Homophone duration in spontaneous speech: A mixed-effects model

Susanne Gahl

August 2009[i]

1. Introduction

A recent analysis of a corpus of spontaneous speech (Gahl, 2008) showed that homophone pairs differed in duration as a function of word frequency. For example, the high-frequency word *time* was shorter on average than its less-frequent homophone twin *thyme*. This effect persisted when other factors affecting word duration were statistically controlled for in a linear regression model. However, that model had several serious limitations. The goal of the current study is to overcome these limitations and to explore the determinants of word duration further.

The model presented in Gahl (2008) was a linear regression model. Its outcome variable was the average duration of the higher-frequency member of the homophone pairs. Predictors were entered into the model in a blockwise fashion, in three separate blocks. The sole predictor in the first block was the average duration of the lower-frequency member of a homophone pair. The second block contained known determinants of word duration in connected speech, such as contextual speaking rate, the probability of a word given neighboring words in an utterance, orthographic regularity, and proximity to pauses. On the third and final block, the frequency of the higher-frequency member of the homophone pair (i.e. the frequency of the word whose duration was to be predicted by the model) was entered, to ascertain whether word frequency was a significant predictor of word duration over and above other known factors. That question has theoretical implications for linguistic and psycholinguistic models of language production, which are discussed in Gahl (2008).

The modeling strategy of predicting the average duration of the higher frequency member of the homophone pair (e.g. *time*) from the lower frequency member (e.g. *thyme*) is problematic in a number of ways. For one thing, information specific to the lower frequency homophone never entered into the model predictions, except indirectly, via the duration of the lower-frequency homophone. For example, while the orthographic regularity of the high-frequency homophone was a predictor in the model, the orthographic regularity of the lower-frequency homophone was not. This meant that the homophone twin was not a perfect control for the effect of phonemic content on word duration, since the duration of, for example, *thyme* in part reflect the orthographic regularity of that word. Properties specific to the low-frequency words were never entered into the model.

A further problem with the modeling strategy in Gahl (2008) was that information about specific word tokens was lost to the model. All of the predictors in the model represented information about word types, not word tokens. For some variables, this is as it should be. Word frequency, for example, is a property of a word type: The frequency of the word *thyme* is a property of the word type *thyme*, not of an individual token. By contrast, whether the word *thyme* immediately precedes a pause in an utterance, on the other hand, is a property of a specific token of the word. Information about proximity to

pauses was entered into the model by determining the proportion of all tokens of *thyme* that immediately preceded a pause and using that proportion as a predictor in the model. Similarly, the measure of contextual speaking rate in the previous model was the average contextual speaking rate of all tokens representing a word type.

The use of summary measures imposes a serious limitation on the model, in that the proportion of pre-pausal tokens may fail to capture crucial interactions between pausing and the other predictors. It is conceivable, for example, that the effect of orthographic regularity on word duration, which is subtle under the best of circumstances, becomes negligible before pauses or major disfluencies. If that is so, then model will underestimate the overall importance of orthographic regularity. This constitutes a serious limitation, especially since the goal of Gahl (2008)'s analysis was to establish whether frequency was a significant predictor of word duration over and above other known determinants of word duration: Since the effect of frequency on word durations is controversial, it is important to give all other predictors a chance to explain the observed durations, without reference to word frequency. The same goes for other token-specific predictors, such as contextual speaking rate.

The goal of the current study is to explore the various determinants of word duration further, using a regression model that avoids the problems just pointed out. The model is a mixed-effect regression model, with token durations of both members of each homophone pair as the outcome variable. This model is expected to have greater power for detecting the effect of the predictors on token duration.

2. Data

The data for the study come from the time-aligned transcript of the Switchboard corpus (Deshmukh, Ganapathiraju, Gleeson, Hamaker, & Picone, 1998), a collection of 240 hours of recorded telephone conversations between strangers (Godfrey, Holliman, & McDaniel, 1992). Information about the word types, such as syntactic category and several other variables was extracted from the CELEX lexical database (Baayen, Piepenbrock, & van Rijn, 1993). In constructing the database for the current paper, I followed the procedure described in Gahl (2008)'s paper: First, all sets of words with identical phonetic transcription but different orthography were extracted from the CELEX database. Unlike the analysis in Gahl (2008), which took into account only pairs of homophones, the current analysis also took into account sets with more than two members, such as *right/write/rite*. Several classes of homophone sets were then excluded from the database, following Gahl (2008): (1) Sets including function words and names of letters of the alphabet, such as *b/be/bee;* (2) Sets such as *sauce/source*, which are homophones based on the British English transcription used in CELEX, but not in American English; (3) Sets including orthographic forms representing more than one phonemic form. Thus, the pair *read - red* was not included, as *read* represents two different pronunciations; (4) Sets for which only one member was attested in Switchboard. On a final step, the duration of each token of all of the words on the resulting wordlist was then extracted from the Switchboard corpus. This procedure yielded a database of 79,867 tokens.

3. Methods

To facilitate comparison of the current model to that in Gahl (2008), I included all of the predictors that were in that previous model. In addition, the new model included the speaker's age and sex. A description of all variables included in the models and of any transformations applied to the raw values can be found in Table 1.

logDur	The log-transformed, centered duration of the token, in milliseconds. Datapoints with log durations of less than -4, i.e. raw durations of 18 milliseconds, were removed from the dataset. There were only two such data points.

Predictors relating to the context in which the token was found:

lRate, rRate	The local speaking rate in the region before (lRate) or after (rRate) the target token, based on the stretch of speech bounded by a pause or conversational turn and the target token; log-transformed and centered.
lBigramProb, rBigramProb	The bigram probability of the token given the preceding (lBigramProb) or following (rBigramProb) word. The data set used here excluded utterances in which the token had no immediate neighbor, i.e. was initial or final in an utterance. Both measures were log-transformed and centered.
prePausal	A binary variable indicating whether the target immediately preceded a pause of 500 ms or more.
preDisfl	A binary variable indicating whether the target immediately preceded a disfluency, i.e. a hesitation sound such as *um* or *uh.*
age	The age of the talker.
sex	The sex of the talker.

Predictors relating to context-independent properties of the target words:

len	Length in letters.
berndt	Orthographic regularity, based on published grapheme-to-phoneme probabilities (Berndt, Reggia, & Mitchum, 1987).
nQuot	The proportion of nouns in the target word's frequency count in CELEX. For example, the form *stake* has a "noun quotient" (nQuot) of .935, reflecting the much higher frequency of *stake* as a noun than as a verb. As discussed in Gahl (2008) and references cited therein, syntactic category affects word durations, in part

	because nouns are more likely than verbs to occur at the end of a phrase and hence undergo phrase-final lengthening.
logSwbdFq	The word frequency in the Switchboard corpus, log-transformed and centered.

Additionally, the model included the identity of the speaker (**callerId**) and of the pronunciation (**pron**, i.e. the form, which is the same for the two members of a homophone pair) as random-effects factors.

Tables 1A and 1B show summary statistics for the continuous (Table 1A) and categorical (Table 1B) variables in the model.

Variable	Mean	Median	SD
Token duration (ms)	312	290	13.6
Speaking rate preceding the target	3.1 syl/sec	3.6 syl/sec	1.00
Speaking rate following the target	2.9 syllables/second	3.1 syl/sec	0.69
Bigram probability given the previous word	0.04	0.008	0.10
Bigram probability given the following word	0.03	0.004	0.12
Length in letters	4.36	4.00	0.97
Orthographic regularity	0.93	1.00	0.12
Noun proportion	0.38	0.14	0.44
Speaker age	37.2 years	34 years	11.0
Frequency in Switchboard	5754	4066	6124

Table 1A: Summary statistics of homophone tokens in the Switchboard corpus: Continuous variables

Categorical predictors:	
Immediately preceding a pause	True: 6024
	False: 73843
Immediately preceding a disfluency	True: 17307
	False: 62560
Speaker sex	Female: 39122
	Male: 40745

Table 1B: Summary statistics of homophone tokens in the Switchboard corpus: Categorical variables

Figure 1 summarizes the correlational structure of the numerical predictors in a hierarchical clustering plot, using Spearman's rho^2 as a distance measure. Interestingly, frequency and local speaking rate clustered together, with the speaking rate in the region following the target clustering most closely with the frequency measure. Although the bivariate correlations are not very high, the clustering of local speaking rate and frequency confirms the importance of keeping speaking rate under statistical control in an investigation of frequency effects on word duration.

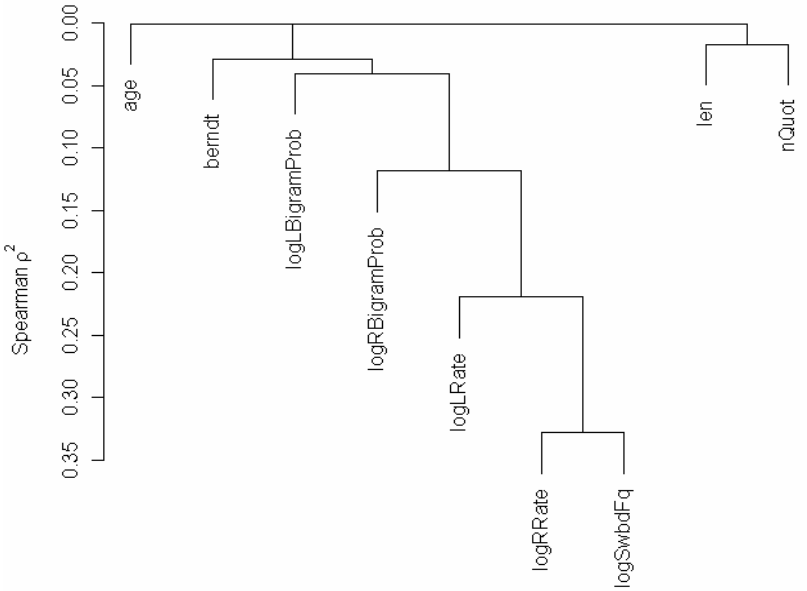


Figure 1. Hierarchical clustering of predictors of word duration

4. Results

The main empirical question in Gahl (2008) was whether "lemma frequency", e.g. the frequency of the word *time* vs. the word *thyme*, was a significant predictor of word duration, once other factors were controlled for. As argued above, the model developed in Gahl (2008) is vulnerable to a number of objections. The current study presents a model of word durations at the token level that addresses these objections. As in Gahl (2008), the current study reports models with and without frequency as a predictor and uses model comparison by means of sequential ANOVAs to decide whether the including frequency in the model led to a significant improvement. The current model uses token-level information wherever possible, in contrast to the fixed-effects model of Gahl (2008), which predicted average durations based on information about word types, such as average local speaking rate. The outcome variable in the current model was the centered, log-transformed duration of each homophone token, for example all tokens of *time* and all tokens of *thyme*, that were found in the Switchboard corpus. Data preparation and statistical analysis was carried out using R (R Development Core Team, 2008), and in particular the Design (Harrell, 2008)and languageR (Baayen, 2008) packages.

Table 2 shows the regression coefficients and summary information for the "baseline" model, i.e. a model including all factors except frequency.

Variable	beta	SE	t
Intercept	0.0442	0.0129911	3.40
cLogLBigram	-0.0084185	0.0006519	-12.91
cLogRBigram	-0.0215032	0.0006842	-31.43
cLogLRate	-0.1279349	0.0225121	-5.68
cLogRRate	-0.0353013	0.0248146	-1.42
cLogBerndt	-0.1183335	0.0326874	-3.62
cLogNQuot	0.1804308	0.0144830	12.46
clen	0.0568600	0.0052771	10.77
cAge	0.0016734	0.0004152	4.03
sexMALE	-0.0793775	0.0090479	-8.77
prePausalTRUE	0.4060329	0.0045876	88.51
preDisflTRUE	0.4105785	0.0030670	133.87

Table 2: Summary of the "baseline" model of word duration, i.e. a model without a frequency term. The model has random intercepts for speaker (s = 0.095) and phonemic content (i.e. a grouping variable that groups together all homophonous tokens, e.g. all tokens of *thyme* and *time* as one group; the standard deviation for that random effect was 0.13); the standard deviation of the residual was 0.32.

The pattern of significant effects was similar to what was found in Gahl (2008): Words tend to lengthen immediately preceding pauses or disfluencies. High contextual probability, high speaking rate in the stretch preceding the token, and high orthographic regularity predict shorter word durations. Male speakers tend to produce shorter word durations than female speakers. Greater length in letters, higher speaker age, and stronger noun-bias promote longer word durations.

Is frequency a determinant of word durations, over and above the other predictors? Table 3 shows the regression coefficients and summary information for a model identical to the baseline model, but with the addition of frequency as a predictor. Again, all factors showed an effect in the same direction as found in Gahl (2008): Pre-pausal position or occurrence right before a disfluency and high noun-quotient (which tends to increase the likelihood of a word's occurring in phrase-final position) are associated with longer word durations. High frequency, high contextual speaking rate in the stretch of speech preceding the token, and predictability, and orthographic regularity are all associated with shorter word durations. Speaking rate in the region following the target yielded a marginally significant p-value ($p = .05$).

Variable	beta	SE	t
Intercept	0.0122744	0.0138638	0.89
cLogLBigram	-0.0079965	0.0006628	-12.06
cLogRBigram	-0.0210636	0.0006950	-30.31
cLogLRate	-0.1512506	0.0230406	-6.56
cLogRRate	-0.0480938	0.0245853	-1.96
cLogBerndt	-0.1418903	0.0312169	-4.55
cLogNQuot	0.1744043	0.0142783	12.21
clen	0.0526141	0.0051433	10.23
cAge	0.0016767	0.0004135	4.06
sexMALE	-0.0793267	0.0090104	-8.80
prePausalTRUE	0.4062573	0.0046101	88.12
preDisflTRUE	0.4108366	0.0030824	133.28
cLogSwbdFq	-0.0110719	0.0025900	-4.27

Table 3: Summary of the mixed-effects model of word duration, including word frequency. The model has random intercepts for speaker ($s = 0.095$) and phonemic content (i.e. a grouping variable that groups together all homophonous tokens, e.g. all tokens of *thyme* and *time* as one group; the standard deviation for that random effect was 0.13); the standard deviation of the residual was 0.32.

The two models – with and without frequency as a predictor – were compared using a sequential ANOVA, evaluating whether adding frequency to the model significantly improved the model. The model comparison shows the addition of the frequency variable to be justified ($\chi^2(1) = 15.95$, $p < .00001$)

Given the theoretical significance of the effect of frequency, and given the small expected size of that effect, the significance levels of the model coefficients was explored further using Markov chain Monte Carlo sampling. Table 4 shows the coefficients and associated statistics.

	Estimate	MCMCmean	HPD95lower	HPD95upper	pMCMC	Pr(>|t|)
Intercept	0.0123	0.0076	-0.0189	0.0361	0.5994	0.3760
cLogLBigram	-0.0080	-0.0080	-0.0093	-0.0068	0.0001	0.0000
cLogRBigram	-0.0211	-0.0211	-0.0225	-0.0198	0.0001	0.0000
cLogLRate	-0.1513	-0.1482	-0.1927	-0.1033	0.0001	0.0000
cLogRRate	-0.0481	-0.0528	-0.1009	-0.0043	0.0322	0.0504
cLogBerndt	-0.1419	-0.1479	-0.2115	-0.0881	0.0001	0.0000
cLogNQuot	0.1744	0.1750	0.1465	0.2023	0.0001	0.0000
Clen	0.0526	0.0551	0.0448	0.0654	0.0001	0.0000
cAge	0.0017	0.0017	0.0009	0.0025	0.0001	0.0001
sexMALE	-0.0793	-0.0794	-0.0960	-0.0620	0.0001	0.0000
prePausalTRUE	0.4063	0.4063	0.3975	0.4155	0.0001	0.0000
preDisflTRUE	0.4108	0.4109	0.4048	0.4169	0.0001	0.0000
cLogSwbdFq	-0.0111	-0.0116	-0.0167	-0.0067	0.0001	0.0000

Table 4: Coefficients and associated statistics for the model including word frequency as a predictor of word duration. "MCMC mean" denotes the mean value of the coefficient across 500 Markov chain Monte Carlo samples of the posterior distribution of the parameters. "HDP lower" and "HPD upper" denote the lower and upper bounds of the Highest Posterior Density interval for 95% of the probability density. "pMCMC" denotes the associated probability value. p(t) denotes the probability based on the t-distribution with 79,855 degrees of freedom.

5. Discussion: Model evaluation

The robustness of the model including frequency was further explored by inspecting the residuals. The current model is a purely linear model, despite the fact that some of the predictors in the model are known to have quadratic and other non-linear effects on word duration. Exploring such non-linear relationships was left for future research.

Figure 2 shows the residuals for the model including frequency as a predictor. The plot suggests that the model is not very good at fitting words that are either unusually long or unusually short: This is to be expected – the model is linear in the middle range of data – one reason why extrapolating beyond the range of the data is not advised.

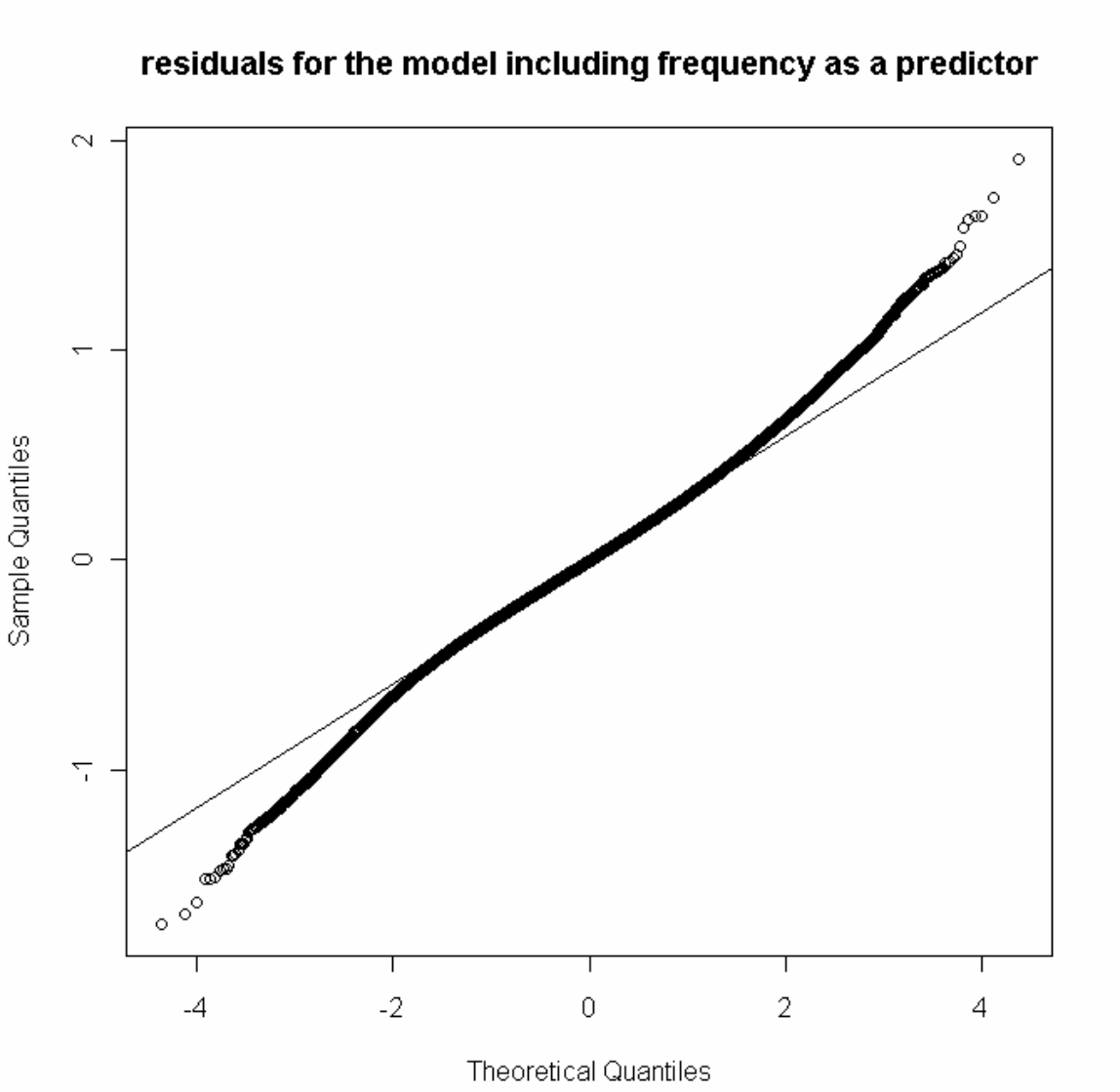


Figure 2. Model residuals for the model including frequency as a predictor of word duration

To ascertain whether the extreme data points were distorting the model estimates, I excluded all words with durations greater than 500 milliseconds or shorter than 100 milliseconds. The pattern of significances in the resulting models was unchanged compared to the original model, and the normality of the residuals was not improved much. I therefore report the model based on all data points here.

Below are plots of the partial effects of the models with and without frequency as a predictor, created with Harald Baayen's plotLMER.fnc (Baayen, 2008). The broken lines

represent MCMC-based HPD intervals. The solid line shows predicted values. The 95% HPD shows what to expect for the model predictions, including how "linear" the relationships are predicted to be. The predicted relationships appear to be fairly robustly linear relationships, with the exception of the effect of orthographic regularity ("cLogBerndt"), and possibly also the speaking rate in the region following the target ("cLogRRate").

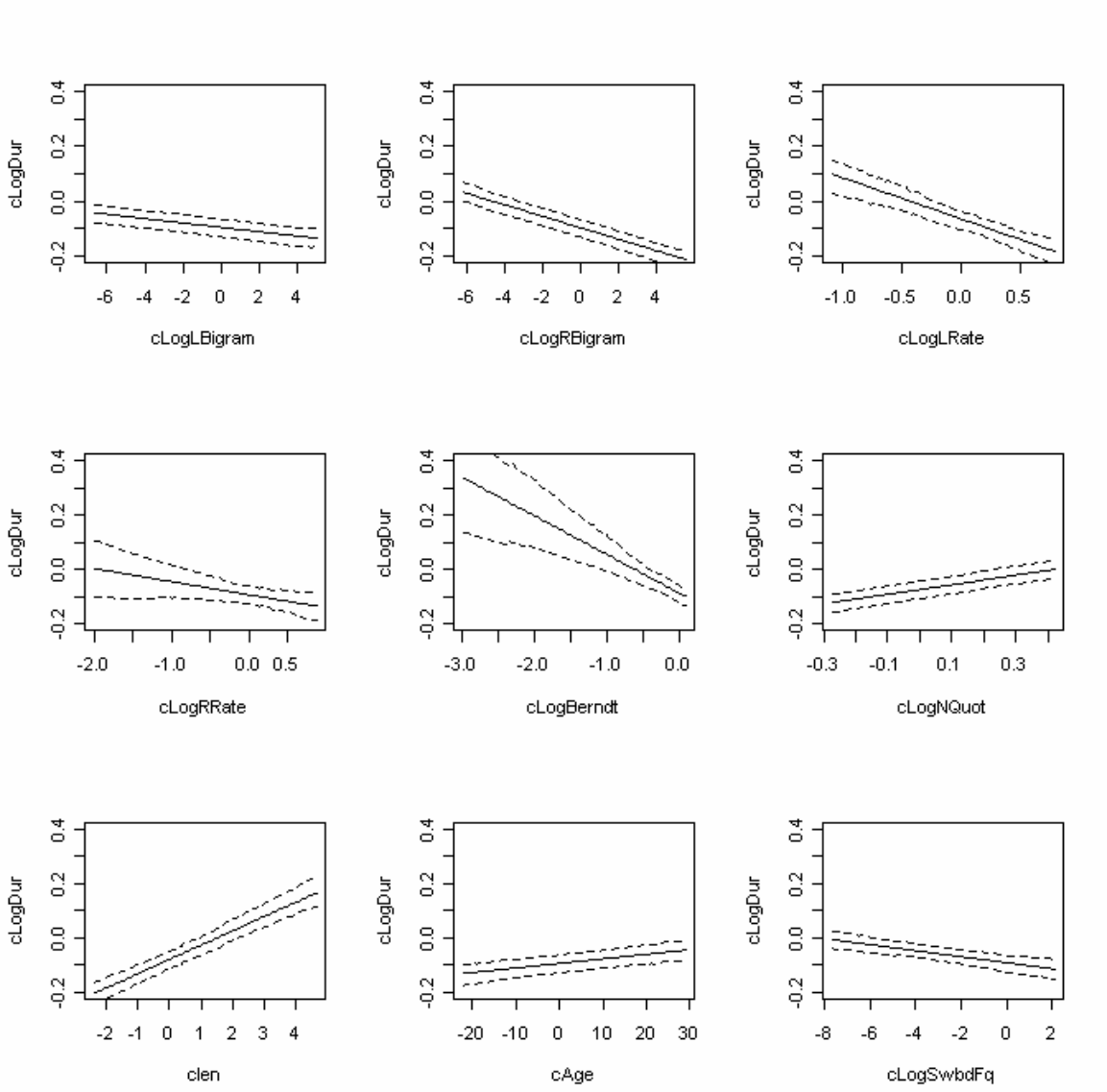


Figure 3. Partial effect plots of numerical predictors in the model of word duration

The model presented so far takes into account the same predictors and assumptions as the model in Gahl (2008), since one of the goals of the current study is to check whether the averaging procedure in Gahl (2008) produced a spurious result. The resulting model

replicates the findings in Gahl (2008) and suggests that the conclusion in that study still stands.

I next consider ways in which the model could be improved, first by excluding very short and very long tokens; then by taking into account the non-linear relationship between local bigram probabilities and word duration.

The model considered so far doesn't work very well for long token durations. One possible reason for this might be the fact that long tokens occur before disfluencies and/or pauses, and that those effects drown out all other effects for tokens before disfluencies or pauses. The mean duration of tokens preceding disfluencies was 415 milliseconds (vs. 284 for tokens not preceding disfluencies). Similarly, the mean duration of tokens immediately preceding a pause was 408 milliseconds, compared to 304 milliseconds for tokens that did not precede pauses. Moreover, the presence of pauses and disfluencies may well affect the predictive power of some of the other predictors in the model, such as bigram probabilities and local speaking rate. I therefore fit a separate model to the subset of the data that excluded tokens before disfluencies and pauses (n = 56536 out of the whole set of 79867). The pattern of fixed effects is unchanged, with the exception of the effects of speaking rate in the region following the target, which is now highly significant. Removing disfluencies and pauses is conceptually well-motivated and results in the removal of many data points near the top of the word duration scale; therefore removing such datapoints might be expected to improve model fit in the extreme data ranges. Table 5 summarizes that model.

Variable	beta	SE	t
Intercept	0.0031404	0.0154416	0.20
cLogLBigram	-0.0072150	0.0007872	-9.17
cLogRBigram	-0.0291421	0.0008164	-35.70
cLogLRate	-0.1264128	0.0256116	-4.94
cLogRRate	-0.1246426	0.0282804	-4.41
cLogBerndt	-0.1285027	0.0359311	-3.58
cLogNQuot	0.2049346	0.0162473	12.61
clen	0.0571749	0.0058635	9.75
cAge	0.0023477	0.0004133	5.68
sexMALE	-0.0538263	0.0090052	-5.98
cLogSwbdFq	-0.0105186	0.0029475	-3.57

Table 5: Summary of the mixed-effects model of word duration when tokens immediately preceding disfluencies or pauses were excluded. Like the previous models, the model has random intercepts for speaker (s = 0.091) and phonemic content (i.e. a grouping variable that groups together all homophonous tokens, e.g. all tokens of *thyme* and *time* as one group; the standard deviation for that random effect was 0.13); the standard deviation of the residual was 0.32.

Disappointingly, the distribution of residuals does not improve. Figure 4 shows the residuals of a model without disfluent or pre-pausal tokens, side-by-side with the residuals when all tokens are included.

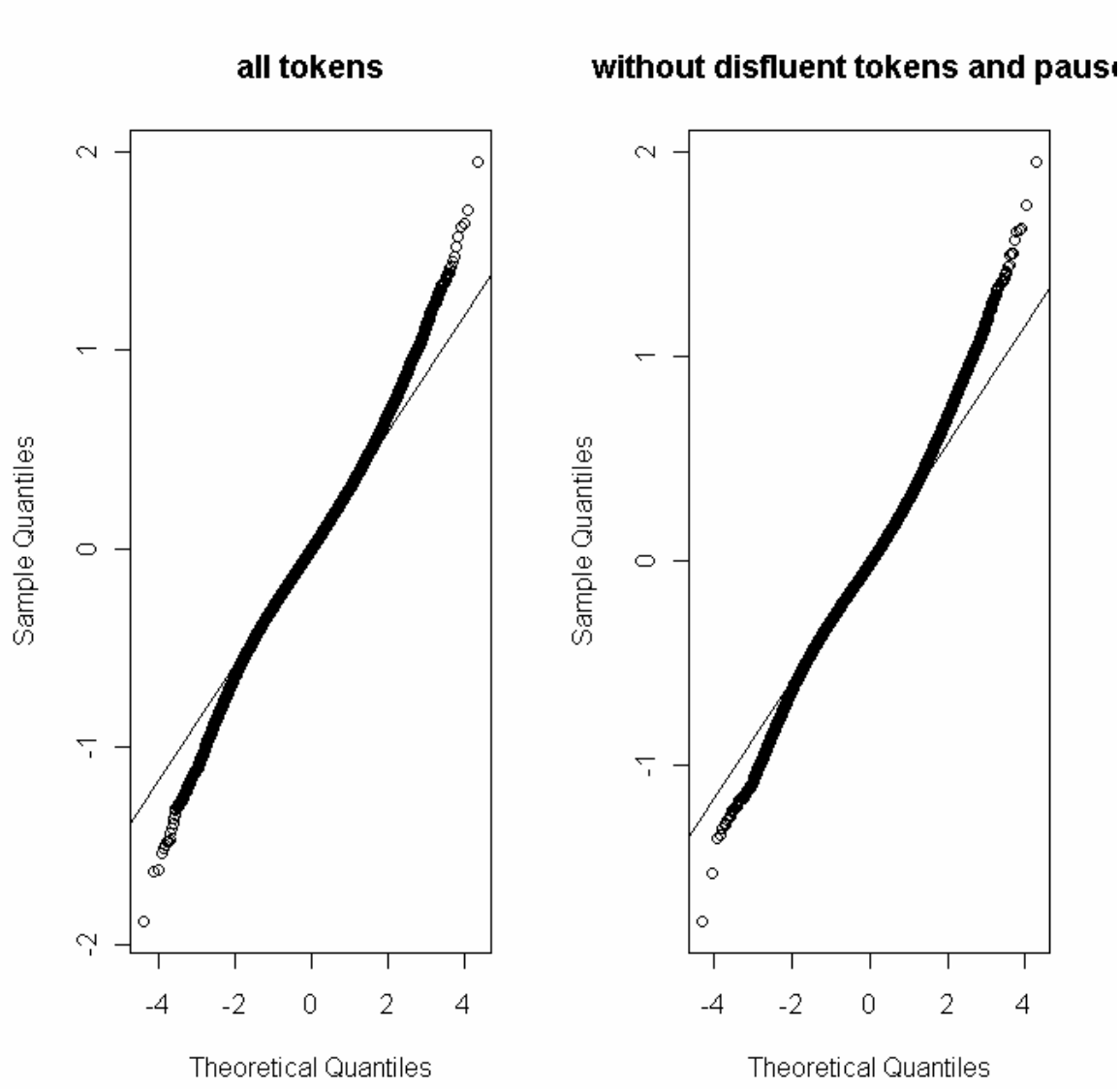


Figure 4. Model residuals for the full data set (left panel) and the subset of the data set excluding tokens immediately preceding pauses or disfluencies (right panel)

What about the shortest word durations, which the model also predicts poorly? Very short word durations likely represent tokens that pose special problems for the time-alignment, which may add unacceptable levels of noise to the short duration values. Notice that the raw word durations include durations barely long enough for a single segment. I therefore re-fitted the models, this time excluding tokens with durations below 120 milliseconds. Table 6 gives the descriptive statistics for the subset of the data with word durations above 120 ms.

	Mean	SD
Outcome variable: token duration	318 ms	
Continuous predictor variables:		
Speaking rate, preceding	3.1 syllables/second	
Speaking rate, following	3.11 syllables/second	
Bigram probability, given prev.	0.04	
Bigram probability, given foll.	0.03	
Length in letters	4.39	
Orthographic regularity ("berndt")	0.93	
Noun proportion	0.38	
Speaker age	37.2 years	
Frequency in Switchboard	5754	

Categorical predictors:

Immediately preceding a pause	True: 6024 False: 73843
Immediately preceding a disfluency	True: 17265 False: 60608
Speaker sex	Female: 38336 Male: 39537

Table 6: Descriptive statistics for the dataset excluding word durations below 121 ms.

Table 7 summarizes the model when tokens below 121 milliseconds were excluded. The number of observations meeting that criterion was 54590. All predictors in the model are still significant, in the expected direction.

Variable	beta	SE	t
Intercept	0.0134904	0.0149837	0.900
cLogLBigram	-0.0034344	0.0007527	-4.563
cLogLRate	-0.1189210	0.0244233	-4.869
cLogRRate	-0.0747063	0.0269802	-2.769
cLogBerndt	-0.1214651	0.0347610	-3.494
cLogNQuot	0.1951226	0.0154983	12.590
clen	0.0504931	0.0056772	8.894
cAge	0.0020900	0.0003759	5.561
sexMALE	-0.0503839	0.0081858	-6.155
cLogSwbdFq	-0.0253404	0.0027947	-9.067

Table 7: Summary of the mixed-effects model of word duration, including word frequency. Tokens with durations below 121 milliseconds, as well as tokens immediately preceding pauses or disfluencies were excluded. The model has random intercepts for speaker (s = 0.08) and phonemic content (i.e. a grouping variable that groups together all homophonous tokens, e.g. all tokens of *thyme* and *time* as one group; the standard deviation for that random effect was 0.13); the standard deviation of the residual was 0.30.

Figure 5 shows the residuals for that model, side by side with the residuals when all tokens are included. The residuals are definitely much improved in the lower quartiles. For the higher quartiles, normality of the residuals does not improve. This may be due to one or more of the predictors having a non-linear relationship to word durations, or it may be due to the distribution of the outcome itself.

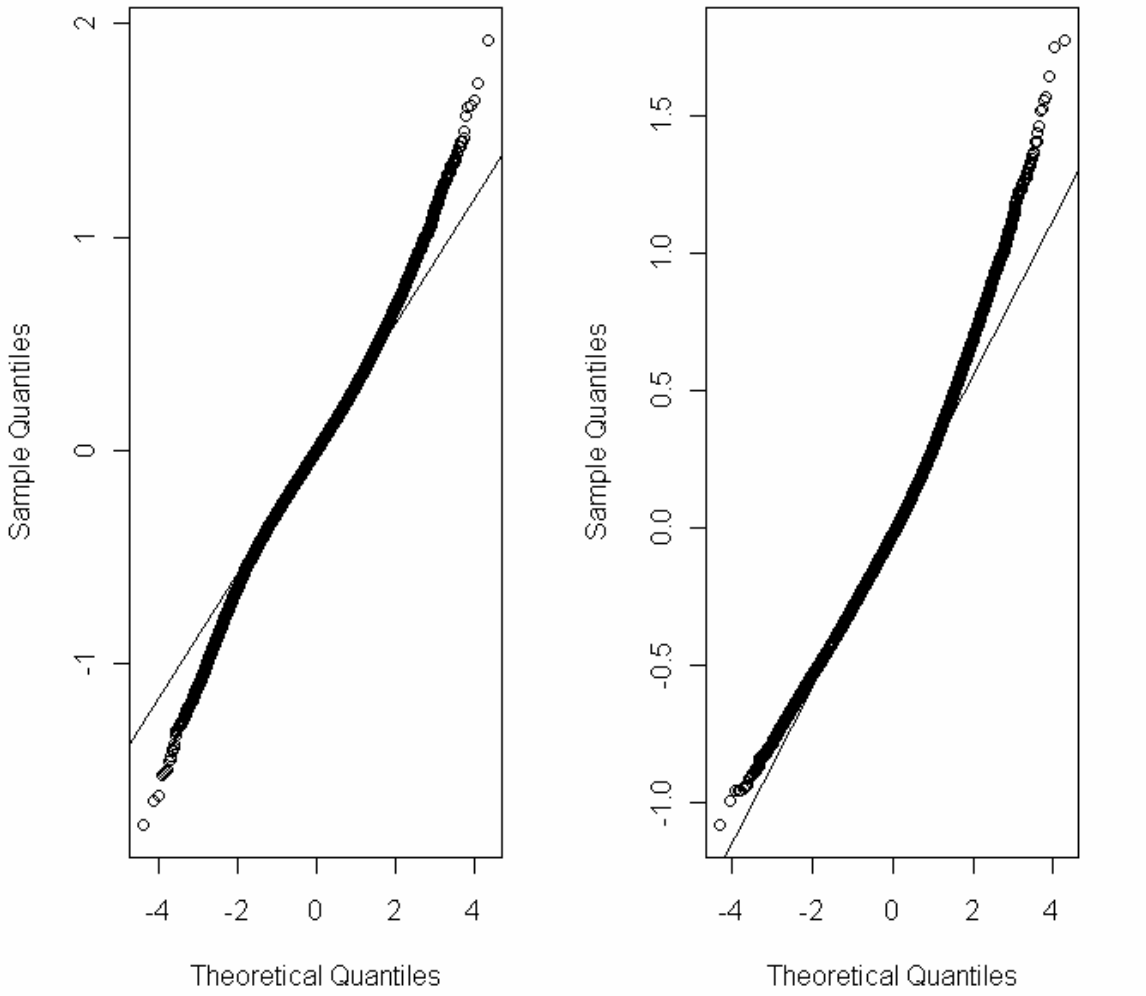


Figure 5: QQ-plots of the residuals of frequency-based models of all data points (left panel) and the subset of the data excluding token durations below 120 milliseconds and tokens immediately preceding pauses or disfluencies.

How successful are the models at explaining variability in word durations? And is the model's success at capturing variability just due to the random adjustments, or does adding the fixed effects improve the model's ability to capture variability? This second question merits special attention: One way to evaluate models is to ask what the relative contribution of random and fixed effects is: If model fit is achieved primarily via random adjustments, then this may cast doubt on the importance of the fixed effects. On the other hand, one of the random effects in the current model is based on "pronunciation", i.e. phonemic content. One would expect phonemic content to be an excellent predictor of word durations, even absent any information about the fixed effects. Therefore, the question isn't so much how much variability in the data can be captured by the fixed

effects that cannot be captured via random adjustments ("Can the random effects do it all?"); the question is, rather, how much the variance of the random adjustments changes depending on whether the fixed effects are in the model ("Can the fixed effects keep the random effects from doing all the work?").

	All tokens
Variance accounted for:	homoBaseline
By fixed and random effects	0.47 (baseline) 0.46 (including frequency)
By callerId only	0.07
By random effects only	0.30
…plus fixed effects for pauses and disfl	0.46

Table 8: Amount of variability in word duration captured by models with and without fixed effects.

Table 8 shows the amount of variability explained by models with and without fixed effects. that the full model accounts for 46% of the variance in word durations. A good part of that variance is accounted for just by the random effects: 30/47 = 66.0%. Adding the fixed effects for pauses and disfluencies means that 46 % of the variance is accounted for, leaving practically nothing[1] (0.0044) to explain for the linguistically more interesting variables. Clearly, the importance of the fixed effects does not lie in improving the overall model fit.

The question is: How much do the random adjustments change when fixed effects are added? Table 9 shows the random adjustments in a model without fixed effects (durations above 120 milliseconds, excluding pauses and disfluencies), compared to the adjustments in the full model. The variances for callerId and for pronunciation are indeed smaller for the model that includes fixed effects: The variance for the adjustment by Speaker ("callerId") goes down by 17% ((0.008028 - 0.0066523) / 0.00828). The variance for the adjustment by Pronunciation goes down by 53% ((0.0184661 – 0.039007) / 0.039007). In answer to our question, then: These numbers suggest that the fixed effects are capable of keeping the random adjustments from doing all the work.

[1] > cor(fitted(homo.lmer), d$cLogDur)^2 - cor(fitted(homoFixedPauseDisfl.lmer), d$cLogDur)^2
[1] 0.00441292

Without fixed effects:			With fixed effects:		
Random effects:			Random effects:		
Groups Name	Variance	Std.Dev.	Groups Name	Variance	Std.Dev.
callerId (Intercept)	0.008028	0.089599	callerId (Intercept)	0.0066523	0.081562
pron (Intercept)	0.039007	0.197503	pron (Intercept)	0.0184661	0.135890
Residual	0.092682	0.304437	Residual	0.0905292	0.300881

Table 9: Variance and standard deviations of random effects in models with and without fixed effects

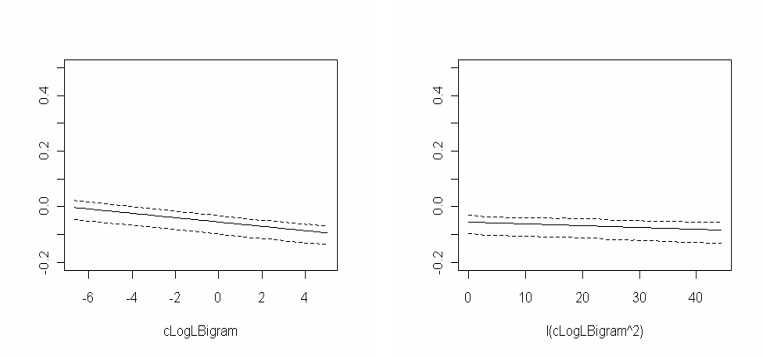
The models discussed so far all assume linear relationships between predictors and outcome. The partial regression plots suggest that that assumption is reasonable; yet, previous studies of word durations have noted non-linear relationships between bigram probability and word duration (Bell, Brenier, Gregory, Girand, & Jurafsky, 2009). I therefore fitted a model taking into account a quadratic effect of bigram probabilities. Table 10 summarizes the model.

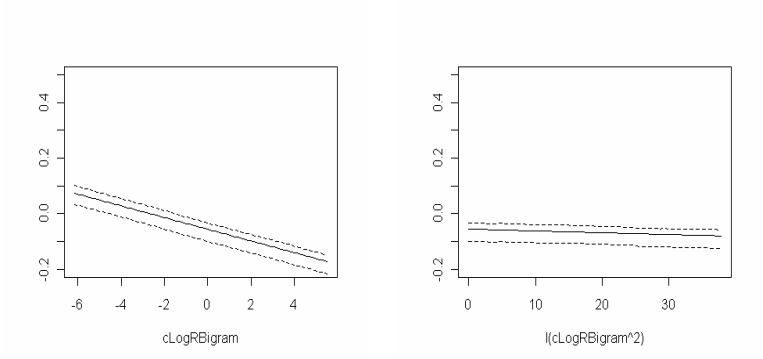
Variable	beta	SE	t
Intercept	0.0328851	0.0154164	2.13
cLogLBigram	-0.0078023	0.0006626	-11.78
cLogLBigram^2	-0.0006452	0.0002149	-3.00
cLogRBigram	-0.0210377	0.0006964	-30.21
cLogRBigram^2	-0.0006479	0.0002156	-3.01
cLogLRate	-0.1072044	0.0231190	-4.64
cLogRRate	-0.1404621	0.0242616	-5.79
cLogBerndt	-0.1253817	0.0337309	-3.72
clen	0.0460915	0.0054989	8.38
cAge	0.0016649	0.0004150	4.01
sexMALE	-0.0790579	0.0090449	-8.74
prePausalTRUE	0.4074274	0.0045930	88.71
preDisflTRUE	0.4120113	0.0030761	133.94
cLogSwbdFq	-0.0165871	0.0027480	-6.04

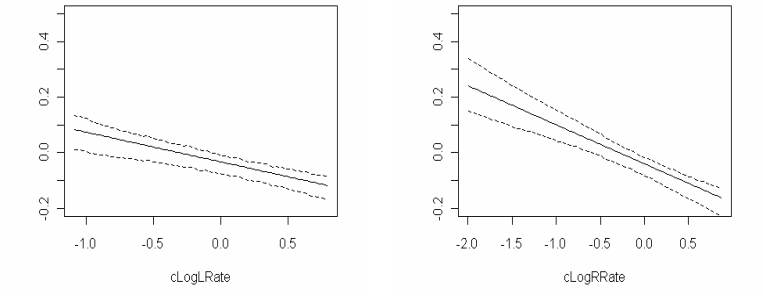
Table 10: Summary of the mixed-effects model of word duration including a quadratic term for local bigram measures. Like the previous models, the model has random intercepts for speaker (s = 0.095) and phonemic content (s = 0.14); the standard deviation

of the residual was 0.32. Total variance accounted for by random and fixed effects was .46.

Including the quadratic terms results in a slight improvement, inasmuch as the effect of local speaking rate emerges as significant, as one would expect it to be. All other coefficients are significant, in the expected direction: Higher bigram probabilities, faster speaking rate, greater orthographic regularity, and higher word frequency are associated with shorter word duration. Male speakers tended to produce shorter word durations, other things beign equal. Greater length in letters, higher speaker age, and neighboring pauses and disfluencies are associated with longer word durations.







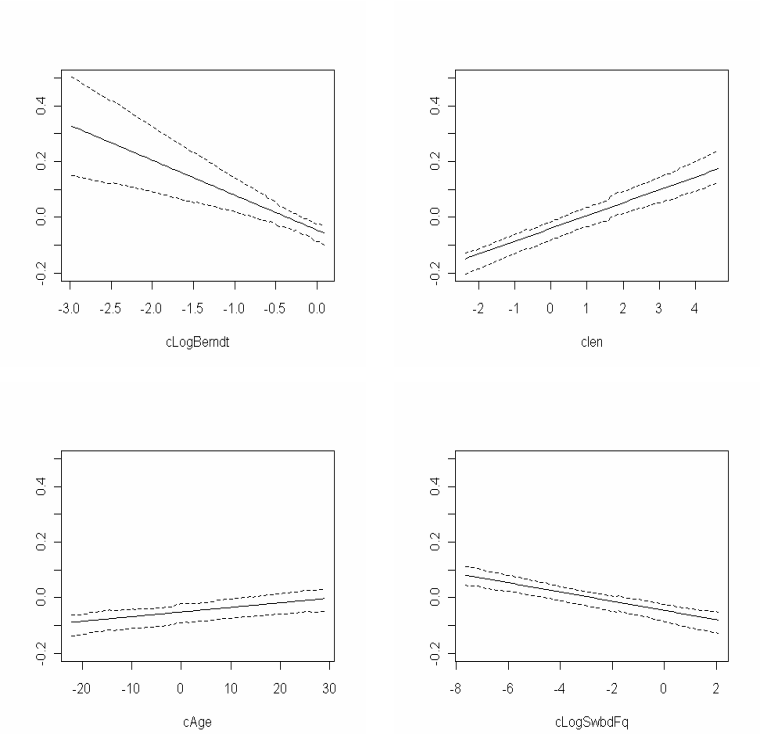


Figure 6: Partial residuals for the model including quadratic terms for bigram probability, given the target token's immediate left and right neighbors. Values on the y-axis represent predicted

The current model does not include a quadratic term for the effect of speaking rate on word durations, even though previous studies have reported quadratic relationships between speaking rate and word duration (Jaeger, p.c.), and between speaking rate and error rates in automatic speech recognition systems (Goldwater, Jurafsky, & Manning, 2008). I tested this, but found speaking rate preceding the target, as well as the quadratic term for the speaking rate in the region following the target to be non-significant in the resulting model. Including the quadratic terms for speaking rate also did not result in improved residuals or amount of variance accounted for. The impression I get is that the "quadratic" effect really just separates the tokens with the lowest speaking rates (which tend to be the ones near disfluencies) from all other tokens – so looking at the fluent tokens separately, or having disfluencies in the model renders the quadratic terms for speaking rate superfluous.

Conclusion

The purpose of the current study was to ascertain whether the effect of lemma frequency on word durations reported in Gahl (2008) would hold up to scrutiny. It did.

Baayen, H. (2008). *Analyzing linguistic data: A practical introduction to Statistics using R*. Cambridge: Cambridge University Press.

Baayen, R. H., Piepenbrock, R., & van Rijn, H. (1993). The CELEX Lexical Database (CD-ROM). : Linguistic Data Consortium, University of Pennsylvania, Philadelphia (PA).

Bell, A., Brenier, J., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language, 60*(1), 92-111.

Berndt, R. S., Reggia, J. A., & Mitchum, C. C. (1987). Empirically derived probabilities for grapheme-to-phoneme correspondences in English. *Behavior Research Methods, Instruments & Computers, 19*(1), 1-9.

Deshmukh, N., Ganapathiraju, A., Gleeson, A., Hamaker, J., & Picone, J. (1998). *Resegmentation of Switchboard.* Paper presented at the International Conference on Spoken Language Processing, Sydney, Australia.

Gahl, S. (2008). "Time" and "thyme" are not homophones: Word durations in spontaneous speech. *Language, 84*(3), 474-496.

Godfrey, J., Holliman, E., & McDaniel, J. (1992). *Switchboard: Telephone speech corpus for research and development.* Paper presented at the International Conference on Acoustics, Speech, and Signal Processing.

Goldwater, S., Jurafsky, D., & Manning, C. D. (2008). Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase ASR error rates, *ACL-HLT*. Columbus, OH.

Harrell, F. E. (2008). Design: Design Package. R package version 2.1-2. http://biostat.mc.vanderbilt.edu/s/Design, http://biostat.mc.vanderbilt.edu/rms (Version R package version 2.1-1).

R Development Core Team. (2008). R: A language and environment fo statistical computing. Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

[i] This material is based in part upon work supported by the National Science Foundation under Grant Number IIS-0624345 to Stanford University for the research project 'The Dynamics of Probabilistic Grammar"' (PI Joan Bresnan). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.