2124	904
هذي	1215	إلى	1008	في	2079	2008
وش	808	حسنا	718	يا	1550	1463
راح	777	أو	473	من	1233	1639
بس	676	أريد	333	هذا	1137	1334
الحين	603	أنه	329	والله	976	745
اه	596	أنت	301	لا	930	1918
شي	548	ليس	301	له	833	614
عشان	426	أخي	297	الله	822	824
اي	394	إنه	278	الخير	656	534

Table 8: Top words in GLF only, MSA only, and overlapped words between both GLF and MSA

Figure 4: The word cloud represents the most frequent words in GLF that do not appear in MSA

DA	Count	MSA	Count	Overlap	DA Count	MSA Count
اللي	1115	أن	2131	ما	2713	842
هيك	1050	إذا	725	في	2185	1883
شو	1022	إلى	642	يعني	1337	767
بس	997	أو	575	من	1163	1790
هون	873	الآن	522	الله	941	911
راح	756	أيضا	519	يا	686	723
كمان	595	ماذا	459	هو	642	721
هلا	528	هكذا	444	هذا	635	973
عنا	507	أنا	405	هي	544	394
عم	500	التي	388	طبعا	502	244

Table 9: Top words in LEV only, MSA only, and overlapped words between both LEV and MSA

Figure 5: The word cloud represents the most frequent words in LEV that do not appear in MSA

DA	Count	MSA	Count	Overlap	DA Count	MSA Count
اللي	1247	أن	1524	ما	2135	650
ديال	920	إلى	769	في	1716	2437
غادي	580	أنا	480	من	1081	1616
شي	580	عندما	435	لي	1001	1097
باش	452	أي	362	واحد	828	84
صافي	420	تلك	322	على	631	726
ديالي	397	أو	309	يعني	623	293
بزاف	354	أنني	304	مع	549	530
بحال	342	شيئا	277	قلت	548	522
را	296	ليس	275	غير	535	68

Table 10: Top words in MGR only, MSA only, and overlapped words between both MGR and MSA

Figure 6: The word cloud represents the most frequent words in MGR that do not appear in MSA

B Annotation Guidelines for Human Evaluation

This appendix describes the guidelines followed by human annotators for evaluating DA-MSA translations in our study. Four annotators, each a native speaker of the four regional dialects (Egyptian, Gulf, Levantine, and Maghreb), evaluated samples corresponding to their dialect.

Annotation Procedure

For each sampled sentence, the annotator was presented with:

- The dialectal Arabic (DA) source transcript (from their dialect)
- Its automatically generated Modern Standard Arabic (MSA) translation

Annotators rated the translation on a scale from 1 to 5 for each of the following criteria:

1. **Accuracy:** How far does each translation accurately convey the meaning of the source text?
2. **Fluency:** How far is each translation easy to understand and follow?
3. **Style and Tone:** How far does each translation match the style and tone of the source text?
4. **Cultural Suitability:** How far is each translation culturally appropriate for the target audience?
5. **Terminology:** How far does each translation use specialized terminology accurately and consistently?

Each criterion was scored according to the following scale:

- 5 = **Completely** (meets the criterion perfectly)
- 4 = **Mostly**
- 3 = **Somewhat**
- 2 = **Slightly**
- 1 = **Not at all**