To appear in *Language* (Commentary Section), in the June or the September issue of 2018

Time and *Thyme* are NOT homophones: a closer look at Gahl's work on the lemma frequency effect including a reanalysis

ARNE LOHMANN

Heinrich-Heine-Universität Düsseldorf

ABSTRACT

The article '*Time* and *Thyme* are not homophones' (Gahl 2008), published in *Language*, reports a frequency effect differentiating the durations of homophones, e.g. *time* vs. *thyme*. The article is of fundamental theoretical relevance, as the finding reported has significant implications for research on homophones and effects of frequency in general. However, as I show in the present paper, the main analysis in Gahl 2008 does not provide quantitative evidence for the effect. The same is true of a follow-up study (Gahl 2009). The present paper provides a reanalysis based on the original dataset, which shows that the frequency effect reported in the original paper is real.[*]

Keywords: homophones, frequency effects, frequency inheritance, regression modeling, acoustic phonetics, speech production

1. INTRODUCTION. In a much-noticed paper, Gahl (2008) puts forth the claim that lemma frequency impacts the duration of homophones, based on a corpus-analysis of spontaneous speech. The main finding reported is that 'high-frequency words like *time* are significantly shorter than their low-frequency homophones like *thyme'* (Gahl 2008: 474). In a follow-up study based on the same dataset, but applying a more sophisticated modeling approach, the claim is repeated (Gahl 2009). This possible result is of great theoretical relevance, as it has major implications for the representation of homophonous words in the mental lexicon and in models of speech production. With 252 citations (according to a google scholar search on 1/8/2018), the paper by Gahl (2008) can be considered one of the cornerstone articles in the literature on frequency effects. Interestingly,

[*]I thank Susanne Gahl for sharing her Switchboard homophone data with me. I am furthermore grateful to Gero Kunter and Filip Nenadic for discussing several aspects of the analyses depicted with me. Thanks also goes to Melanie Bell, Sonia Ben Hedia, Matthew Kelley, Ingo Plag and Benjamin Tucker for providing helpful comments on earlier versions of this article. I furthermore want to thank Andries Coetzee and Roger Levy at *Language* for valuable comments. Roger Levy's suggestions in particular resulted in important improvements of the quantitative analyses provided. All remaining errors are mine. Funding for this study by the *Deutsche Forschungsgemeinschaft* is gratefully acknowledged (grant LO-2135/1-1).

earlier research had found the effect to be elusive, with as many negative as positive findings reported in the literature (see Gahl 2008 for an overview).

In the following, I show that the main statistical models reported in both Gahl 2008 and Gahl 2009 do not provide evidence for the claim put forth. This is due to methodological issues that result in the crucial hypothesis of differences in duration between homophones not being directly tested.

The present paper critically analyzes and discusses the statistical models reported in the two papers and provides a refined reanalysis of Gahl's data. In addition, the same calculation is carried out on a separate dataset of homographic noun-verb homophones (slightly adapting the analysis originally published in Lohmann 2017). The result of both analyses is that the lemma frequency effect is real, so that the main theoretical claims following from Gahl's work still stand.

2. BACKGROUND: HYPOTHESES ABOUT FREQUENCY EFFECTS ON WORD DURATION. In order to point out the relevant methodological aspects of the analyses presented in Gahl 2008, 2009, it is helpful to first distinguish between different hypotheses about effects of frequency on word duration. The general hypothesis that underlies the studies is that frequency has a reductive effect on duration. Based upon this general hypothesis, two relevant sub-hypotheses on homophonous and non-homophonous words can be distinguished.

Hypothesis 1 (H1): All other things being equal, two words that are not phonological homophones differ in duration depending on their difference in token frequency, with the more frequent word being pronounced with shorter duration, e.g. duration of *time* < duration of *sage*.

Hypothesis 2 (H2): All other things being equal, two words that are phonological homophones differ in duration depending on their difference in frequency, with the more frequent word being pronounced with shorter duration, e.g. duration of *time* < duration of *thyme*.

H1 and H2 are two specifications of the general frequency-shortening hypothesis, with H1 having only non-homophonous words within its scope and H2 being only about homophonous words. A number of articles have provided evidence for H1 by showing that token frequency co-determines the acoustic realization of non-homophonous words (see Wright 1979 for an early study and Jurafsky 2003 and Gahl 2008 for overviews). To the extent of my knowledge this hypothesis is largely uncontroversial (but see Baayen et al. 2016 for a skeptical opinion on the role of frequency

in speech processing in general). Most importantly, H1 is not the hypothesis the articles by Gahl (2008, 2009) are about. It is listed here merely for expository reasons, as it will be referred to below during the discussion of the empirical analyses.

In the current context, the more relevant hypothesis is H2, as it is this hypothesis that the studies by Gahl (2008, 2009) aim to test. It states that the general relationship between frequency and reduction holds also for homophonous words. This hypothesis is more controversial than H1 and arguably of greater theoretical relevance, because in previous psycholinguistic research it has been found that the token frequency of homophonous word pairs is inherited from one member of the pair to the other. The phenomenon of so-called 'frequency inheritance' has been shown to produce a number of empirical effects in which the low-frequency member of a homophone pair behaves like its high-frequency twin, due to inheritance of its high frequency (see Middleton et al. 2015 for an overview). This inheritance effect can be taken to predict that the a low-frequency word with a high-frequency homophone will be pronounced with the same short duration as the high-frequency word. Hypothesis H2 and frequency inheritance are thus in direct conflict, with positive evidence for H2 constituting evidence against a frequency inheritance effect on word duration.

Since the main aim of Gahl (2008, 2009) is to provide evidence for H2, it will be discussed in the following to what extent the statistical models reported in these papers achieve that aim.

3. A CLOSER LOOK AT GAHL 2008. The empirical analysis reported in Gahl 2008 is based on 220 heterographic homophone pairs extracted from the Switchboard corpus of telephone conversations, which are instantiated by approximately 80,000 tokens (Gahl 2008: 480). A global comparison of mean durations shows that, on average, the low-frequency homophones are of greater duration than the high-frequency homophones. However, as Gahl (2008: 481) points out, this comparison does not consider possibly confounding variables, which is why she calculates a multifactorial regression model predicting word duration. This model is the crucial piece of evidence for the lemma frequency effect that is put forth in Gahl 2008. The main issue with this quantitative model is that it does not test for duration differences between the homophones being related to differences in frequency, as stated in H2 above. Therefore, although the paper states otherwise, no conclusive evidence for high-frequency words being shorter in duration than their low-frequency homophones is provided.

In the following, the model presented in Gahl (2008) will be analyzed in greater detail. The dependent variable is the average duration of only the high-frequency members of the individual homophone pairs. While the durations of the low-frequency homophones are thus not predicted, their average durations serve as one predictor of the duration of their high-frequency twins. The general logic of the model is to show that the average duration of the low-frequency member predicts the duration of its homophone counterpart only imperfectly, and, most importantly, that token frequency is a statistically significant predictor that fills this gap. The model output is reproduced in Table 1 below (using more transparent variable names than in the original article).

Fixed-effect predictors	Coefficient	Std.Error	*t*
Intercept	-0.5247	0.104	-5.07
Speaking rate	-0.0492	0.020	-2.42
Duration of low-frequency homophone	0.2141	0.040	5.42
Bigram probability	-0.0171	0.005	-3.21
M-score (grapheme-phoneme probability)	-0.2213	0.073	-3.02
Noun proportion	0.1034	0.024	4.29
Pause ratio	0.2813	0.137	2.06
Logged token frequency of word	-0.0297	0.001	-4.43

TABLE 1. Summary of the regression model fitted to log-transformed average durations of the high-frequency homophones in the dataset (see Gahl 2008: 486)

In order to show that a possible difference in duration between the homophones is not due to other factors impinging on the duration of the words in the sample, Gahl (2008) employs a number of control variables that enter the model (see Table 1). For a detailed overview and discussion of these co-variates, see Gahl 2008. It is important to note that the control variables capture only properties of the high-frequency but not of the low-frequency homophones, because the model is fitted to the average durations of only the high-frequency words (see also Gahl 2009: 279 on this point). Consequently, these variables control only for differences between the high-frequency words, but not for differences between the actual homophones. As a result, the control variables do not achieve their primary aim, because differences between the homophones are the main point of the study.[1]

Moving on to the main variables of interest, the crucial predictor in the model is the logged frequency of the high-frequency words (bottom row in in Table 1 above). Gahl interprets the significance of this predictor[2] in the presence of the control variables as evidence for H2 (see Gahl 2008: 486). However, this conclusion is not warranted, as this predictor contains only the frequencies of the high-frequency homophones in the sample, for example, *time* (7,312), *night* (1,020), and *blue* (161), all of which are high-frequency members of their respective pairs. As a

consequence, the significance of this predictor indicates that high-frequency words with a higher frequency have a shorter duration than high-frequency words with a lower frequency, e.g. *time* has a shorter duration than *blue* or *night*. This is evidence for H1, i.e. for frequency effects impacting the durations of non-homophonous words. However, it is not evidence for H2, as information about the frequency of the low-frequency member, or about the difference in frequency between the homophones is not included in the model.

A further predictor in the model that is relevant for the discussion of H2 is the average duration of the low-frequency homophone. Gahl (2008: 486) interprets the less than perfect correlation of this predictor with the dependent variable as evidence that the word pairs are not homophones: 'A striking aspect of the model is the small contribution of homophone duration as a predictor of word duration. Homophones are usually defined as sets of words that sound alike. Given that definition, one would expect the duration of a word like *thyme* to predict the duration of its twin *time* perfectly.' However, there is another explanation for the small contribution of this predictor, which has to do with the composition of the sample employed. Like the dependent variable, the predictor capturing the duration of the low-frequency member is based on average durations. Now obviously, in the sample drawn from the Switchboard corpus, the low-frequency words are far outnumbered by their high-frequency twins. In fact, the median frequency of the low-frequency homophones in the sample is only 4 (Gahl 2008: 480), which means that 50% of the low-frequency words are instantiated by fewer than 4 tokens. A case in point is *thyme*, which occurs exactly once in the corpus.[3] Assuming that token durations of a given word type are quite variable in spontaneous speech, employing a very low number of tokens will result in a fairly unreliable duration average for a word. Therefore, the less than perfect predictive performance of this predictor is not necessarily due to *thyme* and *time* and other pairs not being homophones, but may simply be due to sampling issues.

In conclusion, the model presented does not provide evidence for low-frequency homophones being systematically pronounced with greater duration than their high-frequency twins, because the frequency predictor employed in the model tests only for differences between the high-frequency words, and not between the homophones. In other words, H2 is not explicitly tested. Only evidence for H1 is provided, which is explicitly NOT the topic of the paper. Employing the average duration of the low-frequency word as a further predictor does not provide

unambiguous evidence for the non-homophone status of the target words either, as this predictor is based on critically low numbers of tokens, undermining its reliability.

4. A CLOSER LOOK AT GAHL 2009. The recalculation reported in Gahl 2009 is based on the same sample, but employs the more sophisticated modeling technique of mixed-effects modeling. Moreover, the new model is based on tokens and is fitted to the word durations of both the high-frequency and the low-frequency members of the homophone pairs. This alone, as Gahl (2009: 279) points out, avoids certain obvious drawbacks of the model that predicts only the average durations of the high-frequency words, discussed in the previous section. However, there are still significant shortcomings, which result in this recalculation still not providing conclusive evidence for H2.

The model is fitted to the word duration of all homophones in the sample (N= 79,867). Gahl includes a large number of control variables as fixed effects, most of which are the same as in the 2008 article (for a complete list and explanation of all predictors, see Gahl 2009: 282-283). Since the sample contains the high-frequency and the low-frequency tokens, these variables now control for possible differences between the homophones that may affect their duration, in contrast to the model reported in Gahl 2008. The random effects structure includes random intercepts for speaker and phonemic content, which is "a grouping variable that groups together all homophonous tokens, e.g. all tokens of *thyme* and *time* as one group" (Gahl 2009: 285). Partial results for the fixed-effect predictors are shown in Table 2, reproduced from Gahl 2009: 286.

Fixed-effect predictors	Coefficient	Std.Error	*t*
Intercept	0.012	0.014	0.89
Speech rate (left context)	-0.151	0.023	-6.56
Speech rate (right context)	-0.048	0.025	-1.96
Length in letters	0.053	0.005	10.23
Right bigram probability	-0.021	0.001	-30.31
Left bigram probability	-0.007	0.001	-12.06
Age of speaker	0.002	0.001	4.06
Sex of speaker (male)	-0.079	0.009	-8.80
Target word precedes a disfluency	0.411	0.003	133.28
Target word precedes a pause	0.406	0.005	88.12
Orthographic regularity	-0.142	0.031	-4.55
Noun proportion	0.174	0.014	12.21
Token frequency of word in Switchboard	-0.011	0.003	-4.27

TABLE 2. Summary of the fixed-effects predictors of the model fitted to the log-transformed word durations as reported in Gahl 2009

The crucial variable in the model is the log-transformed frequency of the word ('Token frequency of word in Switchboard' in Table 2). This variable contains frequency values for each word in the sample, i.e. not only for the high-frequency words but also for their low-frequency twins. When entering this predictor into the model along with the control variables, it significantly improves model fit, based on an ANOVA calculation ($p<0.001$, see Gahl 2009: 287). Gahl interprets this result as evidence for H2.

At first glance this seems to be a correct interpretation, as both members of each pair are now part of the sample analyzed and have different frequencies. However, the reasons for this predictor being significant are not necessarily due to differences in frequency between the homophones within the word pairs, because the frequency predictor tests for effects of frequency in a global fashion, i.e. between all words in the sample, not specifically between the homophones. As a consequence the significant result of this predictor may be due to certain frequent words in the sample having a shorter duration than certain infrequent ones, irrespective of whether they belong to the same homophone pairs.

In fact, the result may also come about if all homophones were pronounced the same and there was simply an effect of more frequent pairs being pronounced with shorter duration than less frequent pairs. In order to demonstrate that, I created a new variable that captures the predictions of a frequency inheritance account, namely that all homophones are affected by the same cumulative frequency. For this new variable I added up the frequency counts of both members of each homophone pair, so e.g. *time/thyme* now have the same cumulative frequency, which of course is very close to the frequency of *time*. I term this new variable *pair frequency*, as frequency differs only between homophone pairs but not between individual words or lemmas. The correlation between *pair frequency* and the original frequency predictor is extremely high in the sample ($r=0.98$, based on word frequencies obtained from Switchboard). No matter which of the two is used as a predictor, the model returns a significant result. The extreme correlation is the result of the cumulative frequency count being very close to the frequency of the respective high-frequency homophone and the frequency imbalance in the data already discussed in Section 3, with the high-frequency words contributing many more tokens than the low-frequency words. The important point of this correlation is that the frequency predictor employed in the model in Gahl 2009 does not test the predictions of H2, as it is driven to a large extent by frequency differences between pairs, but not by frequency differences within pairs.

The upshot of this discussion is that the model reported in Gahl 2009 does not provide conclusive evidence for a duration difference between the homophones, because the significant result of the frequency predictor could simply be an effect of frequent word pairs being pronounced with shorter duration than infrequent ones, but with no actual difference in duration between the homophones in a pair.

5. A REANALYSIS. The problem with the models presented in Gahl 2008 and 2009 is that the frequency predictor employed in them either tests for effects of frequency differences between only the high-frequency words (in Gahl 2008), or between all words in the sample but not specifically between the homophones (in Gahl 2009). Consequently, as pointed out above, no compelling evidence for the crucial hypothesis H2 is provided. The discussion in Section 4 has shown that, in order to obtain this kind of evidence, it is necessary to disentangle possible frequency effects at work between different homophone pairs from the frequency effect differentiating the homophones in any given pair. In the following, I present a reanalysis of Gahl's dataset that does exactly that. In addition, I will run the same analysis on a separate dataset of noun-verb homophones in English, e.g. *face*(V) vs. *face*(N). The latter analysis is based on the study reported in Lohmann 2017 and employs data from the Buckeye corpus (Pitt et al. 2007); see Lohmann 2017 for details.

For the sake of exposition, the reanalysis will be conducted in several steps. In a first step I calculate mixed-effects models predicting word duration that contain all control variables (for explanations of the individual variables and their operationalization see Gahl 2009 and Lohmann 2017) and as a frequency-related predictor, the logged mean frequency of both members of each homophone pair.[4] That is, for a given pair, e.g. *time* and *thyme*, or *face*(V) and *face*(N), one common frequency is assumed. As a consequence, the resultant models are ignorant as to frequency differences between the homophones, but consider frequency differences between pairs. The point of these models is to take into account known influences on word duration, with the only exception being the crucial frequency effect differentiating the homophones, which will be tested in a second step. One way to get at this effect is to investigate the residuals of the models, i.e. the difference between observed and fitted values. Since the models assume the same mean frequency value for both the low-frequency and the high-frequency member of the individual pairs, they should systematically overestimate the duration of the high-frequency words and underestimate the duration of their low-frequency counterparts, in case there is a systematic difference in duration as predicted by H2. Consequently, the residuals (observed minus fitted values) should be positive for

the low-frequency words and negative for the high-frequency words. In case there is no difference in duration as predicted by H2, there should be no systematic difference in the direction of the residuals between the low-frequency and the high-frequency words.

I fitted two models - one for Gahl's dataset and one for the dataset of N/V homophones - to the log-transformed (base 10) word duration in milliseconds using the *lmer* function of the *lme4* package (Bates et al. 2014) in R (R Core Team 2014). All scalar independent variables were log-transformed (base 10), centered, and standardized through dividing by two standard deviations (see Gelman & Hill 2007: 56). The models include random intercepts for speaker and phonemic content, i.e. homophone pair (see Section 4) and also a random intercept for the specific word (lemma). In following a design-driven rather than data-driven approach, no attempt was made to simplify the random effects structure (see Barr et al. 2013 on that point). *P*-values for the fixed effects were obtained by comparing the fit of the model with and without the additional fixed-effect predictors using likelihood-ratio tests as implemented in the *anova* function in R. Fixed-effect predictors that do not significantly contribute to model fit were kept in the model in case they display effects in the predicted direction (following established model fitting guidelines, as e.g. by Gelman & Hill 2007: 69). Since all tested variables fulfill this criterion, no fixed-effect predictor was removed. Multicollinearity was not an issue with either model, with Variance Inflation Factors <2. Tables 3 and 4 provide the random and fixed-effects summaries of the models calculated for the two datasets.[5]

Random effects	Variance	SD
Speaker (intercept), 520 groups	1.695E-3	4.117E-2
Phonemic content (intercept), 206 groups	1.685E-3	4.105E-2
Lemma (intercept), 412 groups	8.366E-4	2.892E-2
Residual	1.922E-2	1.387E-1

Fixed-effect predictors	Coefficient	Std. Error	*t*	*p*
Intercept	2.4100	0.00892	270.156	
Speech rate (left context)	-0.0216	0.01559	-1.386	=0.17
Speech rate (right context)	-0.0057	0.00847	-0.675	=0.49
Length in letters	0.0571	0.00644	8.878	<0.001
Right bigram probability	-0.0395	0.00134	-29.514	<0.001
Left bigram probability	-0.0150	0.00123	-12.170	<0.001
Age of speaker	0.0154	0.00385	4.013	<0.001
Sex of speaker (male)	-0.0346	0.00393	-8.811	<0.001
Target word precedes a disfluency	0.1783	0.00134	133.329	<0.001
Target word precedes a pause	0.1761	0.00200	88.117	<0.001
Orthographic regularity	-0.1590	0.00493	-3.223	<0.01
Noun proportion	0.0432	0.00574	7.530	<0.001

Mean frequency of pair	-0.0625	0.00808	-7.726	<0.001

TABLE 3. Random and fixed-effects summary of a regression model predicting the log-transformed word durations in the Gahl dataset (N=79,166), including mean frequency of the homophone pairs as a predictor

Random effects	Variance	SD
Speaker (intercept), 40 groups	2.281E-3	4.776E-2
Phonemic content (intercept), 63 groups	2.220E-3	4.712E-2
Lemma (intercept), 126 groups	1.224E-3	3.498E-2
Residual	1.212E-2	1.101E-1

Fixed-effect predictors	Coefficient	Std. Error	*t*	*p*
Intercept	2.6050	0.07141	36.475	
Speech rate	-0.0097	0.00450	-2.146	<0.05
Length in segments on CV-tier	0.0979	0.01324	7.394	<0.001
Right bigram probability	-0.0226	0.00475	-4.760	<0.001
Left bigram probability	-0.0067	0.00438	-1.527	=0.13
Position (Phrase-final)	0.0352	0.00750	4.690	<0.001
Position (Clause-final)	0.1034	0.00560	18.450	<0.001
Pitch range	0.0669	0.00434	15.427	<0.001
Target word precedes a pause	0.0920	0.00926	9.938	<0.001
Mean frequency of pair	-0.0347	0.01523	-2.277	<0.05

TABLE 4. Random and fixed-effects summary of a regression model predicting the log-transformed word durations in the N/V dataset (N=3,435), including mean frequency of the homophone pairs as a predictor

The bottom rows in Table 3 and 4 illustrate the effect of mean frequency of the homophone pairs, which is statistically significant in both datasets. Its negative coefficient indicates the expected reductive effect of frequency, with pairs of higher frequency being pronounced with shorter duration.

In order to test the predictions of H2, I aggregated the residuals of the models for the low-frequency and the high-frequency groups in both datasets. This was done by first calculating the by-lemma mean residuals and based on that the mean residuals for the two frequency groups.

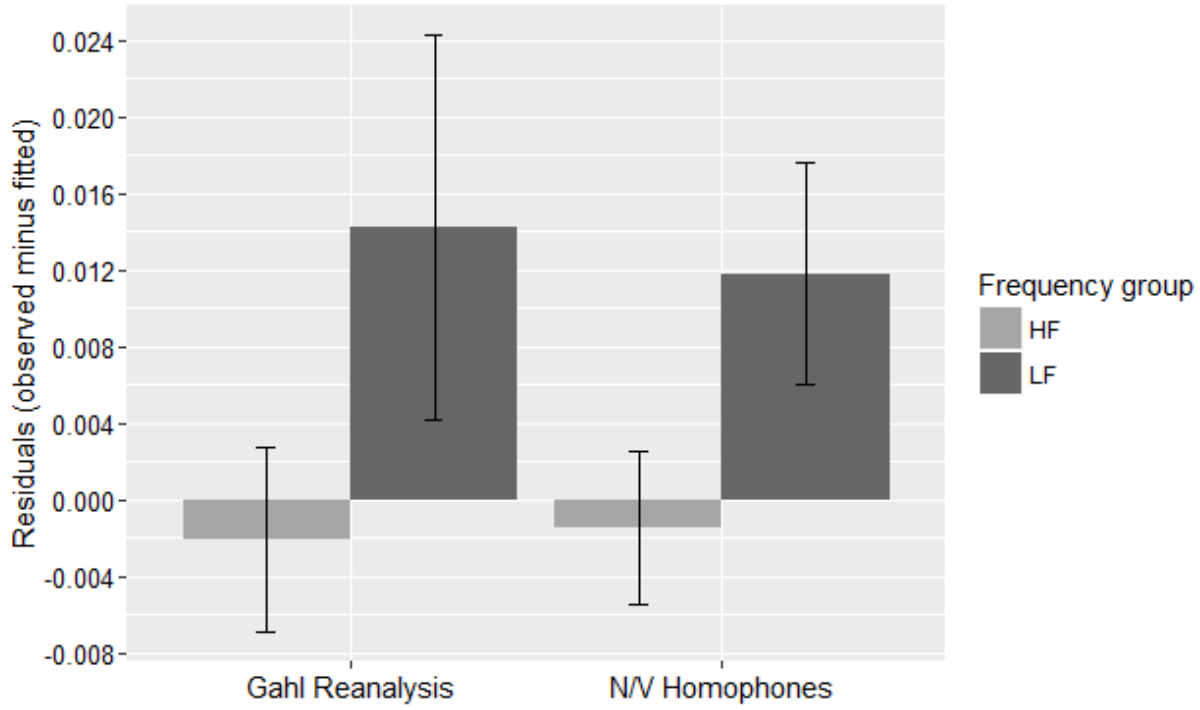

FIGURE 1. Mean residuals by low-frequency (LF) and high-frequency (HF) homophones (the error bars indicate the 95% confidence intervals)

As shown in Figure 1, the residuals are on average negative for the high-frequency group, but positive for the low-frequency group, indicating that the models do indeed underestimate the durations of the low-frequency tokens and overestimate the durations of the high-frequency tokens, as predicted by H2. Paired *t*-tests show the difference in mean residuals to be significant in both datasets (Gahl reanalysis: $t = 2.80$, df = 205, $p < 0.01$; N/V homophones: $t = 3.24$, df = 62, $p < 0.01$). The reason for the residuals of the high-frequency group being closer to zero for both models is the aforementioned imbalance in the datasets, with more high-frequency than low-frequency tokens. The models try to reduce the residuals overall, and this is achieved by optimizing predictions for the much larger group of high-frequency words.

The predictions of H2 can also be directly tested using a predictor that captures the difference in frequency between the homophones, which is then added to the models reported in Tables 3 and 4. One way to do this is to create a binary predictor that indicates whether a word is the low-frequency or the high-frequency word relative to its homophone twin. Alternatively, the relative frequency of the homophones can also be expressed on a scale: for each pair of homophones I calculated a logged frequency ratio arrived at through dividing the frequency of the

word in question by the frequency of its homophone twin. This results in a negative value for the low-frequency word and a corresponding positive value for the high-frequency word, with its size indicating the size of the difference in frequency. See the example below for the pair *fair/fare*, which have the following frequencies in SUBTLEX-US, *fair* = 314, *fare* = 4,832.

Logged frequency ratio (*fair*) = log10(4,832 / 314) = 1.187
Logged frequency ratio (*fare*) = log10(314 / 4,832) = -1.187

I tested both the binary and the scalar predictor by separately adding them to the models presented in Tables 3 and 4 along with random slopes by speaker and phonemic content for the additional fixed-effect predictor.

Crucially, the additional predictors – both the binary and the scalar variant – yield statistically significant results in the respective models (at $\alpha = 0.05$), showing that there is in fact a statistically significant duration difference in the expected direction, with the high-frequency homophones being pronounced with shorter duration. Tables 5 and 6 show the complete output for models including the binary predictor.

Random effects	Variance	SD
Speaker (intercept), 520 groups	1.713E-3	4.139E-2
Speaker (Frequency, relative)	5.240E-5	7.239E-3
Phonemic content (intercept), 206 groups	2.646E-3	5.144E-2
Phonemic content (Frequency, relative)	1.447E-3	3.804E-2
Lemma (intercept), 412 groups	1.480E-5	3.847E-3
Residual	1.922E-2	1.386E-1

Fixed-effect predictors	Coefficient	Std. Error	*t*	*p*
Intercept	2.4050	0.00910	264.337	
Speech rate (left context)	-0.0252	0.01549	-1.625	=0.10
Speech rate (right context)	-0.0049	0.00837	-0.587	=0.54
Length in letters	0.0547	0.00639	8.556	<0.001
Right bigram probability	-0.0394	0.00134	-29.396	<0.001
Left bigram probability	-0.0148	0.00123	-12.019	<0.001
Age of speaker	0.0160	0.00383	4.163	<0.001
Sex of speaker (male)	-0.0337	0.00391	-8.628	<0.001
Target word precedes a disfluency	0.1783	0.00134	133.339	<0.001
Target word precedes a pause	0.1761	0.00200	88.127	<0.001
Orthographic regularity	-0.1592	0.00493	-3.231	<0.01
Noun proportion	0.0423	0.00567	7.464	<0.001
Mean frequency of pair	-0.0641	0.00801	-7.952	<0.001
Frequency relative to homophone twin (low-frequency)	0.0145	0.00493	2.941	<0.01

TABLE 5. Random and fixed-effects summary of a regression model predicting the log-transformed word durations in the Gahl dataset (N=79,166), including a binary predictor capturing the difference in frequency between the homophones

Random effects	Variance	SD	
Speaker (intercept), 40 groups	2.155E-03		4.642E-02
Speaker (Frequency, relative)	4.989E-05		7.063E-03
Phonemic content (intercept), 63 groups	3.877E-03		6.226E-02
Phonemic content (Frequency, relative)	1.718E-03		4.145E-02
Lemma (intercept), 126 groups	1.680E-17		4.099E-09
Residual	1.211E-02		1.100E-01

Fixed-effect predictors	Coefficient	Std. Error	*t*	*p*
Intercept	2.6080	0.07093	37.049	
Speech rate	-0.0097	0.00450	-2.145	<0.05
Length in segments on CV-tier	0.1002	0.01311	7.642	<0.001
Right bigram probability	-0.0226	0.00474	-4.757	<0.001
Left bigram probability	-0.0071	0.00435	-1.624	=0.10
Position (Phrase-final)	0.0322	0.00749	4.292	<0.001
Position (Clause-final)	0.1008	0.00559	18.026	<0.001
Pitch range	0.0668	0.00433	15.411	<0.001
Target word precedes a pause	0.0928	0.00924	10.041	<0.001
Mean frequency of pair	-0.0377	0.01500	-2.510	<0.05
Frequency relative to homophone twin (low-frequency)	0.0270	0.00781	3.451	<0.01

TABLE 6. Random and fixed-effects summary of a regression model predicting the log-transformed word durations in the N/V dataset (N=3,435), including a binary predictor capturing the difference in frequency between the homophones

The bottom rows in Table 5 and 6 show a statistically significant effect of the additional predictor, its positive coefficient indicating a greater duration of low-frequency homophones, as predicted by H2.[6] A further noteworthy result of the models is that the variance accounted for by the random intercept for lemma, i.e. specific homophone, is reduced considerably compared to the models reported in Tables 3 and 4. This indicates that much of the unexplained variance on the level of the individual lemma is indeed due to differences in frequency between the homophones. In order to estimate the size of the frequency effect, I fitted the models reported in Tables 5 and 6 also to the untransformed word durations, which indicate a difference in duration of 15.1ms and 21.2ms on average between the homophone members of the individual pairs in the Gahl dataset and the N/V dataset respectively.[7]

6. SUMMARY AND CONCLUSION. The present paper demonstrates that the main analyses in Gahl 2008, 2009, although widely cited as doing so, do not provide explicit evidence for a difference in duration between homophones that is contingent on differences in frequency. A reanalysis of the data collected by Gahl (2008, 2009), however, does provide evidence for this effect. In conclusion, the theoretical implications with regard to the lexical representation of homophones, extensively discussed in Gahl 2008 and numerous papers citing it, still stand. Furthermore, the same reanalysis is carried out on a sample of homographic noun-verb conversion homophones, which provides further evidence for the lemma frequency effect, a finding discussed in detail in Lohmann 2017. The latter result suggests that the frequency effect on homophone duration is not contingent on differences in spelling.

REFERENCES

BAAYEN, R. HARALD. 2010. Demythologizing the word frequency effect: A discriminative learning perspective. *The Mental Lexicon* 5, 149–157.

BARR, DALE; ROGER LEVY; CHRISTOPH SCHEEPERS; and HARRY J. TILY. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68. 255–278.

BATES, DOUGLAS; MARTIN MAECHLER; BEN BOLKER; and STEVEN WALKER. 2014. lme4: Linear mixed-effects models using Eigen and S4. Online: http://CRAN.R-project.org/package=lme4.

BRYSBAERT, MARC, and BORIS NEW. 2009. Moving beyond Kucera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods.* 41. 977–990.

GAHL, SUSANNE. 2008. *Time* and *Thyme* are not homophones: The effect of lemma frequency on word durations in spontaneous speech. *Language*. 84. 474–496.

GAHL, SUSANNE. 2009. Homophone duration in spontaneous speech: A mixed-effects model. *UC Berkeley Phonology Lab Annual Report*. 279–298. Online: http://linguistics.berkeley.edu/phonlab/annual_report

GELMAN, ANDREW, and JENNIFER HILL. 2007. *Data analysis using regression and multilevel, hierarchical models (Analytical methods for social research).* Cambridge: Cambridge University Press.

JURAFSKY, DANIEL. 2003. Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. *Probabilistic linguistics,* ed. by Jennifer Hay, Rens Bod, and Stefanie Jannedy, 39–95. Cambridge, MA: MIT Press.

LOHMANN, ARNE. 2017. *Cut*(N) and *cut*(V) are not homophones - Lemma frequency affects the duration of noun-verb conversion pairs. *Journal of Linguistics.* Online via *JL First View*: https://doi.org/10.1017/S0022226717000378

MIDDLETON, ERICA; QI CHEN; and JAY VERKUILEN. 2015. Friends and foes in the lexicon: homophone naming in aphasia. *Journal of Experimental Psychology: Learning, Memory, and Cognition.* 41. 77–94.

PITT, MARK; LAURA DILLEY; KEITH JOHNSON; SCOTT KIESLING; WILLIAM RAYMOND; ELIZABETH HUME; and ERIC FOSLER-LUSSIER. 2007. *Buckeye Corpus of Conversational Speech (2nd release).* Columbus, OH: Department of Psychology, Ohio State University.

R CORE TEAM. 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Online: http://www.R-project.org/.

WRIGHT, CHARLES E. 1979. Duration differences between rare and common words and their implications for the interpretation of word frequency effects. *Memory & Cognition.* 7. 411–419.

[1] A further complication with the control variables arises through fitting the model to the average duration of many tokens, as many control variables influence only a certain share but not all tokens of a particular type, for example, whether a pause follows the word or not. Gahl addresses this issue by calculating average values for all control variables, e.g. the ratio of tokens followed by a pause. This means that mean values averaging over a large number of tokens are used. Given that the duration of an individual token is affected by many situational and contextual variables, this approach is deemed to lose a great deal of information about the actual factors that influence the durations of the words making up the sample. It is partly for this reason that Gahl conducts the reanalysis on a token basis, reported in Gahl (2009).

[2] Gahl tested for significance of this predictor by comparing the fit of the model with and without it via the calculation of an ANOVA, which returns a *p*-value of $p<0.001$ (Gahl 2008: 484). The frequencies of this predictor are taken from the Switchboard corpus.

[3] My own search of the Switchboard corpus reveals that this one occurrence is part of the compound *lemon thyme*. This renders the question of how reliable this token is for the calculation of an average duration of the word *thyme* even more acute, since being part of a compound may have certain prosodic effects impacting the duration of this token.

[4] The predictor is based on frequencies that I retrieved from the SUBTLEX-US corpus (Brysbaert & New 2009). Gahl performed her original calculations based on frequencies obtained from Switchboard. Since Switchboard is a fairly small corpus, I retrieved frequency values from the larger SUBTLEX-US corpus as I consider it more representative of the linguistic experience of the speakers recorded in Switchboard. Using the original Switchboard frequencies does not change the pattern of results as reported in the following.

[5] When calculating the model on the dataset by Gahl, I noticed that for 12 out of the 220 homophone pairs, only one member of the pair was instantiated in the sample (probably due to the low-frequency member not occurring in the Switchboard corpus), which is why I excluded these 12 pairs. For another two pairs SUBTLEX-US frequencies could not be retrieved for both homophones, so these two pairs were also excluded. As a result, the model reported here is based on 206 homophone pairs, which explains the slightly lower sample size compared to Gahl's original model reported in Section 4.

[6] A potential cause for concern regarding the results of the models are the bigram probabilities, as one would expect high-frequency words to be more predictable than low-frequency words. Since the coefficients of regression models express the rate of predicted change when other variables are held constant, holding bigram probabilities constant is tantamount to testing the effect of frequency differences in potentially unusual environments not representative of the typically higher predictability of high-frequency words and the typically lower predictability of the low-frequency words. In order to address this concern, I also calculated models without the bigram probabilities. Crucially, the predictors capturing the frequency difference are still statistically significant and the direction of the frequency effect does not change, i.e. the same duration difference between the homophones is observed. I thank Roger Levy, who reviewed this paper, for drawing my attention to this issue and suggesting this additional analysis.

[7] The random effects structure of these additional models includes only random intercepts but no random slopes, because their inclusion resulted in nonconvergence of the models.