FACULTY OF HUMANITIES
UNIVERSITY OF COPENHAGEN

Word frequency and sound change in groups and individuals

Nicolai Pharao

LANCHART – The Danish National Research Foundations Center for the Study of Language Change in Real Time

Usage based approach to sound change

Studies in the usage based approach to sound change (Bybee (2002)) have shown that reduction affects high frequency items before low frequency items. This is supported by studies of corpora of speech from a variety of languages and has also been observed as a general tendency in reductive changes arising in Copenhagen Danish throughout the 20th century (Brink & Lund (1975), pp. 729-730). These findings support an exemplar based model of the mental lexicon (e.g. Johnson (1997) and Pierrehumbert (2002)).

An important aspect of these models in relation to sound change is that the frequency effect should be observable at the level of the individual language user, in order for word-form frequency to be explanatory and relevant in the description of representation in an exemplar based mental lexicon (Pierrehumbert (2001) p. 148).

Sound change is most often studied by comparing the speech of speakers from different generations, and inferring from differences in speech behavior whether a sound change is going on – such studies are referred to as *apparent time* studies (following Labov (1966). Recently, a growing number of studies chart on-going sound change by examining changes in speech behavior within the speech of individual speakers. This is done by comparing the behavior of the same individuals at different points in time through the analysis of recordings made several decades apart. Such studies are referred to as *real time* studies (cf. Sankoff 2006 for a review).

In this study the frequency effect was tested by conducting both an apparent time and a real time study of the deletion of [w] before syllabic [ð̩] in Copenhagen Danish. The following questions were asked:

1) Is deletion of [w] more likely in high frequency word-forms than in low frequency word-forms?
2) Are younger speakers more likely than older speakers to delete [w] in relatively low frequency word-forms?
3) Can the frequency effect be observed at the level of individual speakers?
4) When a speaker continues to participate in the process across the lifespan, does the spread of [w] deletion follow the path of word-form frequency observed in apparent time?

[w]-deletion in Copenhagen Danish

In Copenhagen Danish, original [w] is deleted before syllabic [ð̩]. Deletion of [w] in this context can be interpreted as a process of reanalysis by the listener (cf. Ohala (1981): since Danish [ð] is slightly velarized in any style of speech, this velarization can become perceptually more salient and the velar characteristics of the approximant [w] can be attributed to the following [ð], that is the rounding in [w] is in effect drowned by the velarization, leading to deletion of [w] in running speech (see also Grønnum (2005), p. 334).

A previous study of this reduction process (Brink & Lund (1975)) identified 2 conditioning phonetic factors:

Preceding vowel quality: [w] deletion is more likely after front vowels than after back vowels, e.g. more likely in 'skrevet' ['sgˀʁæ:wð̩] *have written* than in 'sproget' ['sbʁɔ:ˀwð̩] *the language*

Stress: [w] deletion is more likely in unstressed syllables, than in stressed syllables, e.g. more likely in 'kirkelivet' ['kʰiɐgˀəli:ˀwð̩] *church life* than in 'livet' [li:ˀwð̩] *life*

The LANCHART corpus

Sociolinguistic interviews with 22 middle class subjects from 1987 and re-recordings of the same speakers made in 2006 were analyzed. The subjects were divided with respect to age and gender .

Generation	Older	Younger
male	6	5
Female	6	5
Total	12	10

Table 1 – Distribution of subjects by gender and age group
Subjects in the "Older" generation were born between 1946 and 1962, and subjects in the "Younger" generation were born between 1967 and 1973

References
Brink, L.& J. Lund (1975) *Dansk Rigsmål 1 & 2*, Gyldendal, Copenhagen, Denmark.
Bybee, J. (2002) *Phonology and Languag Use*, CUP, Cambridge, UK.
Grønnum, N. (2005) *Fonetik & Fonologi – almen og dansk* (3rd ed.) Akademisk, Copenhagen, Denmark.
Johnson, K. (1997) "Speech perception without speaker normalization" in K. Johnson & J.W. Mullenix (eds.) *Talker Variability in Speech Processing*, Academic Press, San Diego, USA, pp. 145-166.
Labov, W. (1966) *The Social Stratification of English i New York City*, CUP, Cambridge, UK.
Ohala, J. J. (1981) "The Listener as a Source of Sound Change" in Masek, Hendrick & Miller (eds.) *Papers from the Parasession on Language and Behavior*, Chicago Linguistics Society, Chicago, USA, pp. 178-203.
Pierrehumbert, J. (2001) "Exemplar dynamics: Word frequency, lenition adn contrast" in J. Bybee & P. Hopper (eds.) *Frequency and the Emergence of Linguistic Structure*, John Benjamins, Philadelphia, USA, pp. 137-157.
Mendoza-Denton, N. , J. Hay & S. Jannedy (2003) "Probabilistic Sociolinguistics" in R. Bod, J. Hay & S. Jannedy (eds.) *Probabilistic Linguistics* MIT Press, Cambridge, USA, pp. 97-138

Data analysis

Two phonetically trained listeners coded all tokens of [w] before [ð̩]. The tokens were coded for two target variants: deletion, and realization as [w]. Deletion was defined as 'no audible trace of a rounded, non-syllabic segment intervening between the nucleus of the syllable containing underlying [w] and the [ð̩] of the suffix'. For the statistical modeling, only tokens in which both coders were certain of the classification were included.

Phonetic factors

In order to be able to take the previously described phonetic factors into account in the statistical analysis of the data, each token was also coded for the quality and height of the preceding vowel as well as the degree of stress on the syllable containing the underlying [w].

Word frequency

To study the frequency effect on [w]-deletion in Copenhagen Danish, word form frequency was included as a factor in the statistical analysis. Frequency counts from spontaneous speech were used to model the effect. The frequencies were calculated on the basis of the entire LANCHART Corpus of sociolinguistic interviews. This corpus comprises speech from 6 different locations in Denmark, and represents the speech of more than 500 speakers, i.e. the counts are not based solely on the speech of the 22 subjects sampled for the study of [w]-deletion. The counts were based on the phonetic annotations of the orthographical transcripts of the interviews and those represent frequencies of spoken word forms. A total of 51.979 different word forms occur in the corpus, with a total of just over 3 million tokens in all (3.088.350). Word forms were assigned relative frequencies by dividing the number of tokens of a given word form with the total number of tokens in the LANCHART Corpus, and the relative frequency counts were log transformed following Mendoza-Denton, Hay & Jannedy (2003).

[w]-deletion across generations

The full set of 418 tokens of word forms containing [w] before [ð̩] in the recordings from 1987 was reduced to a dataset of 336 (56 discarded due to background noise, 26 discarded due to disagreement between the two listeners). The results were analyzed statistically using mixed-effects multiple logistic regression in R 2.7.2 with subject and word form as random effects. The results show an overall effect of year of birth with the probability for [w]-deletion increasing with year of birth, as shown in the graph. The difference between the two generations is also significant ($p = 0.001$) with 73 % of tokens showing deletion for the older generation and 88 % of tokens showing deletion for the younger generation.

Phonetic factors in 1987 – the apparent time study

The effect of the phonetic factors on the deletion of [w] as estimated in the logistic regression models are given in the two tables of co-efficients below, one table for each generation

Table 2 - Phonetic factors in 1987, the older generation

Factor	Estimate	Std. Err	Pr(>|z|)
(Intercept)	-10.79	3.20	0.002 **
unstressed	1.29	0.46	0.005 **
front vowel	3.35	0.86	<0.0001 ***

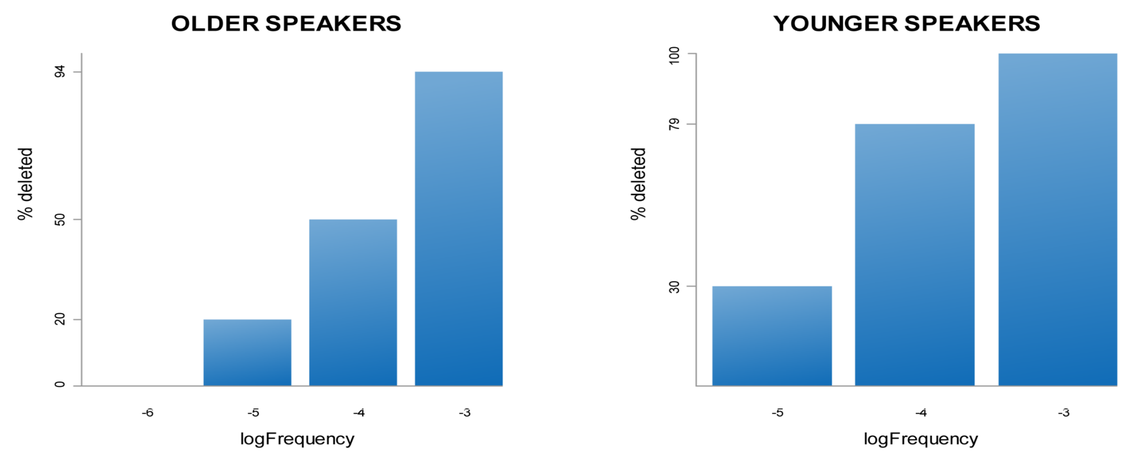
Table 3 - Phonetic factors in 1987, the younger generation

Factor	Estimate	Std. Error	Pr(>|z|)
(Intercept)	2.42	12.80	0.85
unstressed	1.44	0.82	0.07 .
front vowel	5.79	2.11	0.007 **

The phonetic factors show the expected effects for the older generation, with [w]-deletion being more likely in unstressed syllables than in stressed syllables with full stress, and deletion is more likely after front vowels than after back vowels.

For the younger generation, the effect of stress has disappeared, in comparison to the older generation, but the effect of vowel quality is also significant for the younger generation, with [w]-deletion again being more likely before front vowels than before back vowels.

Deletion of [w] as a function of word form frequency in 1987



Word frequency across generations

The factor word form frequency emerges as significant in the models both for the older and the younger generation. The results are given in the graph above. Each column represents a frequency category and give the proportion [w]-deletion within each category.

The graphs show that [w]-deletion is more likely in high frequency words than in words of lower frequency. Furthermore, the group of younger speakers have a higher rate of deletion in mid to low frequency word forms than the group of older speakers. The results thus confirm that reduction spreads from high frequency to low frequency words and that this pattern is also operative in the transmission of [w]-deletion from the older to the younger generation.

Individuals and word form frequency

To investigate whether the effect of word form frequency found at the group level also holds for individual speakers, the rate of [w]-deletion for each speaker in the two generations in table 4 and 5.

Word frequency in the older generation

For two subjects in the older generation the number of tokens was so small that percentage of tokens with deletion were not calculated, and they have therefore been omitted from the table.

Table 4 – Percentage of tokens with [w]-deletion by log frequency for individual subjects in the older generation

Subject	*Number of tokens*	*% deleted log frequency -6*	*% deleted log frequency -5*	*% deleted log frequency -4*	*% deleted log frequency -3*
LAL	11	-	0	0	75
CNI	20	-	-	20	100
TNI	33	-	33	50	87
PTK	38	0	0	37	94
ASA	11	0	-	40	100
EAF	11	-	50	-	100
HTH	20	-	0	50	100
MJE	15	-	0	50	100
CEL	17	-	100	100	87
MFL	22	0	0	100	90

For the remaining speakers the results show a clear tendency for the rate of deletion to be highest in high frequency words and to decrease gradually with word form frequency. However, for two of the speakers, CEL and MFL, the pattern does not strictly follow word frequency, since they actually have a lower rate of deletion in high frequency words than in mid and low frequency words.

Word frequency in the younger generation

For four of the speakers in the younger generation, the total number of tokens was too low to allow for an analysis of the rate of [w]-deletion as a function of word frequency, but for the remaining six speakers in this group the pattern can be seen in the table. And here the result clearly shows that they all have the highest rate of deletion in high frequency words, in fact [w] deletion is obligatory in words of this type for all six speakers.

Table 5 – Percentage of tokens with [w]-deletion by logfrequency for individual subjects in the younger generation

Subject	*Number of tokens*	*% deleted log frequency -6*	*% deleted log frequency -5*	*% deleted log frequency -4*	*% deleted log frequency -3*
MIP	18	-	-	50	100
JJE	12	-	-	67	100
PKJ	20	-	0	75	100
DBE	17	-	0	80	100
JOR	14	-	-	100	100
MPT	16	-	-	100	100

The frequency effect at the level of the speaker

Taken together the results for the two generations show that there is an overwhelming tendency for the frequency effect that was observed at the group level to also hold true for individual speakers. This supports the interpretation of the frequency effect in sound change as a reflection of an exemplar based mental lexicon, in which representations of word forms are incrementally updated as the words are used in daily speech.

[w]-deletion across the lifespan

The study of [w]-deletion also addressed the question of stability of speech behavior across the lifespan by examining the same process in re-recordings of the 22 subjects made in 2006. The overall results show that even when phonetic factors are controlled for there is a general increase in the rate of [w]-deletion, and that deletion is still more likely in the younger generation than in the older generation, despite the fact that all speakers were almost 20 years older at the time of the re-recordings. Again the tendency for [w]-deletion increases with word form log frequency ($\beta = 1.15$ for the older and $\beta = 1.56$ for the younger generation; $p < 0.001$ for both generations). The effect of frequency can be seen in the graph below, where the rate of [w]-deletion as a function of frequency is compared for the two generations. Deletion of [w] is increasingly likely in high frequency words, and it spreads to words of even lower frequency than was observed in the 1987 recordings.

This is a general trend for all speakers in the sample, but only three speakers show a statistically significant increase in their rate of [w]-deletion across the lifespan, mainly because the sequence of [w] and It would be interesting to study a more frequently occurring phonetic variable to see whether the pattern for individuals observed here is a general feature of sound changes involving reduction.

Summary

Taken together, the results from both the apparent time and the real time study of [w]-deletion in Copenhagen Danish supports the hypotheses of a usage based approach to sound change. The analyses further show that the frequency effect also obtain at the level of individual speakers, which supports an exemplar based model of lexical representation where word forms are updated incrementally as they are encountered and used in running speech.

Deletion of [w] as a function of word form frequency in 2006