A phonemic comparison of dialects of Arabic

Brandon Rhodes

May 6, 2018

1 Introduction

The distinction between language and dialect is murky: it is often not clear how we decide that two languages stand in the language-dialect relation or the language-language relation. It is a question buried in politics, societal factors, and history — in addition to linguistic criteria. It is also not obvious what *would* be a good way to distinguish between the two. This paper will not provide (an) answer(s) to this question; however, this paper will take seriously the idea of using *some* quantitative measure to characterize differences between several dialects of Arabic, seeing what comes of it. In short, we will characterize differences between dialects by examining the differences in the distributions of phonemes in a crude way: we will be comparing percentages of occurrence for phonemes in the several dialects. The goal is to see what all this basic method can tell us about the (dis)similarity of the dialects, as it is not as clear as comparing, say, lexical differences. Additionally, we will look at a way of embedding dialects into a vector space using these percentages and then measuring the distances between dialects. However, before we get into these details, let us get a brief background on Arabic dialectology.

1.1 Brief background on recent approaches in Arabic dialectology

Arabic is one of the most widely spoken languages in the world, known for its many dialects. All Arabic-speaking countries have a major state of diglossia where Modern Standard Arabic

(MSA) is used in formal registers and the local dialect is the language acquired and the language used in every day interactions. There has been a recent increase in interest in the study of dialects of Arabic, primarily due to the demand coming from advances in technology. Arabic natural language processing is a field which has been especially interested in dialects because with the advent of social media content online is no longer only written in Modern Standard Arabic. Now, a lot of content is written in the several dialects of Arabic, and the problem with which most recent work has been concerned is identifying when a given text is written in dialect and, if possible, what dialect it comes from (Zaidan and Callison-Burch (2014); Meftouh et al. (2015); Biadsy et al. (2009); Biadsy and Hirschberg (2009); Mehrabani et al. (2010)). These approaches, as most within natural language processing, have had specific, practical and performance driven goals in mind and have not been concerned with a general characterization of the different dialects. For example, pitch distances are used to compared dialects of Arabic, but they are used specifically in the context of dialect identification in Mehrabani et al. (2010). One work that brushes up against this notion of generally characterizing dialects of Arabic is from Zaidan and Callison-Burch (2014); in fact, the present paper will be a direct extension of a concept they introduced: *dialectness factor*. We turn to this now.

1.2 Zaidan and Callison-Burch (2014)'s *dialectness factor*

In their paper, Zaidan and Callison-Burch introduced a measure called *dialectness factor*. This value is defined as the ratio of two percentages: the numerator is the percentage of occurrence for a phoneme or word in a given dialect and the denominator is the percentage of occurrence of that same phoneme or word in Modern Standard Arabic (MSA), as seen in (1). If this value is greater than 1, it indicates that the phoneme or word of interest occurs more in the given dialect than it does Modern Standard Arabic, and if this value is less than 1, it means the opposite. We can see a plot of the *dialectness factor* for the Jordanian dialect in 1. Zaidan and Callison-Burch (2014) did not further pursue this idea; they were using it

to show what phonemes and words *appeared* to be more dialectal, presumably with the hope of identifying text written in dialect more accurately.

(1) Dialectness factor: $\frac{\%\text{ of phoneme/word in dialect}}{\%ofphoneme/wordinMSA}$

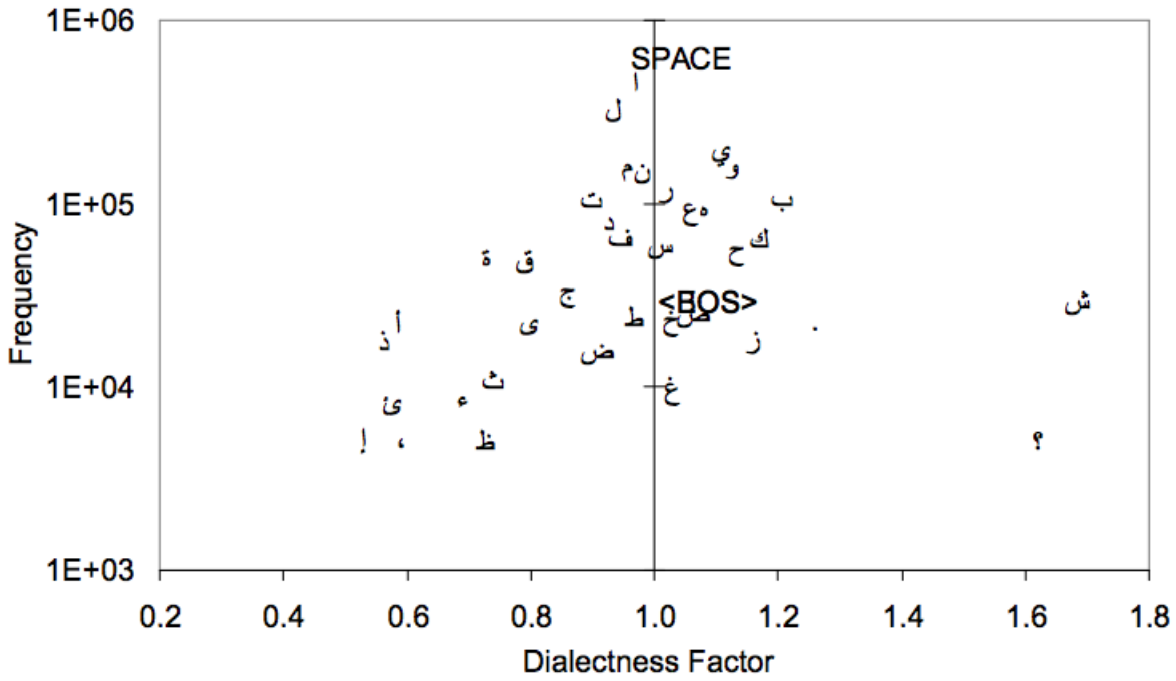

Figure 1: Zaidan and Callison-Burch (2014)'s plot of *dialectness factor* against token frequency for letters in Jordanian Arabic.

This notion of *dialectness factor* is interesting, though. It has a simplicity which is appealing: how far can we go with characterizing dialects based just upon the frequencies of their phonemes? It is straightforward to take a corpus and compute the various values of the *dialectness factor* of a phoneme, but when is it the case that this value indicates a significant difference between the dialects of Arabic and Modern Standard Arabic? Expanding upon and pushing this idea will constitute the bulk of this paper.

1.3 Brief outline of the paper

The paper will proceed as follows. First, we will introduce the methods we will use to make inferences on the differences of distributions in the phonemes in the several dialects of Arabic. This will entail introducing a basic linear model that we will fit to data acquired from a corpus translated into five dialects of Arabic alongside Modern Standard Arabic. In doing this, we will also elaborate on how we can think of this model in terms of *dialectness factor*. Then, we will look at the results of fitting this model to the data from the corpus. Last, we will use the data from the corpus to compute the *dialectness factor* of the given phonemes in the dialects of Arabic, allowing us to embed these dialects into a vector space where we can compute distances between them.

2 Methods and mathematical model

2.1 Background information on PADIC corpus

The Parallel Arabic Dialect Corpus (PADIC) was part of a larger Algerian national research project called 'TORJMAN', led by the Scientific and Technical Research Center for the Development of Arabic Language and funded by the Algerian Ministry of Higher Education and Scientific Research. The main goal for this project was to create dialectal resources for natural language processing tasks, and the PADIC corpus was crafted with a focus on machine translation.

The corpus has approximately 7000 sentences which are translated in parallel for the following dialects of Arabic: Algiers, Annaba, Moroccan, Palestinian and Syrian, along with Modern Standard Arabic. The first two dialects, Algiers and Annaba, are from costal cities in Algeria: Algiers, the capital, is on the central north coast, and Annaba is a city on the northeast coast. The data from Moroccan, Palestinian and Syrian dialects are intended to be representative of the respective countries and not of any particular city in the country.[1]

[1]The translations were obtained via Mechanical Turk (Meftouh et al. (2015)).

Sentences in this corpus come from various television shows, movies and recorded conversations in Algeria (Meftouh et al. (2015)). The corpus has approximately $40,000$ word tokens and around 9000 types for each dialect, with the actual numbers in figure 2. Each of the sentences was translated by a native speaker of one of the dialects, and all of the translations were done in Buckwalter format. Buckwalter format is a transliteration scheme used for representing the characters in Arabic script. Tim Buckwalter developed this ASCII only transliteration scheme which is unique for the reason that it was the first scheme that had a one-to-one mapping of Arabic characters to ASCII codes. We will use Buckwalter format in some of the plots in this paper (for ease of typesetting), so the differences in Buckwalter format from IPA are shown in (3).

(2) Dialects of PADIC:

(i)	Algiers	ALG
(ii)	Annaba	ANN
(iii)	Modern Standard Arabic	MSA
(iv)	Moroccan	MOR
(v)	Palestinian	PAL
(vi)	Syrian	SYR

(3) IPA / Buckwalter differences

IPA		Buckwalter
t^ς	$\mapsto$	T
d^ς	$\mapsto$	D
s^ς	$\mapsto$	S
ʔ	$\mapsto$	E
ħ	$\mapsto$	H
ɣ	$\mapsto$	g
ʃ	$\mapsto$	$

Corpus	#Types	#Tokens
ALG	8966	38707
ANB	9060	38428
TUN	10215	36648
SYR	9825	37259
PAL	9196	39286
MSA	9131	40906

Figure 2: PADIC statistics from Meftouh et al. (2015)

2.2 Data collection

The letter frequencies were obtained in batches of one hundred sentences. The decision to use batches of this size centered around reducing the probability that we would observe no counts for a letter in a batch of sentences from the corpus. To achieve this, a 'worst-case' scenario strategy was used, working in the following way: (i) a program obtained the counts for each letter in the corpus; (ii) we chose from the letters of interest the letter y which exhibited the least amount of counts; (iii) we estimated the probability that this letter would appear in a sentence; (iv) we chose a batch size of n such that $\mathbb{P}(y \text{ does not appear in sentence}) < 0.005$.[2]

(i) Obtain counts for all letters y in PADIC

(ii) Choose letter y' which has fewest counts $\quad \text{count}(y') \approx 400$

(iii) Estimate probability that this letter y' appears in a sentence

$$\begin{aligned}\mathbb{P}(y' \text{ appears in sentence}) &= \frac{\text{count}(y')}{\#\text{ sentences}} \\ &\approx \frac{400}{7200} = 0.056\end{aligned}$$

[2]This assumes that the probability that y appears in one sentence is independent (and equal) to appearing in another sentence. It is straightforward to see that the probability any letter y is not equal with regard to the sentence (e.g. probability depends on length of sentence), but I think the more important assumption of independence does hold to a greater extent.

(iv) Choose n such that $\mathbb{P}(y' \text{ does not appear in sentence}) < 0.005$

$$(\mathbb{P}(y' \text{ does not appear in sentence}))^n = (1 - \mathbb{P}(y' \text{ appears in sentence}))^n < 0.005$$
$$\left(1 - \frac{400}{7200}\right)^n < 0.005$$
$$n \approx \frac{\log(0.005)}{\log(1 - \frac{400}{7200})}$$
$$\approx 93$$

2.3 Mathematical model

2.3.1 Model and underlying assumptions

We want to answer this question: Is the distribution of phoneme y significantly different in dialect i than in MSA? Mathematically, this is asking the question of whether or not there is some constant $\alpha_i \neq 0$ for each dialect i such that the percentage of phoneme y in that dialect is the 'true' percentage of y (which we will estimate by the percentage in MSA) plus some constant α_i, as in (4).[3] In terms of Zaidan and Callison-Burch (2014), we can think of the question as whether or not the *dialectness factor* of phoneme y is 'far enough' away from 1, shown in (5).

(4) Is there some constant $\alpha_i \neq 0$ such that
% of y in dialect i = (% of y in MSA) + α_i?

(5) Is $\frac{\text{count}_i(y)/\text{count}_i(total)}{\text{count}_{MSA}(y)/\text{count}_{MSA}(total)} \neq 1$

The linear model we will use to answer this question is a one-way layout: it has one categorical factor for dialect of Arabic at $I = 6$ levels with $J = 72$ replicates. This means for each dialect, we have a sample of size 72. We will make the standard assumptions for a linear model, which is to assume that for each observed percentage Y_{ij} there is some random error e_{ij}

[3]We say 'true' in quotations because the decision of what dialect we use estimate it is arbitrary; in other words, we could also estimate the 'true' percentage with any of the other dialects, and the results from our model would not change with this choice. Our interpretation, however, would have to be adjusted.

with the following properties, listed in (6): (i) each e_{ij} is independent of all other $e_{i'j'}$ for $i' \neq i$ and $j' \neq j$; (ii) each e_{ij} has mean zero; (iii) all the errors e_{ij} have equal variance; and (iv) the e_{ij} follow a normal distribution. The model just explained is seen in (7): the observed percentage Y_{ij} of a phoneme in dialect i from example j will be written as the 'true' percentage μ of phoneme y plus some dialectal constant α_i and multiplied by a random error e_{ij}. Using this model, we want to test whether $\alpha_i = 0$ for all i.

(6) Assumptions on random errors:

(i) All e_{ij} are independent of each other

(ii) $\mathbb{E}[e_{ij}] = 0$

(iii) $Var(e_{ij}) = \sigma^2$ for all e_{ij}

(iv) $e_{ij} \sim N(0, \sigma^2)$

(7) Model:

$$\underbrace{Y_{ij}}_{\text{observed \% phoneme in dialect } i \text{ for example } j} = \underbrace{\mu}_{\text{'true' \%}} + \underbrace{\alpha_i}_{\text{constant for dialect } i} + \underbrace{e_{ij}}_{\text{random error for observed \%}}$$

(8) Hypotheses to be tested:

H_0 : all $\alpha_i = 0$

H_1 : at least one $\alpha_i \neq 0$

2.3.2 Parameterization and its interpretation

The model as we have it now is over-parameterized.[4] A common way to restrict the model to give us unique estimates is to assume that one of the parameters α_i is equal to zero, and we will use this parameterization. The dialectal constant α_i we will make equal to zero is the constant for Modern Standard Arabic α_{MSA}. Using the maximum likelihood method of estimation, we have the following estimates in (10) for the parameters $\theta =$

[4]Formally speaking, this means that the design matrix X is not full rank. An intuitive way of seeing this is that we can write $Y_{ij} = (\mu+c)+(\alpha_i-c)+e_{ij}$ for any c; so, for the space of parameters $(\mu, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6)$, if $(1, 0, 0, 2, 0, -2, 1)$ is a solution, then so is $(2, -1, -1, 1, -1, -3, 0)$.

$(\mu, \alpha_{MSA}, \alpha_{ALG}, \alpha_{ANN}, \alpha_{MOR}, \alpha_{PAL}, \alpha_{SYR})$, where we will index the α_i as $(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6)$, seen in (9).[5,6] Note that $\overline{Y}_{i.} = \frac{1}{J}\sum_{j=1}^{J} Y_{ij}$: it is the average of the observations for dialect i.

(9) Indexing:

$1 \mapsto$ MSA	$3 \mapsto$ ANN	$5 \mapsto$ PAL
$2 \mapsto$ ALG	$4 \mapsto$ MOR	$6 \mapsto$ SYR

(10) M.L.E. estimates given parameterization $\alpha_1 = 0$

$\hat{\mu} = \overline{Y}_{1.}$	$\hat{\alpha}_3 = \overline{Y}_{3.} - \overline{Y}_{1.}$	$\hat{\alpha}_5 = \overline{Y}_{5.} - \overline{Y}_{1.}$
$\hat{\alpha}_2 = \overline{Y}_{2.} - \overline{Y}_{1.}$	$\hat{\alpha}_4 = \overline{Y}_{4.} - \overline{Y}_{1.}$	$\hat{\alpha}_6 = \overline{Y}_{6.} - \overline{Y}_{1.}$

Choosing $\alpha_1 = 0$ to correspond to Modern Standard Arabic, we have the estimate for the 'true' value μ to be the average percentage in Modern Standard Arabic; consequently, the estimates of the parameters for each dialect is in terms of its variation from Modern Standard Arabic. For example, the estimate for the constant α_2 for the Algiers dialect is the difference of the sample mean for the observations in Modern Standard Arabic from the sample mean from the Algiers dialect, as seen in (11). If this estimate is positive, it implies that the phoneme of interest has a higher occurrence percentage in the Algiers dialect than in Modern Standard Arabic; if estimate is negative, it implies the converse. The same interpretation holds for all dialects; furthermore, using these estimates, we can get a sense of the phoneme's *dialectness factor* in the dialect: if the estimate is positive, it implies the *dialectness factor* is greater than 1.

(11) $\hat{\alpha}_2 = \overline{Y}_{2.} - \overline{Y}_{1.} = \frac{1}{72}\sum_{j=1}^{72} Y_{2j} - \frac{1}{72}\sum_{j=1}^{72} Y_{1j}$

[5]Given the parameterization, this means $\alpha_1 = 0$.

[6]We denote the estimate of a parameter θ as $\hat{\theta}$.

(12) Interpretation:

$$\hat{\alpha}_i < 0 \Rightarrow \ \textit{dialectness factor} < 1$$

$$\hat{\alpha}_i > 0 \Rightarrow \ \textit{dialectness factor} > 1$$

$$\hat{\alpha}_i = 0 \Rightarrow \ \textit{dialectness factor} = 1$$

2.3.3 Justification of assumptions on errors

Before we begin our summary of the data, let us ensure that the assumptions we are making on the random variation e_{ij} are plausible. We will briefly address each assumption in (6), explaining, and demonstrating when possible, why each one is justified.

The independence assumption of the error structure is the most important; however, we have the least amount of evidence for this assumption, so it is not innocuous. Deviation from the truth of this assumption would most likely come from the tendencies of the native speakers who translated the sentences into their dialect; for example, suppose one speaker of the Algiers dialect, call him/her speaker 1, has a phonotactic preference to avoid certain consonant clusters while another speaker, speaker 2, of this dialect does not. The random errors for speaker 1 would presumably be correlated in some way. Since PADIC does not indicate which speaker translated which sentence, we can not model this correlation; however, any implausibility of the independence assumption is likely to be drastically reduced given our method of recording the percentages of a phoneme per one hundred sentences. Having an individual point in the sample as a percentage from one hundred sentences allows for that percentage to be composite of multiple speakers' translations, and this makes the independence assumption tenable. To help demonstrate this, consider the residuals for /b/ in each dialect, seen in figure 3. If there were dependence of the kind just described, we would expect to see some pattern in the sample points within each dialect; however, there is no obvious patterning among the points in the sample for any of the dialects. Another source of deviation from independence could come from PADIC being unbalanced, biased in

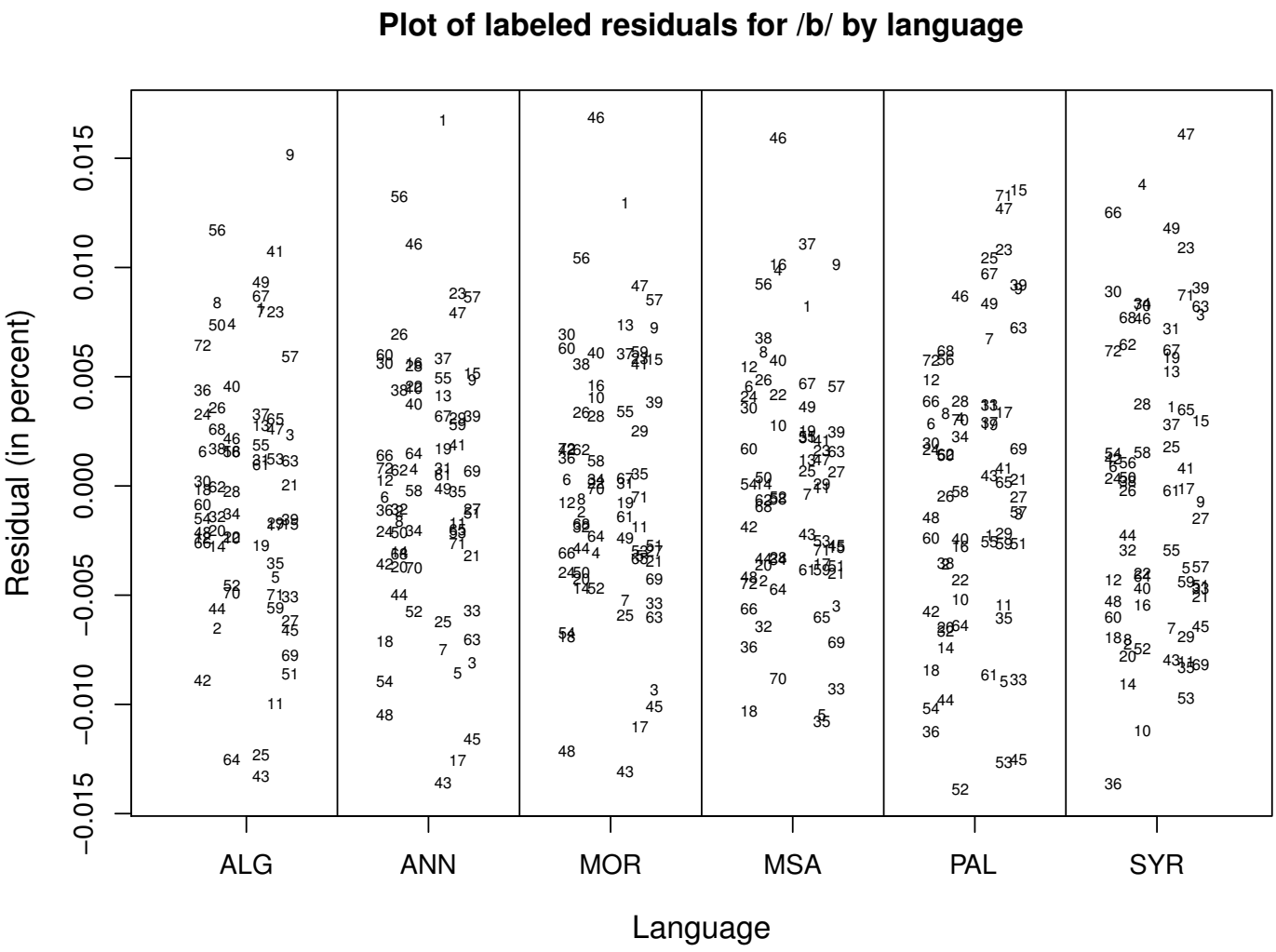

Figure 3: Plot of residuals for each dialect. There are no obvious trends in the data within each dialect.

some way. Based on the construction of PADIC, which is outlined in Meftouh et al. (2015), this does not appear to be a problem.

Next, we can check the last three assumptions in (6) by looking at a few more residual plots, seen in figure 4. First, we can see that for each of the phonemes, we have a distribution which is centered at zero; second, the variance ('spread') of the residuals for each phoneme is approximately equal for all of the dialects; last, each of these residual plots resemble

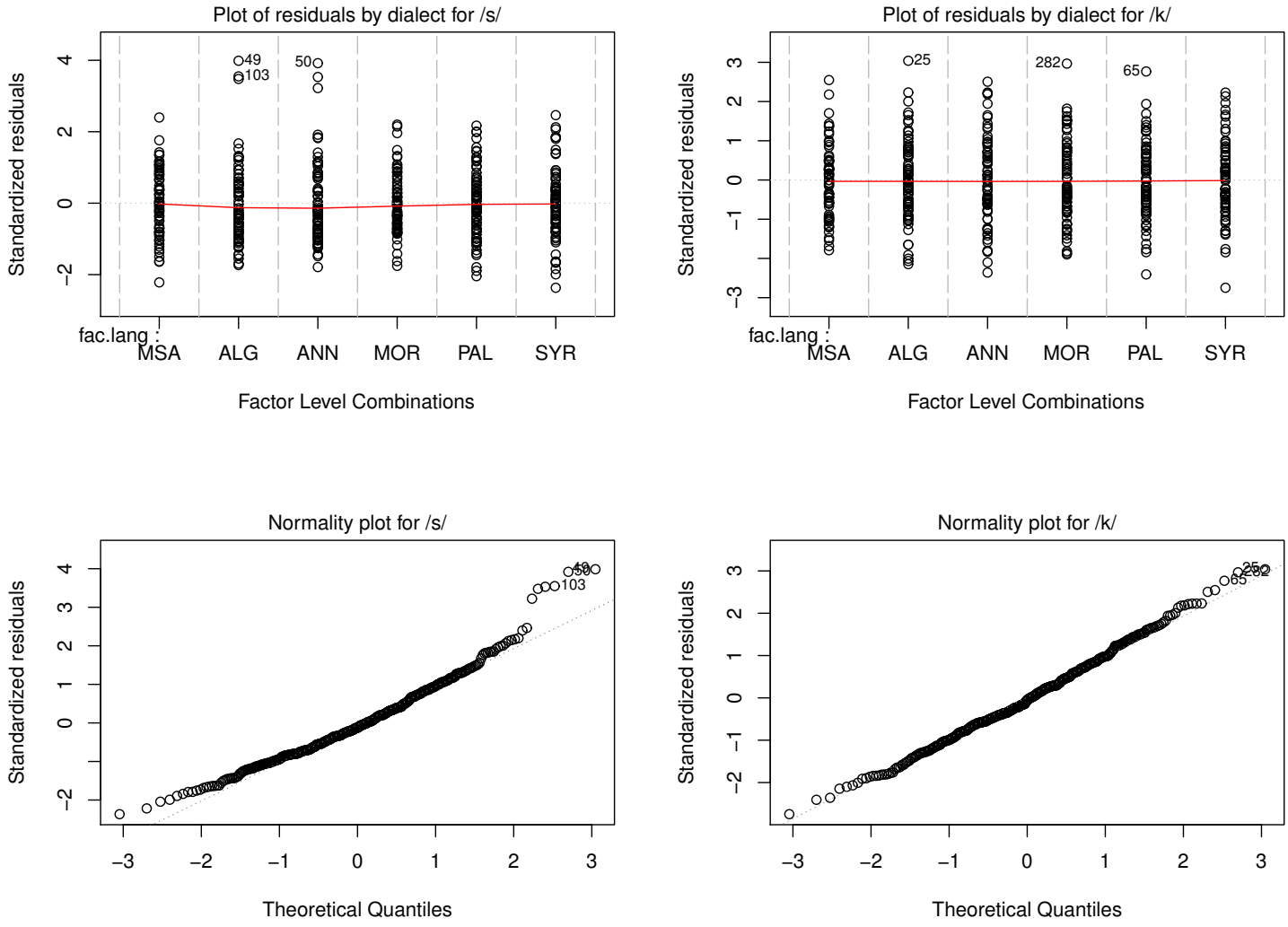

Figure 4: Residual and normality plots demonstrating the plausibility of the last three assumptions of the model.

a normal distribution. These plots are symmetric about their means and have most of the density (shown by darkness in the plot caused by overlapping points) about this mean. Another way to examine the normality assumption is to consider the normality plots in figure 4 (the two plots on the bottom). The plots are close to linear in the theoretical quantiles of a normal distribution, indicating that the normality assumption is plausible. The plots for the selected phonemes here are representative for those not shown.

3 Results and discussion

This section will present the results we have from the PADIC corpus. We will first look at some basic statistics from the corpus; then, we will discuss how we use these statistics when making inferences within the linear model we presented above, focusing on the comparison of the dialects to Modern Standard Arabic using simultaneous t-tests; last, we will present the results.

3.1 Basic statistics and results

The table below in figure 5 gives the sample means (in percent) of the phonemes for each dialect. We've restricted our study to only a subset of the consonant phonemes; this was merely a way to limit the current study, and the methods could be extended to vowels, diacritics and punctuation.[7] The second column (from the left) in the table is the average percent of the given phonemes for Modern Standard Arabic, and this will be the column on which we will base our comparisons. More explicitly, we will be testing whether the difference in percent that each dialect has compared to MSA for a given phoneme is statistically significant. Before we look into the mechanics of how we will use these values, let us look at a few different plots of the results to get a sense of the data.

[7]However, we should be more careful with our language in that case and consider everything a letter — not a phoneme. The letters chosen in this study correspond to those which faithfully represent the underlying phoneme.

Sample mean for each phoneme

letter	MSA	ALG	ANN	MOR	PAL	SYR
m	0.062	0.0521	0.0532	0.0546	0.0601	0.0643
n	0.0652	0.052	0.052	0.0584	0.057	0.0587
l	0.106	0.0925	0.0871	0.094	0.0946	0.096
r	0.0357	0.0477	0.0454	0.0383	0.0385	0.0426
b	0.0311	0.035	0.0345	0.036	0.0434	0.0519
t	0.0495	0.0463	0.0455	0.056	0.0522	0.0548
tˤ	0.0066	0.007	0.0069	0.0076	0.0084	0.0083
d	0.0273	0.0336	0.0332	0.0412	0.0237	0.0257
dˤ	0.0056	0.0021	0.0025	0.0042	0.0043	0.0048
k	0.03	0.0359	0.0367	0.0334	0.0319	0.0343
q	0.0205	0.02	0.0199	0.02	0.0202	0.0167
f	0.0221	0.02	0.0199	0.0246	0.0237	0.0194
s	0.0248	0.0205	0.0206	0.0188	0.0227	0.0238
z	0.0043	0.0066	0.0064	0.0069	0.0065	0.0052
sˤ	0.0101	0.0073	0.0071	0.0098	0.0104	0.0111
ʃ	0.0094	0.0224	0.0203	0.0309	0.0253	0.021
x	0.0085	0.0088	0.0094	0.011	0.0085	0.0093
ħ	0.0189	0.0248	0.025	0.02	0.0241	0.0252
ʔ	0.0339	0.0298	0.0306	0.0296	0.0343	0.0371
ɣ	0.003	0.003	0.003	0.0091	0.0035	0.0037
d͡ʒ	0.0128	0.0121	0.0112	0.0129	0.0107	0.0113

Figure 5: Dialect sample means (grouped by manner of articulation).

First, we can consider plots which incorporate information about each phoneme's *dialectness factor* in each dialect. The plots below will plot the phoneme's mean log(*dialectness factor*) against its mean count in the sample. In other words, we take the entry in the table above a phoneme in a given dialect, we divide this value by the entry for MSA, and we then take the logarithm, as seen in (13). We look at the logarithm of the *dialectness factor* because it will make the scale on the horizontal axis linear, as the *dialectness factor* is a multiplicative

relationship; additionally, it provides a straightforward interpretation of a phoneme occurring less or more frequently by negative and positive values respectively. A vertical line is drawn at zero to indicate MSA's mean *dialectness factor*, which is 1 since $\frac{\%y \text{ in MSA}}{\%y \text{ in MSA}} = 1$ and $\log(1) = 0$. For phonemes which appear to the right of this line, it indicates the phoneme appears more in the dialect, and vice versa for those phonemes which occur to the left.

(13) $\log\left(\frac{\text{table entry for } y \text{ in dialect}}{\text{table entry for } y \text{ in MSA}}\right) = \log(\text{estimated } \textit{dialectness factor})$ POINT IN PLOT

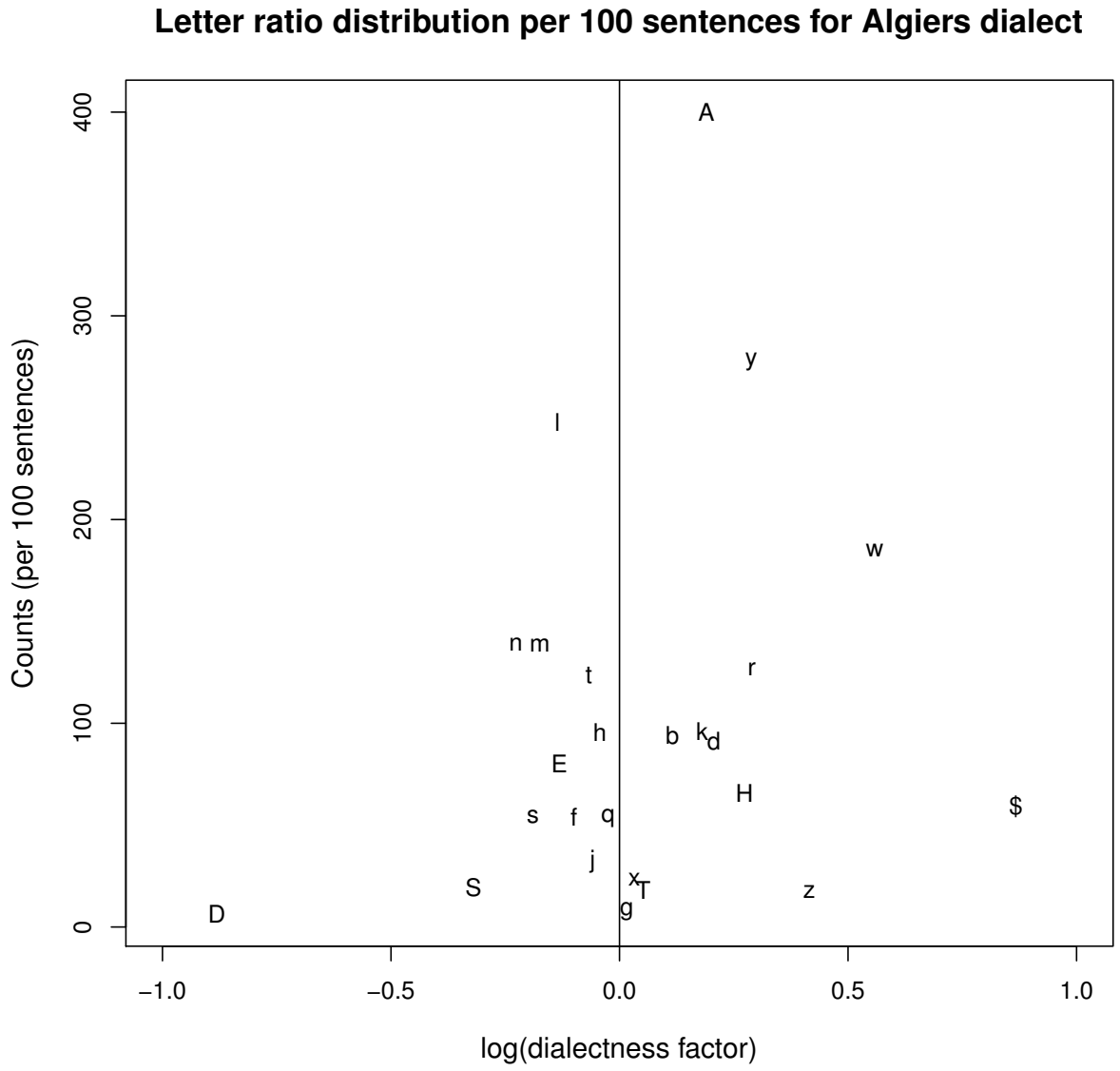

Figure 6: Letter plot for Algiers dialect

3.1.1 Algiers dialect

Most of the phonemes occur within -0.5 and 0.5 which corresponds to a *dialectness factor* of approximately 0.6 and 1.6 respectively; however, we do see that D and \$, [d^{ς}] and [ʃ] in IPA, seem to be markedly farther.

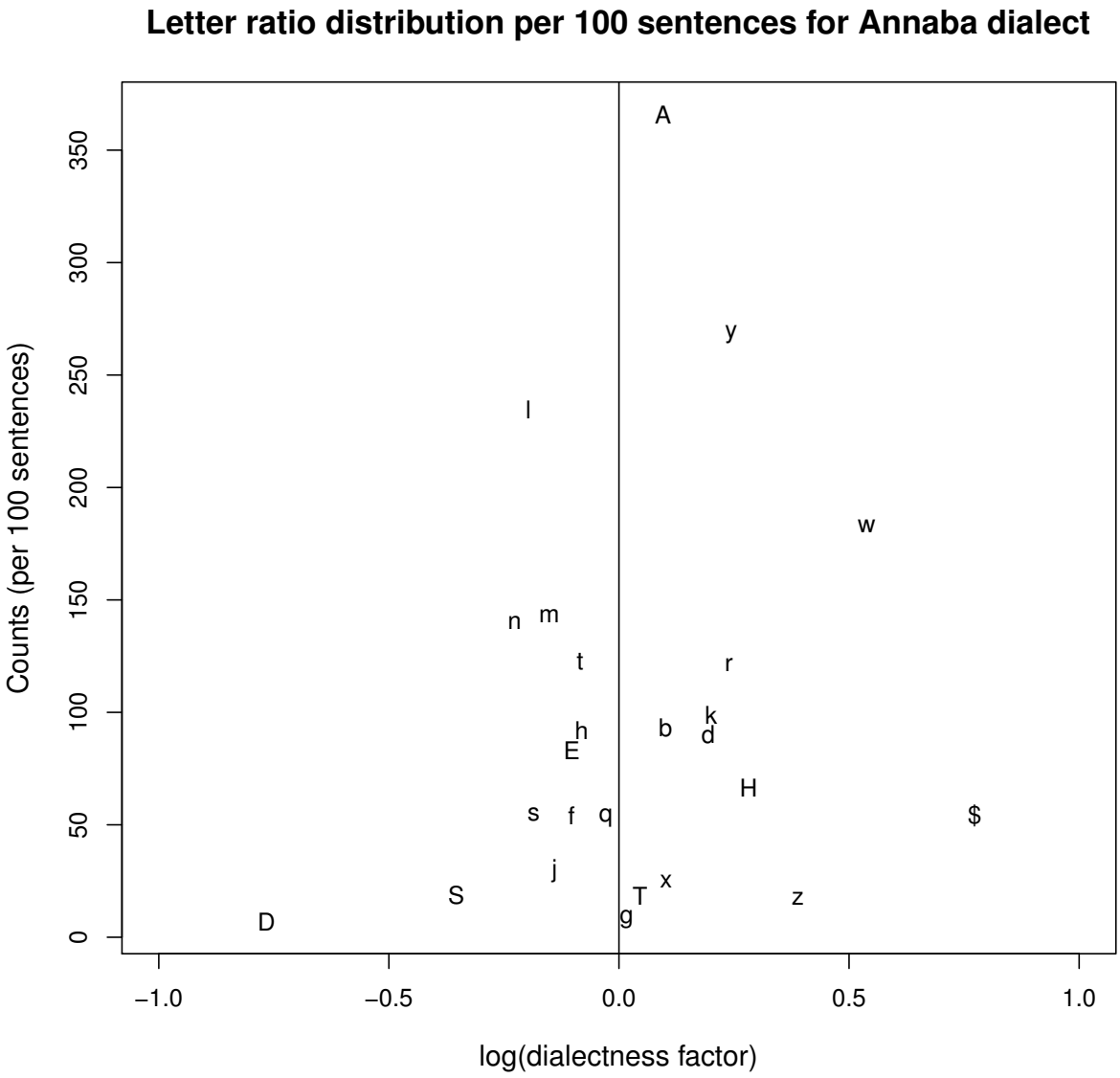

Figure 7: Letter plot for Annaba dialect

3.1.2 Annaba dialect

We see similar behavior from the phonemes as in the previous letter plot (for the Algiers dialect). Both /d$^{\varsigma}$/ and /ʃ/ (D and $) are substantially farther than the rest of phonemes, with the most of the phonemes having a *dialectness factor* in the interval $(0.6, 1.6)$.

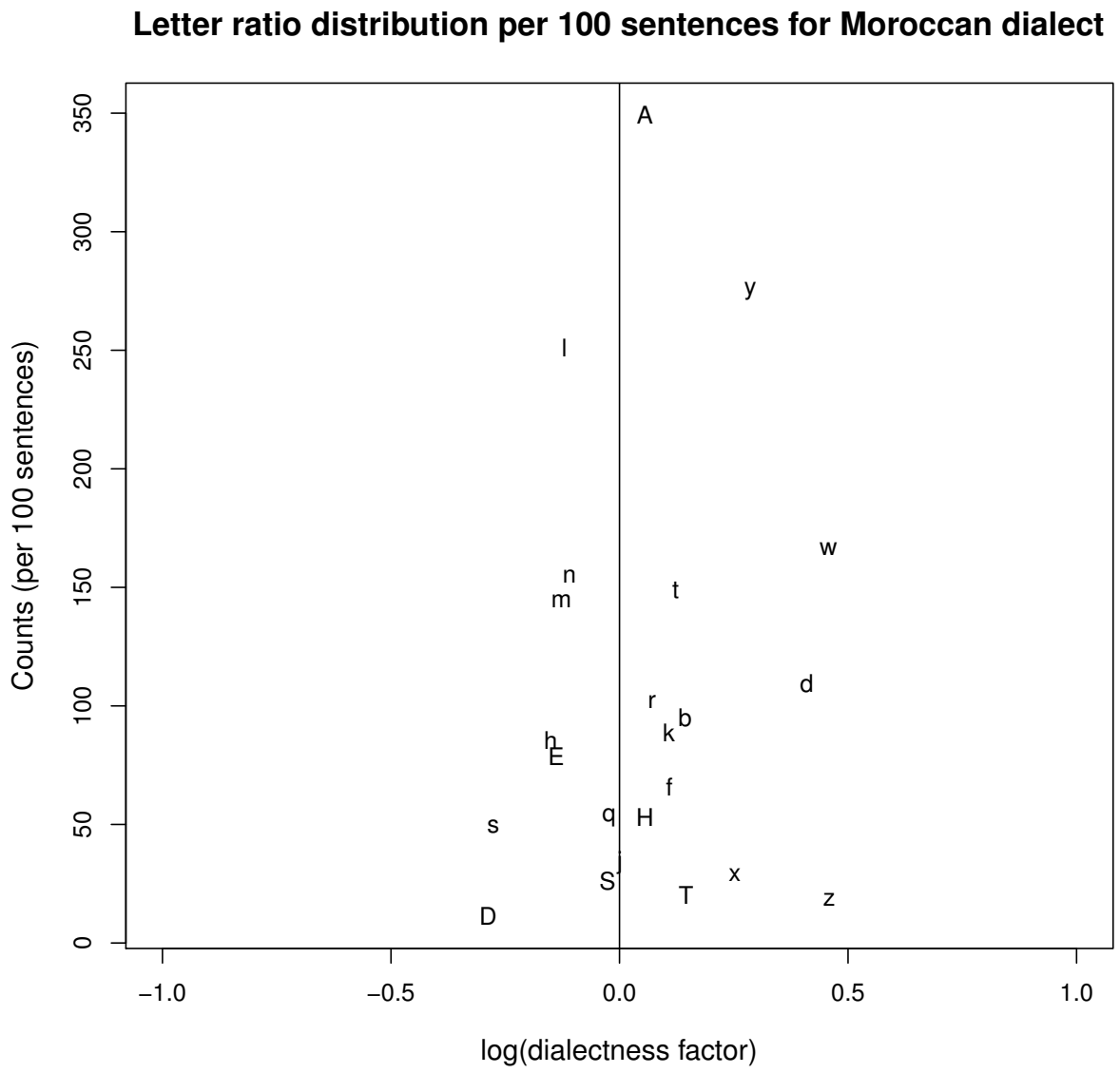

Figure 8: Letter plot for Moroccan dialect (/ʃ/ not shown because of very high *dialectness factor*)

3.1.3 Moroccan dialect

Here we have a slightly different picture: just about all of the phonemes lie within the same interval as the previous two dialects. The palato-alveolar fricative /ʃ/ has such a high *dialectness factor* that we omitted it from the plot to keep the scale the same.

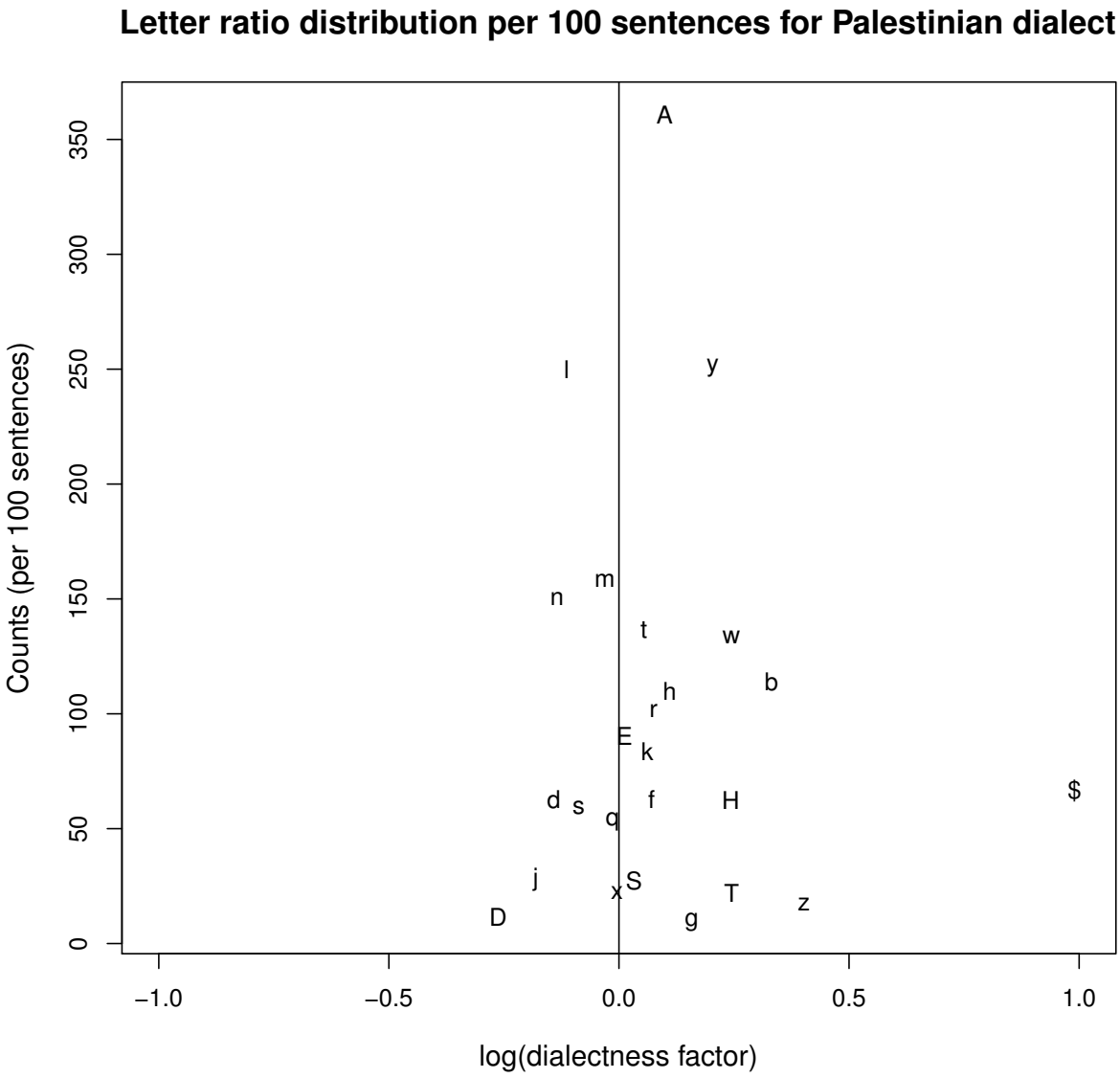

Figure 9: Letter plot for Palestinian dialect

3.1.4 Palestinian dialect

In the Palestinian dialect, we see that the phonemes are concentrated again in the interval of $(0.6, 1.6)$ for *dialectness factor*, but there does seem to be a slight tendency for phonemes to occur more in the dialect than in Modern Standard Arabic — i.e. if we were to imagine this having a center of mass, it appears to be slightly off center to the right.

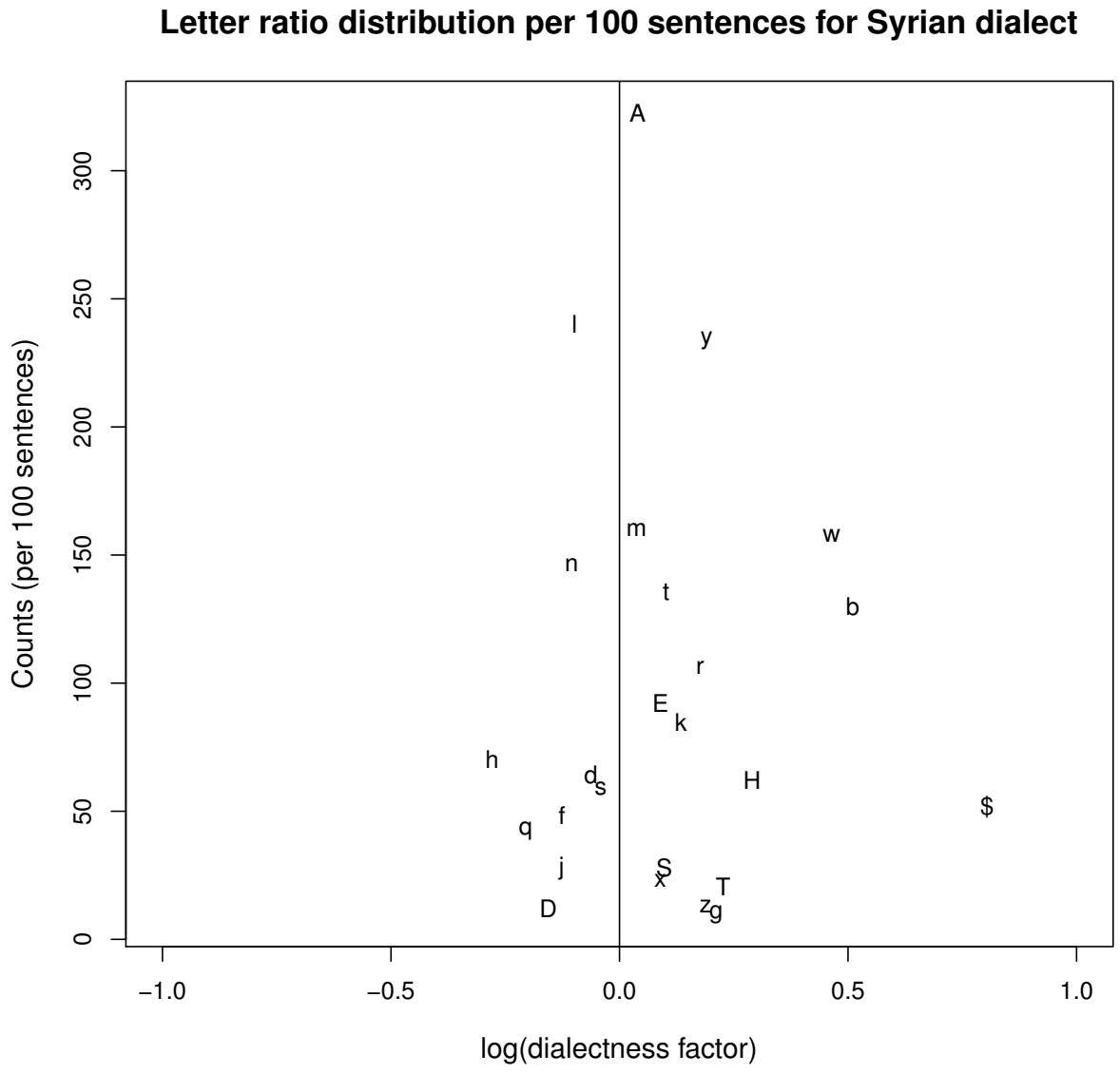

Figure 10: Letter plot for Syrian dialect

3.1.5 Syrian dialect

This distribution seems to be more tightly concentrated in the center than the other dialects. We do, however, see the phoneme /b/ having a substantially higher dialect factor here. Also, the phoneme /d$^{\varsigma}$/ is the closest to Modern Standard Arabic here than in the other dialects.

3.1.6 General remarks on plots

We see the same basic shape for all of the plots: the phonemes with the higher counts appear to be closer to Modern Standard Arabic on average while phonemes which are less frequently occurring seem to have a wider spread. The phoneme /ʃ/ uniformly has a *dialectness factor* greater than 1, while phonemes like the nasals /n/ and /m/ are almost always to the left of the vertical line, indicating that they occurred more in Modern Standard Arabic in the corpus. The Algiers and Annaba dialect appear to have the widest spread on average, and they appear to pattern the most similar way with respect to Modern Standard Arabic, which seems plausible given that they are dialects geographically close to each other. We will soon be more precise about summarizing the data, but we will first get an intuition about the hypotheses we are testing in the data.

3.2 Hypothesis testing

Recall that we want to answer the question: Is the distribution of phoneme y significantly different in dialect i than in MSA? In section 2.3, we talked about this in terms of the α_i and whether or not these $\alpha_i = 0$ or not. This is how we will do it formally, but it may not have most intuitive interpretation, especially with respect to *dialectness factor*, so we will use some plots similar to the ones above to illustrate the different hypotheses in terms of *dialectness factor* that we will see tested in the upcoming sections. Recall first that hypotheses to be tested are the following:

(14) Hypotheses to be tested:

H_0 : all $\alpha_i = 0$

H_1 : at least one $\alpha_i \neq 0$

Note that we can also break the first (null) hypothesis into several parts (recalling that we must assume $\alpha_1 = 0$), as in (15). The hypothesis above is just asking whether or not there is *a* dialect different than Modern Standard Arabic, but, breaking the null hypothesis up, we can ask *which* dialects are significantly different than Modern Standard Arabic, a much more interesting question.

(15) Hypotheses to be tested:

$$H_0 : \text{all } \alpha_i = 0 \quad \rightsquigarrow \quad \begin{aligned} &\alpha_2 = 0 \\ &\alpha_3 = 0 \\ &\dots \\ &\alpha_6 = 0 \end{aligned}$$

In terms of visualizing the hypotheses, we can consider the distribution of our favorite phoneme /b/. The first hypothesis of testing whether there is at least one dialect that is different than Modern Standard Arabic is asking whether we can draw a line such that on one side of the line we have the dialect(s) where the phoneme has a significantly different *dialectness factor* and on the other side the dialect(s) where there is no such significant difference. In figure 11, we can see that one answer we may get is the dashed line, which would imply that the distribution of /b/ in the Palestinian and Syrian dialects is significantly different; the dotted line would imply this only holds for the Syrian dialect; or, there could be no such line, implying none of them are significantly different than Modern Standard Arabic.

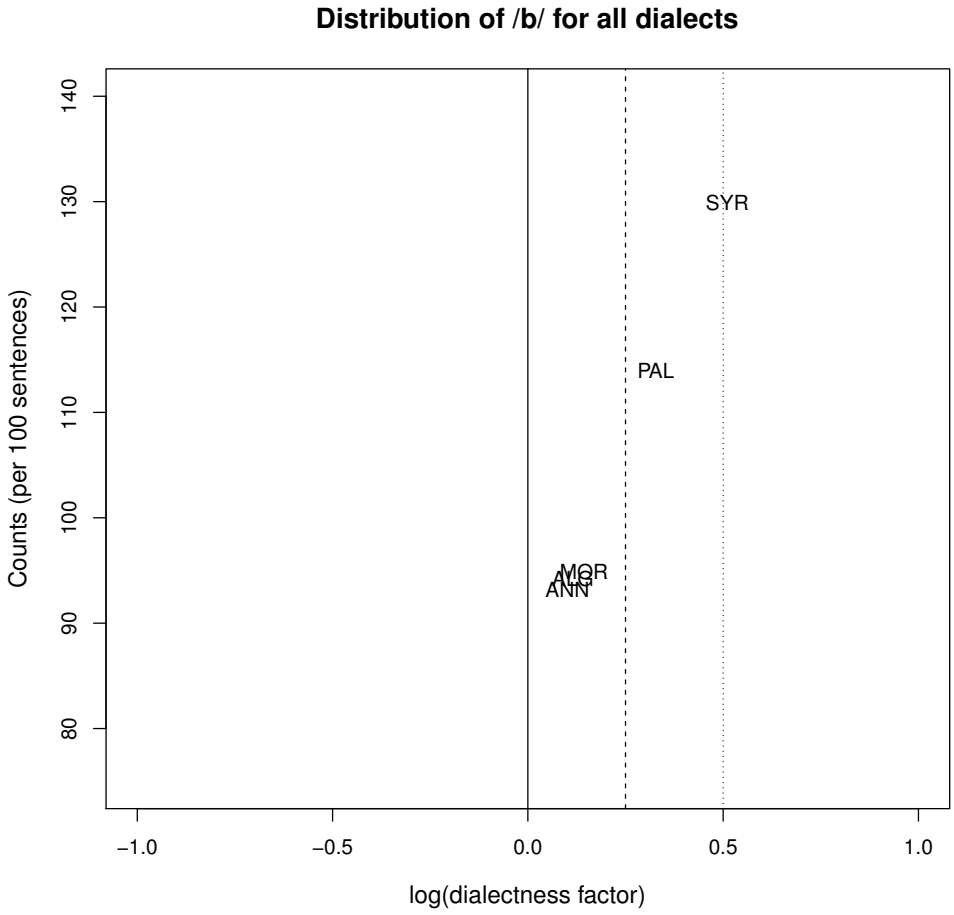

Figure 11: Sample means and possible boundaries for a significant *dialectness factor*.

When we look at the hypothesis with respect to the individual hypotheses which make it up, we can think of it similarly in terms of the same type of plot, but restricted to a specific dialect. The interpretation we have here is that a phoneme's distribution in the given dialect is significantly different than its distribution in Modern Standard Arabic if its center of mass (sample mean) is far enough away from that horizontal line at 0. Take, for example, the distribution of /b/ in the Algiers and Syrian dialects, seen in (12). It is pretty clear the distribution of /b/ in the Syrian dialect is not centered near Modern Standard Arabic (the vertical line); on the other hand, it is not clear that the distribution of /b/ in the Algiers dialect is different than its distribution in Modern Standard Arabic. In general, we do not know the true *dialectness factor* of a phoneme, but using our data we can construct

an interval that we have statistical confidence in saying the true *dialectness factor* lies in this interval. If this interval for the phoneme y in a dialect contains 0, then we will say the distribution phoneme y is not significantly different in the dialect, and if it does not contain 0, we will say it is. For our /b/ example, we could have an interval (dotted lines) for the Algiers dialect which has the vertical line inside of it and an interval for the Syrian which does not, indicating that the distribution of /b/ is significantly different in the Syrian dialect.

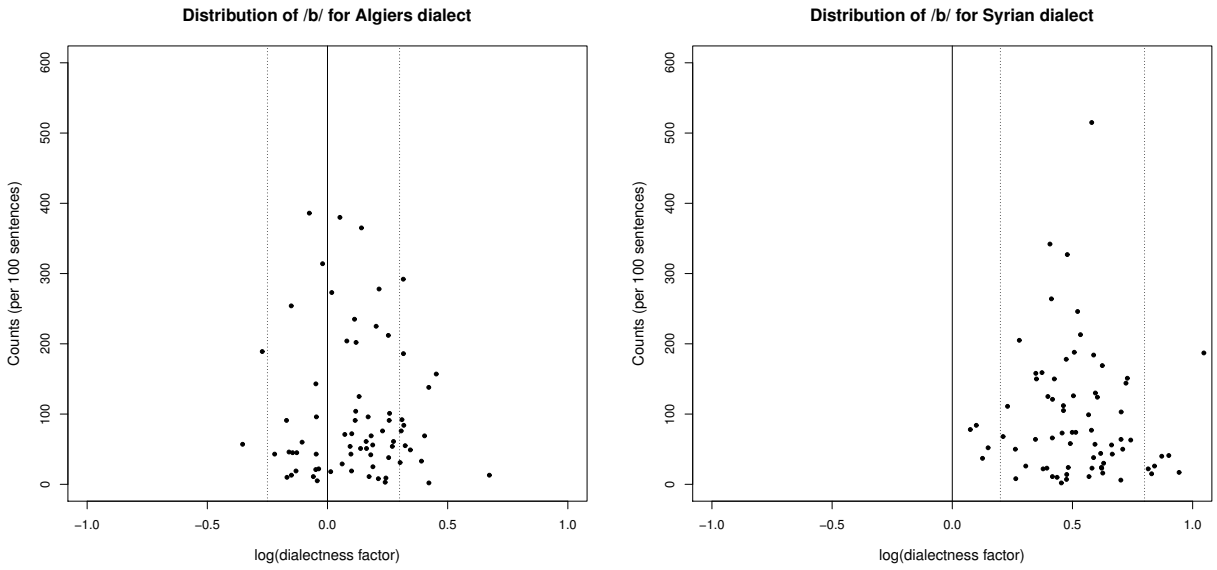

Figure 12: Distribution of /b/ in Algiers and Syrian dialect with possible *dialectness factor* intervals (dotted lines)

3.3 Results

Just above we explained the hypothesis testing in terms of *dialectness factor*: here are the results at a significance level of .05.[8] Red indicates that the phoneme occurs significantly less in the dialect, green indicates that it occurs significantly more and a blank cell indicates there is no significance difference in the distributions of phoneme in the dialect and its distribution in Modern Standard Arabic.

[8] All p-values and intervals have been Bonferroni corrected.

Figure 13: Significant differences (at .05 level) in the distributions of phonemes with respect to Modern Standard Arabic: red indicates less frequently occurring, green indicates more frequently occurring and a blank cell indicates no significant difference.

We see that there are many significant phonemic differences in each dialect with respect to Modern Standard Arabic. Immediately, we see that the Algiers and Annaba dialects differ from Modern Standard Arabic in a near identical way; however, outside of this, there is not very obvious structure to the differences we see. Phonemes which occur significantly more in each dialect are /b/ and /ʃ/ while /l/ and /n/ occur less. It is known that dialects of Arabic do not exhibit grammatical case, and some case suffixes end with /n/, so this may be a reflection of this fact (Zaidan and Callison-Burch (2014)). In addition to comparisons to Modern Standard Arabic, we can make the same comparisons with each pair of dialects.

The results of doing these comparisons are shown in figures 14 and 15.

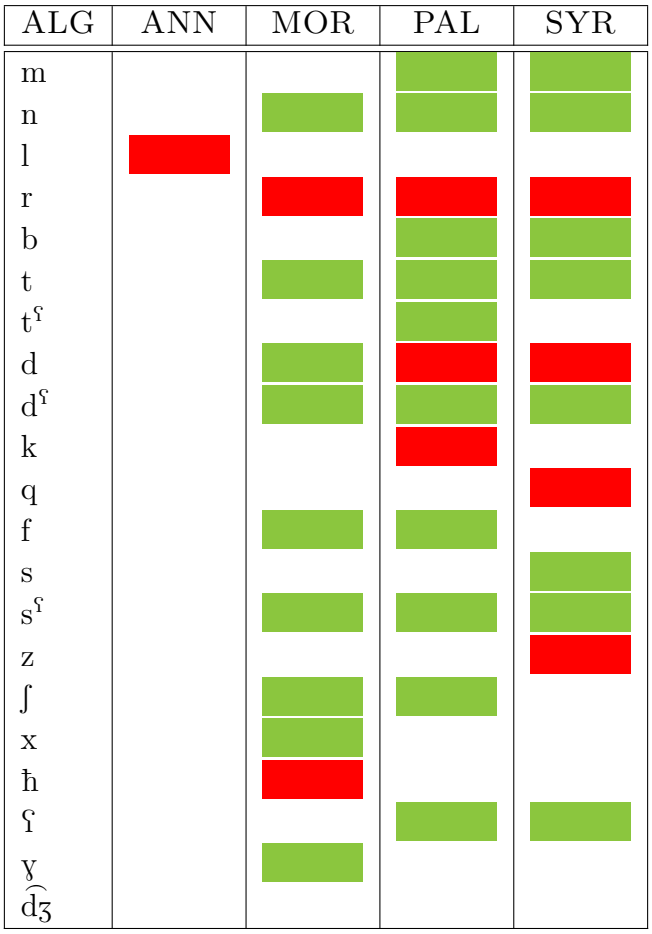

Figure 14: Results of comparing the Annaba, Moroccan, Palestinian and Syrian dialects to the Algiers dialect.

In the tables, the dialect in the heading of the leftmost column is the dialect to which the comparisons are made: for example, the table in figure 14 shows the comparisons of the Annaba, Moroccan, Palestinian and Syrian dialects with respect to the Algiers dialect. One striking characteristic of this table is the Annaba column. In this column we only see one significant difference in phoneme distribution, which seems fitting as Annaba and Algiers are the geographically closest dialects. This also reflects the finding of how the Algiers and Annaba dialects differ similarly from Modern Standard Arabic (figure 13). From the tables in figures 14 and 15, we can note the following: (i) the bilabials /b/, /m/ and the voiced pharyngeal fricative /ʔ/ are significantly more frequent in the Syrian dialect, while

the uvular plosive /q/ is less frequent; (ii) the Palestinian and Syrian dialects have a higher frequency of the emphatic plosives /sˤ/ and /dˤ/; (iii) the voiced velar fricative /ɣ/ has a higher frequency in the Moroccan dialect.

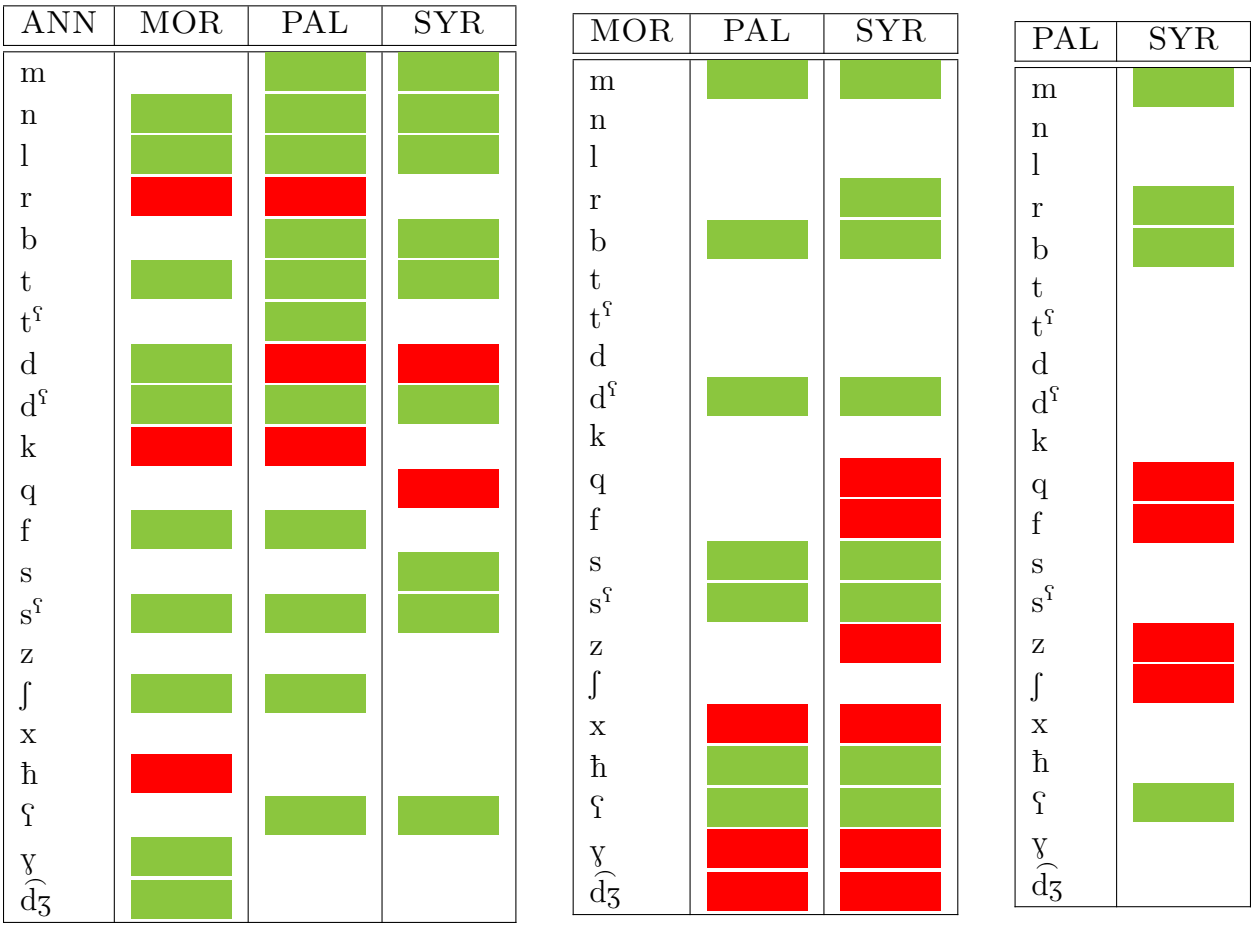

Figure 15: Left table shows remaining comparisons to the Annaba dialect; center table shows remaining comparisons to the Moroccan dialect; right table shows remaining comparison between the Palestinian and Syrian dialect.

4 Using *dialectness factor* for phonemic distance

The results from the previous section are interesting in a more practical sense: they give us an idea of whether the distribution of a phoneme in a dialect is significantly different than its distribution in other dialects, which could possibly have some application (perhaps as a heuristic) for natural language processing tasks. This section will look further into a

more conceptual question. In this section, we will ask the question: Given the *dialectness factor* of a phoneme in each dialect, how are these dialects situated in a high-dimensional vector space, where each dimension corresponds to a phoneme? Which dialects are close to Modern Standard Arabic (the origin)? In which dimensions do dialects appear similar? How 'close' are these points? We will first discuss the idea about how we embed these dialects in the vector space described above, and after we establish that concept, we will look at each dialect's coordinates in this space and look at their distances.

4.1 Phonemic space with *dialectness factor*

Our goal is to have some sense of the phonemic distance dialects of Arabic are from one another: to do this, we need (i) a vector space in which to work and (ii) some way to assign dialects to points in this vector space. The vector space we will use is $\mathbb{R}^{21}$, which we will call Arabic's *phonemic space*, where each of the twenty-one dimensions corresponds to one of the consonant phonemes we have been looking at throughout the paper — i.e. each vector in this space will have twenty-one coordinates with each coordinate representing some relation to a phoneme.[9] The relation we will choose is an adapted version of Zaidan and Callison-Burch (2014)'s *dialectness factor*: we will use the logarithm of this value.

(16) Vector in *phonemic space*: $v = (\xi_1, \xi_2, \ldots, \xi_{21})$ for $\xi_1 \mapsto$ /b/, $\ldots$, $\xi_{21} \mapsto$ /z/

(17) Coordinate assignment: $\xi_y = \log($*dialectness factor* of $y)$, where y is a phoneme

We will use the logarithms of the *dialectness factor* for a couple reasons: (i) the *dialectness factor* of a phoneme is a multiplicative relation, so this makes our computations additive; and (ii) this gives us a intuitive origin for the vector space, which is Modern Standard Arabic. Since a phoneme's *dialectness factor* is defined by its relationship to Modern Standard Arabic, the *dialectness factor* for Modern Standard Arabic will always be 1, making the logarithm 0. As for the first reason, this gives us the notion of positive numbers indicating a phoneme

[9] If we were to extend this analysis to vowels, we would say the vector space in which we are working now is a subset of Arabic's *phonemic space*.

occurring more frequently with respect to the Modern Standard Arabic and negative numbers indicating a phoneme occurring less frequently. In addition, if phoneme has a *dialectness factor* of x and another has a *dialectness factor* of $\frac{1}{x}$, using logarithms gives us symmetry in that the coordinates for this phoneme will be $\log(x)$ and $-\log(x)$ respectively: example (19) demonstrates this.

(18) Implications of (17):

(i) MULTIPLICATIVE $\rightsquigarrow$ ADDITIVE

$\log(\textit{dialectness factor}) = \log\left(\frac{\%\text{ of } y \text{ in dialect}}{\%\text{ of } y \text{ in MSA}}\right) =$

$\log(\%\text{ of } y \text{ in dialect}) - \log(\%\text{ of } y \text{ in MSA})$

(ii) MSA IS ORIGIN

$\log(\textit{dialectness factor}) = \log\left(\frac{\%\text{ of } y \text{ in MSA}}{\%\text{ of } y \text{ in MSA}}\right) =$

$\log(\%\text{ of } y \text{ in MSA}) - \log(\%\text{ of } y \text{ in MSA}) = 0$ for all phonemes y

(19) Interpretation of coordinates:

- $y > 0$: frequency of y in dialect $>$ frequency of y in MSA

  $y < 0$: frequency of y in dialect $<$ frequency of y in MSA

- Suppose *dialectness factor* of $y_1 = 2$ and *dialectness factor* of $y_2 = \frac{1}{2}$. This implies

  $\xi_{y_1} = \log(2) = 0.6931$

  $\xi_{y_2} = \log(\frac{1}{2}) = -0.6931$

The measure we will use is the Euclidean norm $||\cdot||_2$. This means the distance between two dialects D_1 and D_2 will be the square root of the sum of the squared distances between each phoneme's coordinates, as in (20).

(20) Distance between D_1 and D_2: $\quad d(D_1, D_2) = \sqrt{\sum_{i=1}^{21}(y_{1i} - y_{2i})^2}$

4.2 Coordinates for each dialect in *phonemic space*

The plots below will show the coordinates of each phoneme (dimension) in *phonemic space*; having bar plots below helps bring out the contrasts and similarities we found between the dialects in section 3.3. The line at 0 indicates the origin in that dimension, which represents Modern Standard Arabic; hence, a bar extending above this line indicates the phoneme occurred more frequently and vice versa for bars extending below the line. These plots are just another way of visualizing the information from the coordinates, discussed just above in section 4.1.

From these plots, we can get an intuitive idea of which dialects will be close to one another. The more two dialects have bars which extend on the same side of zero and have similar lengths, the more similar the dialects will be. We can see again that the Algiers and Annaba dialect will be closer to each other than the others: the bars for these dialects always extend on the same side of zero and are for the most part similar in length. Another pair of dialects which should be close to each other are the Palestinian and Syrian. Only in the plot for /m/ do we see a difference in which side of zero the bars extend. The Moroccan dialect appears to be somewhere off on its own. In some instances, it patterns alongside both the Algiers and Annaba dialects and other times the Palestinian and Syrian dialects. We can also see from these plots which phonemes have a higher degree of being dialectal. The palato-alveolar fricative /ʃ/ is by far the most dialectal with it occurring much more in every dialect compared to Modern Standard Arabic. The phonemes /ħ/ and /z/ are also dialectal to an above average extent. Last, we see instances of certain phonemes being unique to some dialects. Most notably, we have /ɣ/ in dramatic fashion occurring more frequently in the Moroccan dialect. Likewise, we see that /q/ occurs much less in the Syrian dialect.

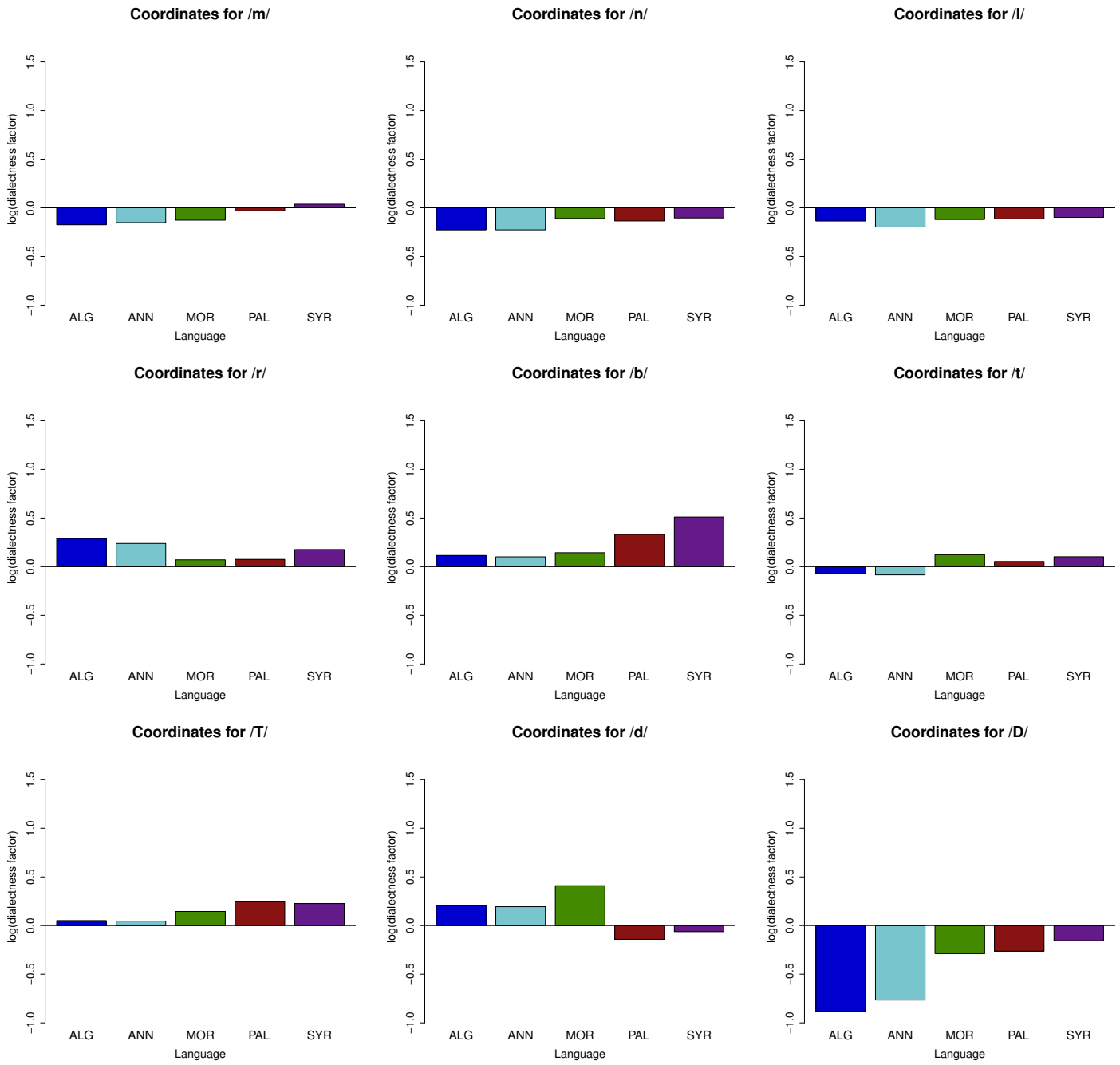

Figure 16: Bar plots of the coordinates in *phonemic space* for each dialect.

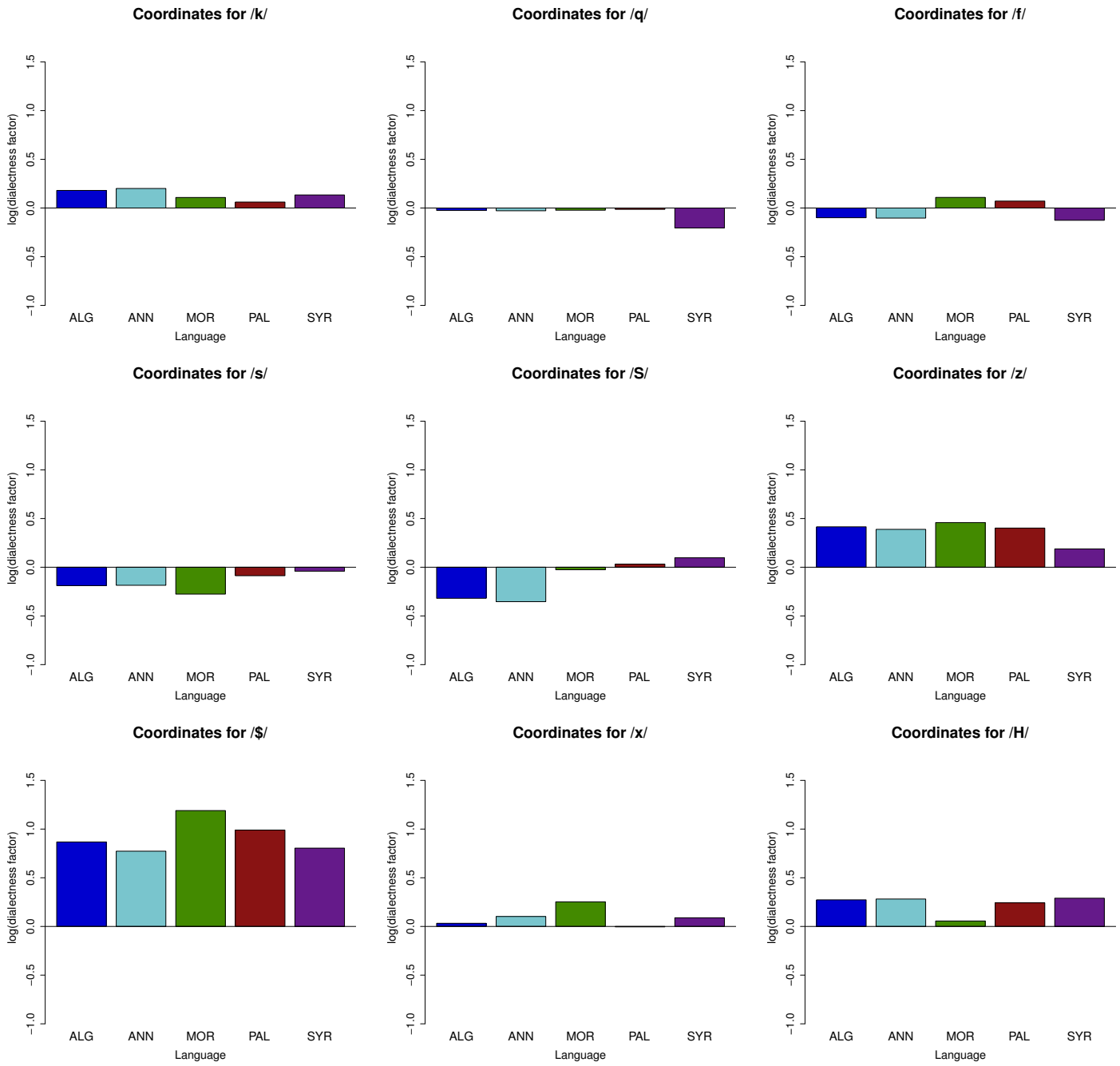

Figure 17: Bar plots of the coordinates in *phonemic space* for each dialect.

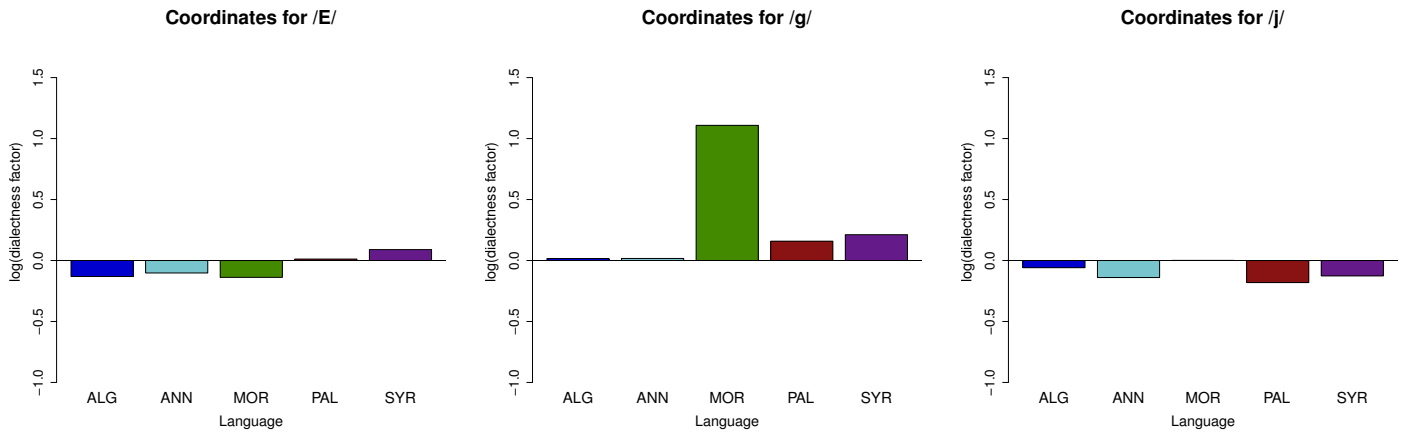

Figure 18: Bar plots of the coordinates in *phonemic space* for each dialect.

4.3 Computing distances in *phonemic space*

As mentioned in (20), the measure we will use to compute distances between dialects is Euclidean distance. The table of distances between each pair of dialects is shown in figure 19. Each cell (i, j) indicates the distance between dialect i and dialect j. We see that the Moroccan dialect is the farthest from Modern Standard Arabic while the Palestinian and Syrian dialects are the closest. In line with our previous observations, we see that the Algiers and Annaba dialect are remarkably close to each other, and we see that the Syrian and Palestinian dialects are close to each other as well.

	ALG	ANN	MOR	PAL	SYR
MSA	1.631	1.503	1.927	1.304	1.291
ALG		0.239	1.443	1.032	1.169
ANN			1.424	0.981	1.08
MOR				1.277	1.335
PAL					0.68

Figure 19: Table of distances between each pair of dialects (units of log(*dialectness factor*)).

In addition to computing the distance using the entire *phonemic space*, we can consider any arbitrary subset of it. Take, for example, the subset corresponding to the phonemes /b/, /q/ and /ʔ/. The dialects appear as in figure 20 in this subspace. However, most

likely, we will only be interested in those subsets which have some linguistic motivation, such as those subsets that pertain to a manner of articulation. For these dialects of Arabic, the phonemic distances based on manner of articulation are given the tables in figure 21. The first table looks at distance with respect to plosives; the second looks at fricatives; and the last looks at the two nasals, alveolar lateral and alveolar trill. From the tables, we see that the distribution of fricatives account for the most differences with each dialect when compared to Modern Standard Arabic. It is also the case that fricatives account for most of the differences between dialects, but we see it is the distribution of stops which accounts for more of the differences between the Algiers and Annaba dialects compared to the Palestinian and Syrian dialects.

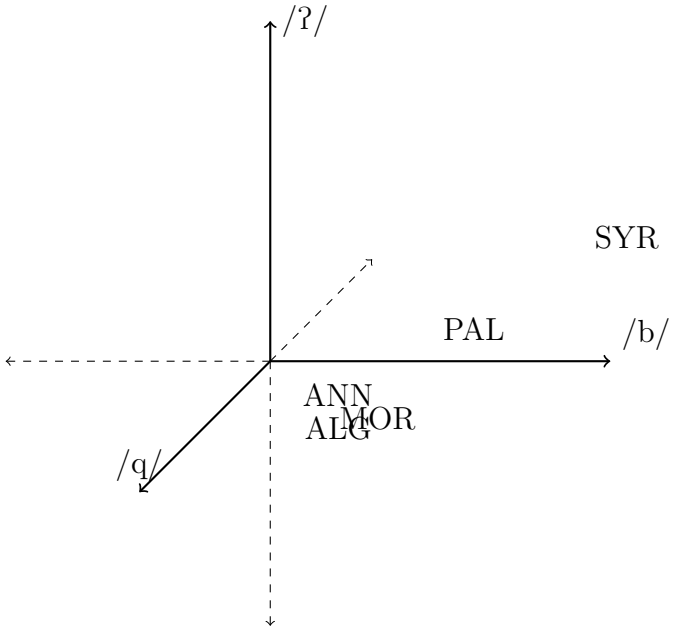

Figure 20: Plot of the points for each dialect in the subset of *phonemic space* which corresponds to /b/, /q/ and /ʔ/.

stops	ALG	ANN	MOR	PAL	SYR
MSA	0.934	0.827	0.565	0.515	0.641
ALG		0.12	0.667	0.785	0.922
ANN			0.582	0.705	0.842
MOR				0.598	0.646
PAL					0.307

fric.	ALG	ANN	MOR	PAL	SYR
MSA	1.08	1.01	1.741	1.113	0.9246
ALG		0.129	1.237	0.471	0.585
ANN			1.267	0.531	0.584
MOR				1.053	1.134
PAL					0.381

other	ALG	ANN	MOR	PAL	SYR
MSA	0.432	0.435	0.218	0.265	0.262
ALG		0.115	0.26	0.3	0.28
ANN			0.261	0.241	0.253
MOR				0.208	0.234
PAL					0.138

Figure 21: Table of distances between each pair of dialects based on manner: the top left table corresponds to plosives; the top right corresponds to fricatives; and the bottom corresponds to /m, n, r, l, d͡ʒ/

5 Conclusion and future directions

This paper took seriously the notion of *dialectness factor* as introduced and briefly discussed in Zaidan and Callison-Burch (2014). In the computation of a phoneme's *dialectness factor*, we use the percentage of occurrence of a phoneme in a dialect of Arabic versus its percentage of occurrence in Modern Standard Arabic. To see if significant differences in phoneme distribution could be captured using percentages, we gathered data from PADIC, a corpus in Modern Standard Arabic and also translated in parallel for the Algiers, Annaba, Moroccan, Palestinian and Syrian dialects. To summarize this data, we chose to fit a one-way linear model. We found that using this model, we could recover many significant differences between dialects based on the distributions of phonemes. Some of these differences seemed to make intuitive sense, such as the Algiers and Annaba dialects being nearly indistinguishable.

The next part of the paper involved using the data from the corpus to compute the *dialectness factor* of each phoneme, and then we went on to discuss how we can use this as a way of embedding the dialects in a high-dimensional vector space where each dimension

corresponded to a phoneme; we called this *phonemic space*. We chose to use the logarithm of the *dialectness factor*, giving us a more natural sense of the geometry of the dialects in these spaces. Once we embedded the dialects into *phonemic space*, we computed the distances between the dialects using the Euclidean measure. The Palestinian and Syrian dialects were found to be closest to Modern Standard Arabic and the Moroccan dialect farthest, based on the *dialectness factor* of the phonemes. We went on to quantify the distance between the various dialects, and then looked at a few meaningful subspaces of *phonemic space*, based on manner of articulation. Looking at these subspaces gave us a more nuanced idea of how dialects differ from Modern Standard Arabic and each other, with most of the differences being due to the distribution of fricatives.

A straightforward extension to this work would be to go up one level of abstraction and look at a language family. It would be interesting to see how languages we have thought to not be dialects of one another behave with respect to looking at just differences in phoneme distribution. As we saw from this study, even languages which are considered dialects of one another have many significant differences in their distribution of phonemes. After looking at the differences in phoneme distributions, we could then compute the *dialectness factor* of phonemes shared by members of a language family, embed them into another vector space and compute the distances between them. If looking at this synchronic relationship between languages in a family is productive, another avenue of this research is to extend it to diachronic relationships. The most straightforward way to do this would be to use the parent language as the basis for comparison — i.e. the parent language would be analogous to Modern Standard Arabic and the child languages would be analogous to the dialects in this paper. This could provide a more nuanced view of how languages have evolved from the parent language; presumably, the child languages are not all equidistant from the parent, contrary to as they appear in many language trees.

Going up yet another level of abstraction, we could consider comparing language families, but there are many difficulties which could crop up in an investigation of that nature. First,

this methodology depends having a non-empty set of shared phonemes, with the hope that this set is not small. When comparing language families, this will certainly not be the case, and it then becomes unclear how to embed the language families meaningfully into a vector space. One way we may approach this would be to restrict our attention to vowels; however, it is not clear that this would avoid the problem. Last, a more practical concern about the two previous extensions to this work is the availability of resources. This work benefitted from the abundance of data in the PADIC corpus, and having access to a sizable corpus (in a friendly format) is not to go overlooked.

References

Biadsy, F. and Hirschberg, J. (2009). Using prosody and phonotactics in arabic dialect identification. In *Tenth Annual Conference of the International Speech Communication Association*.

Biadsy, F., Hirschberg, J., and Habash, N. (2009). Spoken arabic dialect identification using phonotactic modeling. In *Proceedings of the eacl 2009 workshop on computational approaches to semitic languages*, pages 53–61. Association for Computational Linguistics.

Carrilho, E., Magro, C., and Álvarez, X., editors (2013). *Current Approaches to Limits and Areas in Dialectology*. Cambridge Scholars Publishing.

Chambers, J. and Trudgill, P. (2012). *Dialectology*. Cambridge Textbooks in Linguistics.

Davis, L. M. (1990). *Statistics in Dialectology*. The University of Alabama Press.

Linn, M. D., editor (1998). *Handbook of Dialects and Language Variation*. Academic Press, 2 edition.

Meftouh, K., Harrat, S., Jamoussi, S., Abbas, M., and Smaili, K. (2015). Machine transla-

tion experiments on padic: A parallel arabic dialect corpus. In *The 29th Pacific Asia Conference on Language, Information and Computation*, Shanghai, China.

Mehrabani, M., Bořil, H., and Hansen, J. H. (2010). Dialect distance assessment method based on comparison of pitch pattern statistical models. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 5158–5161. IEEE.

Nerbonne, J. (2009). Data-driven dialectology. *Language and Linguistics Compass*, 3(1):175–198.

Nerbonne, J. and Heeringa, W. (2010). Measuring dialect differences. *Language and Space: Theories and Methods. Berlin: Mouton De Gruyter*, pages 550–566.

Schouten, M. and van Reenen, P., editors (1989). *New Methods in Dialectology.* Foris Publications.

Vogelaer, G. D. and Seiler, G., editors (1998). *The Dialect Laboratory.* John Benjamins Publishing Company.

Wieling, M., Margaretha, E., and Nerbonne, J. (2012). Inducing a measure of phonetic similarity from pronunciation variation. *Journal of Phonetics*, 40(2):307–314.

Zaidan, O. F. and Callison-Burch, C. (2014). Arabic dialect identification. *Computational Linguistics*, 40:171–202.