Sound change, priming, salience

Producing and perceiving variation in Liverpool English

Marten Juskan

Language Variation 3

./langsci/graphics/lar

Language Variation

Editors: John Nerbonne, Dirk Geeraerts

In this series:

1. Côté, Marie-Hélène, Remco Knooihuizen and John Nerbonne (eds.). The future of dialects.
2. Schäfer, Lea. Sprachliche Imitation: Jiddisch in der deutschsprachigen Literatur (18.–20. Jahrhundert).
3. Juskan, Marten. Sound change, priming, salience: Producing and perceiving variation in Liverpool English.

ISSN: 2366-7818

Sound change, priming, salience

Producing and perceiving variation in Liverpool English

Marten Juskan

./langsci/graphics/langsci

Marten Juskan. 2018. *Sound change, priming, salience: Producing and perceiving variation in Liverpool English* (Language Variation 3). Berlin: Language Science Press.

This title can be downloaded at:
http://langsci-press.org/catalog/book/210

ISBN: 978-3-96110-119-1 (Digital)
978-3-96110-120-7 (Hardcover)

ISSN: 2366-7818
DOI:10.5281/zenodo.1451308
Source code available from www.github.com/langsci/210
Collaborative reading: paperhive.org/documents/remote?type=langsci&id=210

Cover and concept of design: Ulrike Harbort
Fonts: Linux Libertine, Libertinus Math, Arimo, DejaVu Sans Mono
Typesetting software: XƎLATEX

Language Science Press
Unter den Linden 6
10099 Berlin, Germany
langsci-press.org

Storage and cataloguing done by FU Berlin

./langsci/graphics/storagelogo.pdf

To Daniela

Contents

5 Interview method

Chapter 3 has introduced the four variables that this book focuses on. On the basis of previous work, it has also broadly divided them into a (comparatively) non-salient and a highly salient, even stereotyped and stigmatised group. However, this distinction was largely based on experts' judgements and evaluations, and especially external stereotypes may well "become increasingly divorced from the forms which are actually used in speech" (Labov 1972: 180). This part of the book will therefore try to corroborate the alleged salience of the variables 'from the inside', as it were, and also to go beyond the simple binary distinction of salient/non-salient by ordering our four variables more precisely in relation to each other on the social salience scale.

5.1 Interview structure

Production data for the four variables of interest were obtained in the form of "classical" sociolinguistic interviews. All of these interviews were one-on-one and conducted by the author. Being an outsider to the community entails a number of disadvantages with respect to naturalness of speech of the subjects. However, this was true of all interviews in the same way, so it cannot be a factor influencing inter-group comparisons. The interviews consisted of a free speech section where subjects were asked a number of questions about the area of the city they grew up in, changes in the city, football and other sports, Liverpool's image in the UK and the rivalry with Manchester.[1] Furthermore, subjects were questioned about their use (particularly with respect to themselves) and their understanding of a number of identity labels. See appendix A for the complete questionnaire. Not all questions were asked in all interviews, but all topics were discussed or at least touched upon with every participant, with most of the time

[1] Although this rivalry has historical reasons (cf. Chapter 2), it is today dominated by the rivalry between the football clubs from Liverpool and Manchester in many people's minds. This does not, however, diminish its potential for bringing up questions of identity and local pride in the slightest. Indeed, as Beal (2010: 97) remarks, "[t]he football derby (…) is one of the clearest manifestations of local identity and rivalry in Britain today"

typically devoted to the areas ‘children’s lore’, ‘attachment to Liverpool’, ‘identity’, and ‘Liverpool’s image’.

Towards the end of the interview, participants read out a reading passage (see appendix B) and a list of keywords (appendix C). Most of the test words on the list were also contained in the reading passage for better comparability. Next, subjects were asked to read out the reading passage a second time using their strongest Scouse accent. Not all interviewees wanted to do this or explained they weren’t capable of “putting it on” on demand, but the vast majority of participants completed all three reading tasks. In graphs showing register differences, the data gathered during the accent imitation task will be situated towards the informal end of the style spectrum. I am aware of the fact that the imitation style is almost certainly one where subjects are likely to pay *more* than average attention to their speech. A reviewer quite correctly points out that accent performance probably qualifies as a “frozen, ritualistic” style (Labov 1972) that, in terms of attention, should rather be placed towards the more formal end of the style continuum. However, this task should still – for obvious reasons – trigger the most ‘extreme’ and/or most frequent local variants, even when compared to spontaneous speech, so in that sense I would argue it is quite different from, say, a sermon or some other form of scripted public speech. A linear increase of local variants can be expected from word list through reading and free speech style to accent imitation, so it seems to me the placement of the latter towards the ‘informal’ end of the style spectrum is justified in that respect. Purely for reasons of convenience, accent imitation will occasionally be referred to as the ‘most informal’ speech style, simply because it should be the most ‘vernacular’ register, *not* because I believe subjects paid no attention to their speech.

Finally, subjects were asked a number of questions concerning Scouse, notably whether they thought the accent had changed in their life time and what features they considered most typical. Analysis of these statements can only be qualitative in nature and should be considered an impressionistic snapshot rather than anything close to a representative picture of the relevant groups’ explicit linguistic knowledge. Usually, the interviews lasted between 50 and 60 minutes (40–45 minutes of free speech and 10–15 minutes of reading/accent imitation and metalinguistic comments). Testing took place in a number of locations: pubs and cafés in central Liverpool, cafeterias at Hope University and the University of Liverpool, people’s offices and homes. Not all of these environments were equally quiet, but recording quality was at least acceptable in all cases. All interviews were recorded using a Roland Edirol R-09HR MP3/Wave recorder, and named according to the following pattern:

1. a two digit participant/interview number
2. "F" or "M" to code the participant's gender
3. "MC" or "WC" to code the participant's social class
4. two digits coding the participant's age in years at recording time

"02MWC20", for example, is the code for interview number 2 with a male, working-class subject, who was 20 years old at the time of the interview. These codes will occasionally be used in this study to refer to specific interviews or to attribute quotations to their sources.

5.2 Participants

Participants were recruited through a number of ways. Notes in pubs, cafés, football grounds, community centres, and churches were complemented by e-mail calls for participants through Hope University and the University of Liverpool mailing lists, word-of-mouth advertising and by approaching people in person (mostly students at Liverpool Hope University). Interviews were conducted during two field trips, in September/October 2012 and April/May 2013, respectively. The first 8 subjects participated for free, the remaining ones were offered £10 for their time (some declined). No selection of participants in terms of 'typicality' or 'strength of accent' was made (as opposed to, for example, the 'new NORMs' in Honeybone 2001).

A total of 38 subjects were interviewed. All participants were born and/or had grown up in the Liverpool Urban Area since age 12 or younger. Several subjects had also lived in other cities or towns at one point or another of their life, the reason usually being either job or (university) education related. Most interviewees, however, had spent all their life in Liverpool and its suburbs. Both men and women were interviewed and a rough socio-economic distinction into working class or middle class was made. English was the first (and, with the exception of one participant who was later excluded, also the only) language for all subjects. All participants were White British. The age range was 19–85, with people being classified as belonging to one of three age groups (19–29, 30–55, and 56–85) to mirror social, economic, and cultural change in Liverpool. With the boundaries set as they are the formative years (roughly up to and including the 20s) of most of the participants in the respective group fall together with one of the three phases of the city's development in the latter half of the 20th century (cf. §2.3

and §2.4): 50s and 60s (post-war recovery and Merseybeat era) for the oldest, 70s and 80s (economic depression) for the middle-aged, and 90s and 2000s (regeneration) for the youngest speakers. For reasons of time and space, only 20 interviews could be included in the present study. Interviews entered this sub-sample in the order they had been conducted in until all cells (cf. Table 5.1) were represented by 2 informants (1 in the case of the oldest group). These subjects form what I will call the 'primary sample' for the production part of this study. In total, they contributed almost 19 hours of recorded material. The secondary sample (including all 38 interviews) is the basis for some results in Chapter 8, but other than that all production analyses are exclusively based on the smaller primary sample. Table 5.1 shows how participants in this primary sample are distributed across the categories outlined above.

Table 5.1: Age, gender, and social class of subjects (production)

	19–29		30–55		56–85	
	F	M	F	M	F	M
WC	2	2	2	2	1	1
MC	2	2	2	2	1	1

Figure 5.1 – generated with the QGIS software (QGIS Development Team 2016)[2] – illustrates which part of the city/conurbation the subjects are from or, to be precise, where they currently live. As is clear from the map, most areas of the city are represented although, to be fair, some (suburban) northern parts of Liverpool are underrepresented. There is also a slight bias towards the area around Liverpool districts Aigburth, Mossley Hill and Allerton in the south end of the city (12 subjects in total are from one of these three areas). Note, however, that all age groups are more or less evenly spread across the city.

The study was not restricted to people from within the Liverpool Council boundaries (black line in the map), but also included areas which are administered by other local councils (Sefton, Knowsley, Wirral) and which are, therefore, "technically not Liverpool" as a number of subjects put it. This is indeed, however, more of a technicality since we are talking about a contiguously built up area – just like in most other urban agglomerations. It is clear that invisible lines (sometimes separating one side of a street from the other) can still

[2] Map tiles by Stamen Design, under CC BY 3.0. Data by OpenStreetMap, under CC BY SA. Shapefiles from CDRC 2015 OS Geodata Pack by the ESRC Consumer Data Research Centre; contains Ordnance Survey data © Crown copyright and database right 2015.

figures/liverpool-participant-map.png

Figure 5.1: Geographical distribution of interview subjects

be important for people's identity, but all of the participants in this study self-identified as Liverpudlians or Scousers. This also held for the two subjects who were actually living on the Wirral and who had both been born in Liverpool (and in one case also lived half her life within Liverpool city boundaries). Generally speaking, people in urban areas often move around quite a bit and this might be especially true for Liverpool where many people from inner city areas were actually relocated (sometimes very reluctantly so) to new housing estates on the outskirts of the city during the slum clearances of the 50s and 60s. This is indeed what many of the older participants experienced themselves. For these reasons it was deemed unjustified to restrict the pool of subjects to those living within Liverpool city boundaries only.

5.3 Transcription

All interviews were transcribed orthographically in Praat (Boersma & Weenink 2015) by the author. Since the transcriptions' sole purpose was to serve as input for automatic measuring (cf. §5.4), pauses, intonation, stress, etc. were not marked in the transcripts. Questions and other utterances by the interviewer were also ignored. On separate tiers of the Praat TextGrid, speaking style (word "list", "reading" (passage), "free" (speech), and (accent) "imitation") and topic ("childhood", "Manchester", "identity" etc.) coded, followed by a third one where the participant's speech was segmented into chunks and transcribed. Words containing test tokens and the individual variables themselves were marked on individual tiers called "word" and "variable" respectively. Finally, a sixth tier called "aspiration" was used to mark relevant parts of the consonantal variables (cf. §5.4.1). Figure 5.2 provides an extract from a TextGrid (zoomed to word level) for purposes of illustration.

replaced 'equally'

Figure 5.2: Extract of Praat TextGrid (subject 30FMC44)

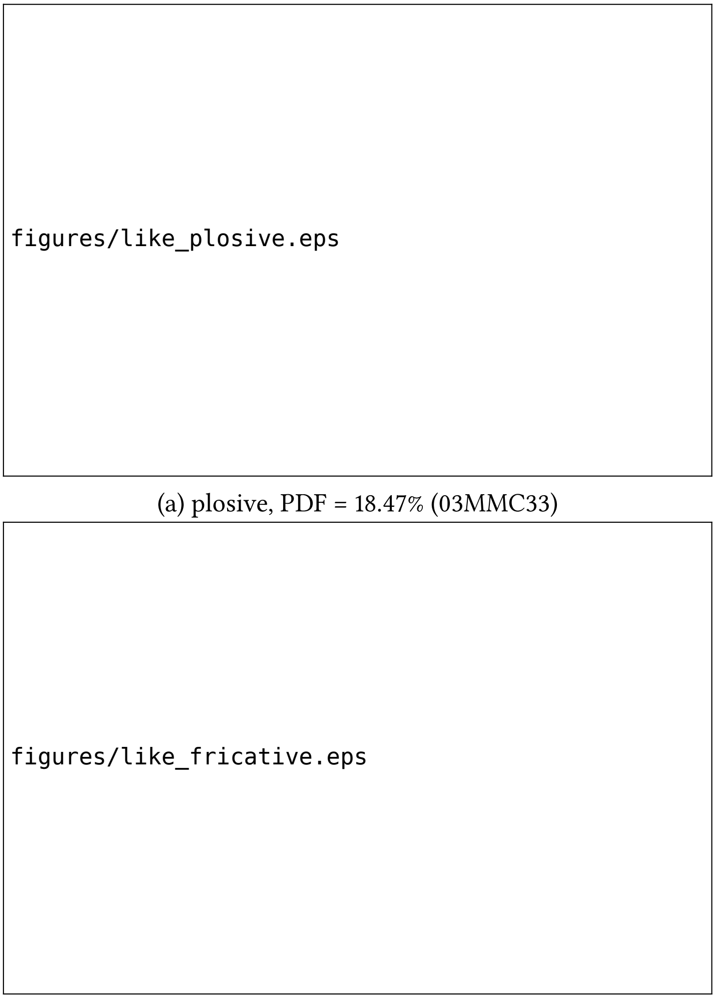
5.4 Measuring

5.4.1 Consonants

The two consonantal variables were analysed both acoustically and auditorily. The method for acoustic measuring of /k/ was heavily inspired by the one used in Sangster 2001 to investigate lenition of alveolar stops. Phonetic plosives have a period of silence, or closure, followed by a burst and friction. For affricates, there is the same silence, but more friction than for plosives, and fricatives have either a very short period of silence or none at all and consist (almost) entirely of friction.

Beginning and end of the friction phase were marked in a Praat TextGrid for every /k/. A script written by the author was then used to automatically measure the duration of these segments as well as the total durations of the plosives (i.e. including the closure phase). /k/ tokens without any friction phase were registered as ‘unreleased’ (and ignored in the analysis). Next, what Sangster calls ‘the proportional duration of friction’ (PDF) was calculated by dividing the duration of the friction phase by the total duration of the plosive. The result is a figure between 0 (or 0%) and 1 (100%), with lower values for more plosive-like realisations and higher values for sounds that are phonetically speaking affricates or fricatives.

The same technique was applied to /ŋ(g)/. This decision might seem strange at first, because the realisational options of /ŋ(g)/ do not seem to be readily comparable to those of /k/. Closer examination, however, reveals that the standard realisation as a nasal [ŋ] involves complete oral closure – just as with [k] – and that for the typical Scouse realisation as [ŋg] this closure phase is followed by a release burst / friction. While the friction of [ŋg] will never be as long as that of a /k/ realised as a fricative, the PDF values will mean the same thing for velar nasal plus as they do for /k/: lower values (no or little friction ⟶ [ŋ]) indicate a standard-like realisation and higher scores (presence of friction ⟶ [ŋg]) mark non-standard, Scouse variants. Alveolar variants of /ŋ(g)/ were coded as “in” and later removed for the quantitative analyses for two reasons. First, [n] is a non-standard variant that is not limited to Liverpool or even a clearly bounded region, but one that is used in all varieties of English English and many others as well. It is also rather salient and commented on by many non-linguists as ‘g-dropping’. However, in order to assess the impact of salience, particularly in perception, this study required a local/regional feature with little or no salience, to compare to the highly salient and local /k/ lenition. Alveolar variants of the <ng> cluster fulfil neither criterion, while [ŋg] realisations tick both boxes. The second reason concerns the method of measurement. Realising <ng> as [n] by definition excludes the presence of even a hint of a plosive, so the PDF measurement outlined above is not applicable. The difference between [ŋ] and [ŋg] (or the devoiced variant [ŋk]), on the other hand, exhibits the same kind of gradualness and, as explained above, can be measured in the same way as /k/ lenition. This parallelism is again crucial for the perception experiment, because it means the stimuli for /k/ and /ŋ(g)/ could be manipulated in a way that was phonetically similar (and thus not a confound). Since linking up data from production and perception is a major interest of this study, the focus in the production part was also exclusively on the [ŋ]-[ŋg] distinction. Figure 5.3 shows two examples and their respective marking in the TextGrid.



(a) plosive, PDF = 18.47% (03MMC33)

(b) fricative, PDF = 81.84% (36FWC20)

Figure 5.3: Spectrograms of /k/ (zoomed to word level)

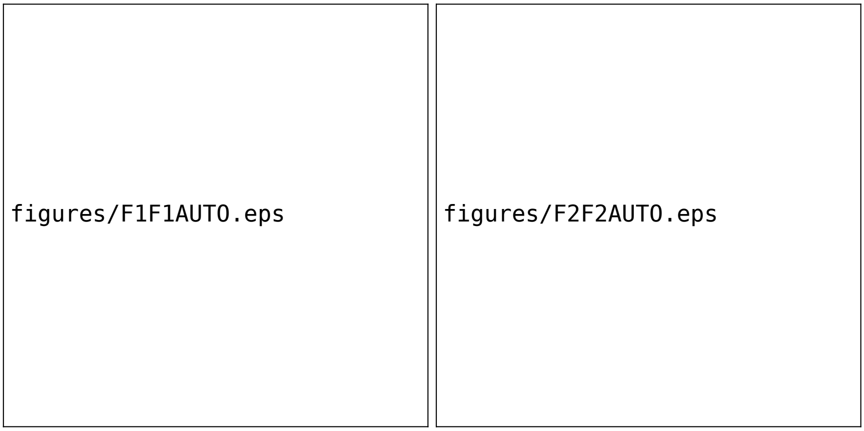
This very precise method of acoustically measuring /k/ and velar nasal plus requires high quality recordings with little to no background noise. As it was unclear at the beginning whether all interviews fulfilled these criteria, the data were also analysed auditorily by the author. Coding was '0' (plosive), '1' (affricate), and '2' (fricative) for /k/, and '0' (nasal) and '1' (nasal plus burst) for /ŋ(g)/. It turned out that all interviews included in this project actually did permit an analysis based on the more precise Sangster method, so the auditory coding was not used in the analysis in the end. It is, however, still accessible for future research.

5.4.2 Vowels

For the measurement of the first two (later three) vowel formants (NURSE, SQUARE, and HAPPY) a Praat script[3] was used to automatise data collection. NURSE and SQUARE were measured first by hand and then in an automated way by the script for the first three (male) subjects. Paired t-tests were then administered to make sure the automated measurements were reliable. Neither test ([t(545) = -0.975, p = 0.330] for F1 and [t(545) = 1.768, p = 0.078] for F2) found a significant difference between hand and automated measurements, although there was a trend for the F2 values. However, the mean difference between hand and automated measurements for F2 was a mere 2.15 Hz. Scatterplots furthermore show a near-perfect correlation of hand and automated measurements, which is why the script was deemed reliable and all formant measurements used in the final analysis were taken automatically only. Clear mismeasurements were later removed from the dataset.

The script took as input pairs of sound files and TextGrids. It then went through each TextGrid and looked for vowel labels in the variable tier. When it found a relevant label it noted the start and end of the segment and measured F1, F2, and F3 at midpoint of the vowel. It then extracted information about the style, topic, carrier word, and the larger context it appeared in from the other tiers and saved all these data into a textfile. F3 was measured because it was needed for one of the normalisation algorithms that were later applied to the raw measurements (cf. §5.4.3). In addition to the three vocalic test variables happy, NURSE, and SQUARE (of which all instances were included), between 10 and 25 tokens of FLEECE and TRAP per subject were also measured. These were taken from the reading passage and word list sections of the interviews since these contexts were considered most likely to produce the most 'extreme' realisations (in terms of the periphery

[3] Generously made available by Mietta Lennes – http://www.helsinki.fi/~lennes/praat-scripts/, last accessed 2013-01-29 – and modified by the author.



(a) F1 measurements

(b) F2 measurements

Figure 5.4: Manual (x-axis) vs. automatic (y-axis) measurements of NURSE and SQUARE

of speakers' vowel spaces). Observations of TRAP were used exclusively as input for normalisation and for comparison of the algorithms (again, cf. §5.4.3). FLEECE measurements were additionally included in the calculation of Pillai scores for happY (cf. §6.1.3).

5.4.3 Normalisation

It is a well known fact among phoneticians and phonologists that there is a huge amount of variation in the acoustic signal that is not due to linguistic or sociolingustic, but rather purely physiological reasons. Even multiple realisations of one and the same phonological sound chain produced by a single speaker in the same style will all be slightly different from one another. In addition to these intra-speaker differences, there are also inter-speaker ones. The most pronounced differences in this area are due to vocal tract length. The length of the vocal tract correlates inversely with vowel formant values. On average, therefore, children (with the shortest vocal tracts) have higher formants than women, who in turn have higher formants than men for one and the same phonological vowel. The potential effect of vocal tract maturation, i.e. changes to length and shape of the vocal tract over the course of an individuals lifetime, further complicates matters (cf. Harrington 2006: 440–441).

It is therefore not possible (or at least not advisable) to directly compare, for instance, women's and men's raw formant values, or those of younger and older speakers. This is where normalisation comes in. According to several articles on the matter (cf., for example, Fabricius et al. 2009; Clopper 2009; Disner 1980; Kendall & Thomas 2009; Thomas 2002), normalisation should ideally achieve four different goals:

1. elimination of differences that are due to physiological reasons
2. preservation of differences that are (socio-)linguistic in nature
3. preservation (or improvement) of phoneme distinctions
4. modelling the process that allows listeners to assign realisations from different speakers to one and the same phoneme

The author is well aware of the irony involved here. This study is, after all, set in an exemplar framework which suggests that listeners do *not* normalise acoustic input, at least not in the same way and to the same degree as is assumed in most other phonological theories. This is most relevant with respect to point 4 in the enumeration above. Sociolinguists, however, usually largely ignore this aspect and focus more on points 1 and 2 (cf. Clopper 2009: 1430; Fabricius et al. 2009: 414–415; Kendall & Thomas 2009), and the present study is no exception. By applying a normalisation algorithm to the data I do not mean to suggest that this procedure mirrors or approximates what happens in listeners' brains. Rather, it is simply the only option one has if the goal is to compare production data of men and women (or those of younger and older speakers) to each other instead of treating them separately.

Normalisation methods are generally categorised with respect to two dimensions: vowel-intrinsic vs. vowel-extrinsic and speaker-intrinsic vs. speaker-extrinsic (cf. Kendall & Thomas 2009). Vowel-intrinsic algorithms extract all data necessary for normalisation from the individual token. Often these methods use F0 and/or F3 to estimate vocal tract length. Vowel-extrinsic algorithms include formant measurements from more than one vowel in their formulas and achieve normalisation with the help of means over several (often all) measured vowels. Speaker-intrinsic methods differ from speaker-extrinsic ones in that the former perform normalisation for each speaker individually (i.e. only taking into account vowels produced by that speaker), whereas the latter include some sort of inter-speaker mean in their calculations (cf., for example, Labov et al. (2006)'s *grand mean*).

A number of algorithms have been proposed over the years, and the question which of those fares best in achieving the goals spelled out above has generated a series of investigations (among others: Hindle 1978; Disner 1980; Adank et al. 2004). Generally speaking, "vowel-extrinsic methods tend to perform better overall (…) for vowel space normalization across talkers", and "vowel-intrinsic methods are appealing as perceptually plausible models of human speech processing" (Clopper 2009: 1440). For this reason, two different normalisation methods were tested in this study, a vowel-intrinsic and a vowel-extrinsic one (both of them speaker-intrinsic). Both normalisations were applied to the raw data using the NORM package for R (Kendall & Thomas 2009). The first, Bark difference, was devised by Syrdal & Gopal (1986), and is a vowel-intrinsic method. Formants are, first of all, transformed into – perceptually "more accurate" (Clopper 2009: 1431–1432) – Bark values using the formula taken from Traunmüller (1990):

$$Z_i = \frac{26.81}{1 + \frac{1960}{F_i}} - 0.53 \tag{5.1}$$

Where F_i is the raw value of a given formant. The Bark rescaled values Z_1 and Z_2 are then substracted from Z_3 to arrive at normalised measures of height and frontness respectively. Syrdal & Gopal originally used Bark-converted F0 instead of F3 for the height dimension, but Kendall & Thomas (2009) argue that a number of things, for instance "[i]ntonation, tone, and consonantal influences affect F0" and consider it preferable to use Bark-converted F3 for both the back-front and the high-low dimension.

The most popular vowel-extrinsic normalisation method among sociolinguists is probably Lobanov (1971). This is unsurprising given the fact that it has frequently been found to be (one of) the most efficient algorithm(s) in reducing physiological and preserving sociolinguistic variation (cf. Clopper 2009: 1440). The main drawback of Lobanov – and many other vowel-extrinsic algorithms – is that it works best when *all* vowels of a system are measured. Constraints of time and resources made this endeavour impractical for the present study. The choice fell on Watt & Fabricius (2002) in its modified version (Fabricius et al. 2009) instead, a method which is "conceptually similar" and deemed "also successful" (Clopper 2009: 1440).

Watt & Fabricius (2002) assume a triangular vowel space with the 'corner' vowels [*i*], [*a*], and [*u′*]. In RP (for which the algorithm was originally designed), these would correspond to FLEECE, TRAP, and GOOSE, but the NORM package automatically chooses the highest/most fronted and the most open vowel available in the sample as [*i*] and [*a*], irrespective of their labels. Obviously, [*i*] and [*a*] should be relatively stable in the variety under scrutiny (cf. Watt & Fabricius

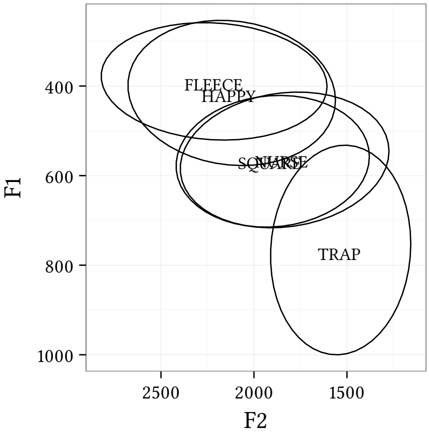
2002: 163). Since I am not aware of any evidence that suggests this is *not* true for FLEECE and TRAP in Scouse, these two were used as corners in this study. From these benchmark vowels, a 'centroid' S or "centre of gravity" (Watt & Fabricius 2002: 164) is then computed as follows:

$$S(F_n) = \frac{[i]F_n + [a]F_n + [u']F_n}{3} \tag{5.2}$$

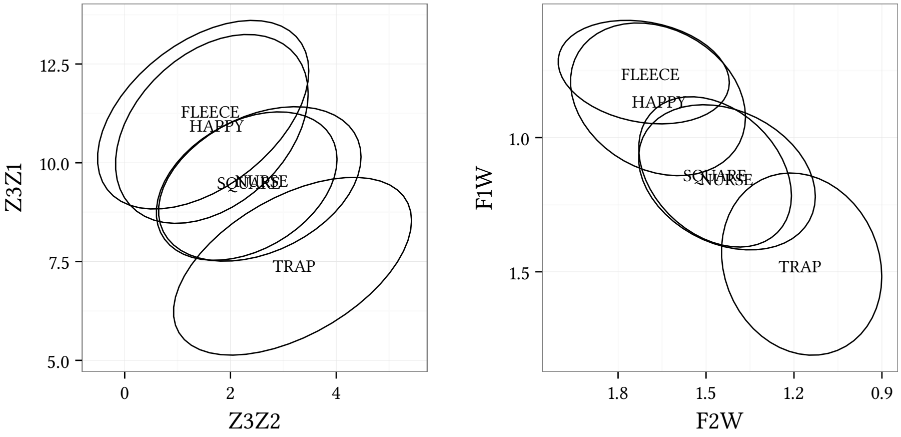
Where F_n is a mean raw formant value of the corner vowels $[i]$, $[a]$, and $[u']$. The centroid value $S(F_n)$ is computed separately for each formant, and normalised values are then expressed as the ratio of the raw measurement to the corresponding centroid: $\frac{F_n}{S(F_n)}$. Note that $[u']$ is not measured, but derived from $[i]$, assuming that $[u']F_1 = [u']F_2 = [i]F_1$. As a result, only FLEECE and TRAP have to be measured. To counter potential skewing due to the fact that TRAP might not be exactly halfway between FLEECE and GOOSE with respect to frontness, $[a]F_2$ is also derived instead of measured in the modified version of the algorithm employed in this study (cf. Fabricius et al. 2009: 420–421; Kendall & Thomas 2009).

With respect to the tests applied to assess the power of the normalisation algorithms, this study largely follows Langstrof 2006. One criterion for determining whether a normalisation process was successful is the degree to which it has reduced variance *within* categories and overlap *across* categories. In our case this would mean that phonemes should be more distinct and that scatter around phoneme means should be reduced in the normalised data. When we look at Figure 5.5b this is not immediately obvious. It should be borne in mind that with respect to NURSE and SQUARE we are talking about a merger for most speakers so we would not necessarily expect these two phonemes to appear more distinct in normalised data. The third vowel under scrutiny here, happY, however, should be more distant from both NURSE and SQUARE in the normalised data. At least for the Bark-difference method, the graph does not suggest that it really is.

It does not look as if scatter for any of the variables had been reduced either. If anything, scatter around the mean seems to have increased, particularly in the front-back dimension. The scatter plot for the Watt & Fabricius normalised data looks a lot more promising. Scatter in both dimensions seems to have been (slightly) reduced, and the phonemes appear to be more distinct. But then again, we are using different scales in the three representations so a purely visual inspection is insufficient. We will thus have a look at variation coefficients next. I am well aware of the fact that the measure of dispersion most commonly used in such cases is the standard deviation. Since we have very different means in our



(a) raw data



(b) Bark difference normalised

(c) Watt-Fabricius normalised

Figure 5.5: Vowel distributions (all subjects pooled)

samples due to different scales, however, comparing standard deviations would be highly misleading as they depend on the means of the samples. We will therefore use variation coefficients which normalise standard deviations by dividing them by the mean of the sample. These variation coefficients can then be meaningfully compared to each other.

Table 5.2 shows a rather mixed picture. While the Bark difference algorithm was successful in reducing variance in the F1 dimension for all three vowels, it ac-

Table 5.2: Variation coefficients for raw and normalised data

	F1			F2		
vowel	raw	Bark	Watt	raw	Bark	Watt
happY	0.187	0.107	0.167	0.120	0.506	0.089
NURSE	0.127	0.101	0.115	0.139	0.356	0.104
SQUARE	0.123	0.101	0.115	0.123	0.363	0.084

tually *increased* variance of F2 considerably for all vowels involved. We thus cannot clearly claim that the Bark difference normalisation reduced within-vowel variance overall. Watt & Fabricius fares a lot better. Reduction in variance for F1 is systematic, if only marginal and slightly less successful than Bark-difference. When we look at F2, however, we see a clear improvement. Again, the reduction in variance is not huge, but Watt & Fabricius does reduce variance systematically across all three vowels, whereas Bark-difference actually *in*creases variance dramatically. On the whole, then, Watt & Fabricius seems to do a better job than Bark-difference in reducing intra-category variance.

Next, we look at Pillai scores as a measure of distance between distributions, in our case of vowel discreteness. Pillai scores were first used in a linguistic study by Hay, Warren, et al. (2006) and their usefulness for sociolinguistic investigations is discussed in Hall-Lew 2010. They are considered to be superior to simple Euclidean distance measures because a Pillai score "takes account of the degree of overlap of the entire distribution" (Hay, Warren, et al. 2006: 467). Pillai scores range between 0 and 1, with values close to 0 (and an accompanying high p-value) indicating a large degree of overlap between the distributions, and values near 1 (and a low p-value) representing distributions that are (almost) completely distinct. The pairing NURSE-SQUARE is not represented in Table 5.3 as these two vowels participate in a merger for most speakers and it can therefore not be assumed that they *should* be distinct in the first place.

Table 5.3 confirms the impression we already got from Figure 5.5: the Bark difference normalisation does not distinguish happY from both NURSE and SQUARE better. We also see, however, that it does not really fare worse. Pillai scores for raw and normalised data are very similar and p-values are close to 0 in both cases. So while the Bark difference normalisation did not make different phonemes appear more distinct, it did at least not result in significantly *less* distinct categories either. Watt & Fabricius, on the other hand, increases the distinctness of happY to both NURSE and SQUARE. Only slightly so for the former, it has to

Table 5.3: Pillai scores for total and female/male vowel distributions

	raw		Bark		Watt	
vowel(s)	Pillai	p-value	Pillai	p-value	Pillai	p-value
happY-NURSE	0.486	< 0.001	0.460	< 0.001	0.550	< 0.001
happY-SQUARE	0.372	< 0.001	0.323	< 0.001	0.425	< 0.001
happY (female/male)	0.446	< 0.001	0.090	< 0.001	0.020	< 0.001
NURSE (female/male)	0.570	< 0.001	0.229	< 0.001	0.131	< 0.001
SQUARE (female/male)	0.659	< 0.001	0.185	< 0.001	0.115	< 0.001

be said, but quite clearly for the latter. As is obvious from the table, p-values are extremely low for all pairings (often too small to be treated as different from zero by the R software), which is in all likelihood simply due to the fact that they are based on comparatively large datasets where even small differences will show up as (highly) significant. When we look at the three vowels individually and compare female to male realisations, we see a clear 'improvement' in the normalised data. For all three vowels, female and male realisations are quite distinct in the raw data, a fact which is reflected in Pillai scores which are comparable or actually higher than those for different phonemes, and p-values that indicate highly significant differences.

All Pillai scores for the Bark difference normalised data are lower than their counterparts calculated for the raw data. Intra-category overlap has increased by 0.356 for happY, 0.34 for NURSE, and 0.474 for SQUARE. This means that female vowels are more similar to male vowels in the normalised data, which is precisely what we would expect a vowel normalisation process to achieve. Female and male distributions are still significantly different from each other in the normalised data, but, (a) again, this could simply be due to the fact that we have a comparatively large data set where even small differences will come out as significant (cf. the very low Pillai score of 0.09 for happY), and, more importantly, (b) we do not want *all* differences to be filtered out, but only the *physiological* ones.

All three vowels are test vowels in this study and we expect to see at least some gender differences which are due to sociolinguistic rather than physiological reasons. This is particularly true for the NURSE-SQUARE merger, a sociolinguistic variable which is considered to be highly salient (cf. Chapter 3) and with respect to which we would therefore not be surprised to see women and men behave differently. happY, on the other hand, is believed to be non-salient (again, cf. Chapter 3), so gender differences are less likely here. This is exactly what the figures in Table 5.3 suggest: a very low Pillai score for happY, indicating almost complete overlap of the distributions (although this small difference is still significant) and considerably larger ones for NURSE and SQUARE, meaning that women and men differ in their realisations of these vowels even *after* normalisation. It thus appears as if the Bark difference normalisation had (largely) eliminated physiological variation but maintained sociolinguistic one, which is just what we want a useful normalisation procedure to do.

The values for our other candidate, however, are even better. Watt & Fabricius increases intra-category overlap for happY by 0.426, for NURSE by 0.439, and for SQUARE by 0.545. The remaining differences between female/male distributions are also less (though still highly) significant in the Watt & Fabricius normalised data. It is possible that Watt & Fabricius has in fact eliminated some information that we are interested in, namely the assumed sociolinguistic gender difference for NURSE and SQUARE. However, the figures show that although the gender distributions are less distinct than in the Bark-difference normalised data, they are still (and by a much larger factor than in the Bark values) more distinct than the normalised female/male distributions of happY. Also, it could very well be that the smaller differences we find for NURSE and SQUARE in the Watt & Fabricius normalised data are simply the more realistic ones. Both algorithms thus create a pronounced relative difference between the female/male Pillai scores of happY and those of NURSE and SQUARE, but in addition Watt & Fabricius decreases inter-phoneme overlap more, so the conclusion with respect to Pillai scores is again that Watt & Fabricius seems to be the preferable choice.

Euclidean distance measures can also be used to determine the usefulness of normalisation procedures. Once again, we cannot directly compare euclidean distances because of different scales. What we can do, however is look at a ratio that is calculated as follows:

$$\frac{d_k}{d_l} = \frac{\langle\sqrt{(F_{1i} - \bar{F}_{1j})^2 + (F_{2i} - \bar{F}_{2j})^2}\rangle}{\langle\sqrt{(F_{1i} - \bar{F}_{1i})^2 + (F_{2i} - \bar{F}_{2i})^2}\rangle} \tag{5.3}$$

Where F_{1i} and F_{2i} are the F1 and F2 values of tokens in phoneme category ‘i’, $\bar{F}_{1i}$ and $\bar{F}_{2i}$ are the mean values of category ‘i’, and $\bar{F}_{1j}$ and $\bar{F}_{2j}$ are the mean values of category ‘j’. d_k is then the average distance of tokens in category ‘i’ to the mean of category ‘j’ (e.g. happY tokens to the mean of NURSE), and d_l is the average distance of tokens in category ‘i’ to their own mean (e.g. happY tokens to the mean of happY). This ratio should always be greater than 1, i.e.

the average distance to the mean of *another* category should be bigger than the average distance to the mean *within* the category. A successful normalisation procedure would have to increase this ratio since both intra-category spread and inter-category overlap should be diminished.

Table 5.4: Euclidean distance ratios for raw and normalised data

vowels	raw	Bark	Watt
happY-NURSE	1.574	1.599	2.326
happY-SQUARE	1.429	1.533	2.187
NURSE-happY	1.694	1.741	2.320
SQUARE-happY	1.579	1.786	2.355

Table 5.4 shows that, while the difference is marginal for the pairing happY-NURSE, all euclidean distance ratios are higher for the Bark difference normalised data than for the raw data. Normalisation using this method was therefore an improvement. Yet the figures for Watt & Fabricius are, once again, even better. While the distance ratio increases on average only by about 0.096 for the Bark-difference normalised data, Watt & Fabricius produces distance ratios that are on average 0.728 higher.

With the current dataset Watt & Fabricius thus yields better results than the Bark-difference method in visual representation (scatter plots), reduction of inter-category variation coefficients, Pillai scores for inter- and intra-category (gender) comparisons, and euclidean distance ratios. Despite the fact that the Bark-difference normalisation is presumably more plausible in perceptual terms, the Watt & Fabricius algorithm will therefore be used in this study whenever normalised vowel values or plots are reported or represented.

5.5 Phonological context

In order to extract the phonological context of the variables under scrutiny orthographical representations of the carrier word and the one following the carrier word were extracted from the transcripts. These orthographic representations were then automatically replaced (in R) by a phonemic transcription that was gathered from the interactive web-based CELEX lexicon database (Baayen et al. 1993). With regard to those transcriptions, CELEX allows the user to choose from four different character sets. For this study, the DISC set was selected because it represents each English phoneme with a single character (vowel length is not coded separately as this bit of information is already included in the vowel quality – in English!). This is highly useful if the transcriptions are going to be processed automatically, as diphthongs and affricates (which are regarded as single phonemes) are represented by a single character. It is therefore impossible to misinterpret the first element of a diphthong (or and affricate) as a simple monophthong (or plosive). This is a crucial advantage, as the software often has no straightforward way of knowing whether it is faced with a diphthong or a sequence of monophthongs. DISC uses a set of simple ASCII characters. In Table 5.5 the characters that are different from IPA are listed together with their IPA equivalents.

Table 5.5: DISC characters and IPA equivalents

consonants		monophthongs		diphthongs	
IPA	DISC	IPA	DISC	IPA	DISC
ŋ	N	ɪ	I	eɪ	1
θ	T	ɛ	E	aɪ	2
ð	D	æ	{	ɔɪ	4
ʃ	S	ʌ	V	əʊ	5
ʒ	Z	ɒ	Q	aʊ	6
ʧ	J	ʊ	U	ɪə	7
ʤ	_	ə	@	ɛə	8
ŋ̩	C	iː	i	ʊə	9
m̩	F	ɑː	#		
n̩	H	ɔː	$		
l̩	P	ɜː	3		

The impact of neighbouring sounds on the test variables *will* be investigated to provide a more complete picture of how these sounds are used in Liverpool. However, since the focus of this study is clearly on independent variables that are social in nature, this part of the analysis will be rather basic. Only the immediately preceding and the immediately following phonemes are considered. In the case of the two consonantal variables, measurements were furthermore restricted, from the start, to cases where /ŋ(g)/ and /k/ occurred either intervocallically or at the end of a word, because these contexts have been identified as the ones where lenition is most likely to occur (cf. §3.3.2). The three test vowels were only measured in content words (and, for NURSE and SQUARE, also exclusively in stressed syllables), both to keep the number of vowels that had to be measured manageable and in order to avoid introducing unnecessary noise into the dataset by including weakened vowels.

All but two measured vowels either occurred at the beginning of a stretch of speech or were preceded by a consonant. NURSE and SQUARE tokens, without exception, either were the last phoneme in a stretch of speech or were *followed* by a consonant, too. Only happY had a sizeable proportion of observations where the test vowel was followed by another vowel. The difference between happY measurements followed by a consonant and those followed by a vowel was small but significiant with respect to the normalised F1 dimension (t(2016.295) = -13.593, p < 0.001), but insignificant as far as the (sociolinguistically more important, cf. Labov 2006: 502) F2 dimension is concerned (t(1935.124) = -1.355, p = 0.176). It was therefore decided to drop happY tokens that occurred before another vowel (along with *any* observation where a test vowel was followed or preceded by silence) when fitting mixed linear effects regression models (see §5.6) because this allowed me to use the same set of phonological predictors (place and manner of preceding and following consonant) for all three vocalic test variables and thereby improved comparability of the models. Measurements of happY still accounted for the largest share of total observations, and in figures and other statistical comparisons (t-tests), the complete data set (including happY followed by another vowel) was used. Of course, this meant that it could not be investigated whether happY formants might be influenced by vowel harmony. This is an interesting question, but given the focus of the present analysis, it was considered outside of the scope of this book anyway, and will have to be addressed in a separate study in the future. Word frequencies were considered for all variables investigated, and operationalised using Zipf scores based on occurrences in SUBTLEX-UK (Van Heuven et al. 2014). See §10.1.1 for a more detailed discussion.

5.6 Statistical analysis

Mixed linear effects models have become a sort of gold standard in recent years, especially in subdisciplines like psycho- and neurolinguistics. Their biggest advantage is the possibility to include so-called random effects. The reasoning behind this is that in most common experimental designs, we have "fixed effects" and "random effects". Fixed effects are the variables the experimenter is primarily interested in and which are, as a consequence, controlled for in the experiment. They are theorised to have the same or a similar impact in the sample that is the basis for the experiment as in the total population which the sample is drawn from. Random effects, on the other hand, are responsible for variation that is not part of the experimental design, but due to the particular sample. As a result, the effects of random factors cannot be extrapolated to the population as a whole (cf. Barr et al. 2013).

In Mixed linear effects models, the impact of random factors is estimated and taken out of the data before the relevance of the fixed effects is calculated. The result is a reduction in noise since variation that is supposed to be due to chance is filtered out. As a common example, consider a hypothetical lexical decision experiment where subjects have to decide whether a particular string of sounds or letters is a word of their language. The words that are presented fall into two intrinsically different groups (e.g. different word class, length, complexity,…) and the experimenter is interested in whether reaction times for these two word groups differ. In such an example it is often found that individual words produce generally higher or lower reaction times across subjects (e.g. due to a non-controlled factor such as frequency or number of similar words in the language). The experimenter, however, is not interested in the effect of particular words but only in the general effect of the group they are part of. The actual words chosen for the experiment are, in this case, considered a random sample of the whole group (→ population). The same goes for the sample of participants, as some people are generally faster or slower to respond than other subjects. Both sources of variation are 'random' because re-running the experiment would (or at least *could*) involve choosing a different sample of words and a different sample of participants (cf. Barr et al. 2013: 259–260). It thus makes sense to filter out variation that is due to individual differences between subjects and test words as it is a characteristic of the sample, and not considered representative of the population.

It was thoroughly considered whether subject and carrier word – the two most straightforward options – should be entered as random factors in an analysis of the production data presented here. Especially with respect to word, this would make some sense. After all, there is no control over which words subjects use the relevant variables in (at least in the free speech part of the interviews which makes up the vast majority of observations). Treating carrier word as a random factor was still deemed problematic, however. This is because the frequency of the carrier word, as well as the sounds directly preceding and following the target sound are factors of theoretical interest here. While word itself could be considered a random factor in this research design, it seems likely that filtering out word effects would also eliminate a lot of *relevant* information that is coded in the variables 'preceding sound', 'following sound', and 'word frequency', as these bits of information (among others) are included in the overall word context of the observation. In the end, the risk was considered worth taking in order to counteract a scenario where (highly frequent) individual words would otherwise unjustifiably dominate the sample, and – possibly – obscure or overlay any more general effects of frequency or phonological environment.

Treating interviewee as a random effect is even more of an issue. As explained above, the reasoning behind treating subject as a random effect in many psycholinguistic experiments is that the group of people that actually took part is a random subset of the population one wants to extrapolate the results to and that individual differences are therefore noise. This crucial assumption, however, is not met in the dataset that is analysed in this chapter. There was no active a priori selection of participants (cf. §5.2) in terms of typicality etc. Nonetheless, the participants that ended up in the sample *are* considered to be representative of their social group. We look at a comparatively small number of middle- and working-class (female/male, old/young…) speakers and analyse their speech because we believe our results *can* be generalised to the group as a whole (at least to a certain extent). This is an essential tenet of any sociolinguistic analysis and argues against treating participant as a random effect.

It is possible to calculate random effects for speaker sub-clusters, e.g. for young working-class women only. This would eliminate the theoretical problem just outlined, as the variation between, say, young working-class women and young working-class men would not be filtered out, but just the differences between individuals *within* the respective sub-groups. This course of action was still rejected, because (a) for the relevant sub-groups (divided by gender and social class) among the young and middle-aged subjects this would mean filtering out the variation between two participants only (which does not really seem worth-while), and (b) more importantly, there is only one subject each in the gender/social class subgroups for the oldest speakers, so there is no other subject to estimate any potential effect of the individual against.

In summary, there are both conceptual and practical problems if one is to consider speaker and/or carrier word as random effects in the production data under scrutiny here. The use of carrier word as a random effect seemed to be more acceptable, though, since this might, in fact, make the results somewhat more representative and comes with less severe downsides. A random intercept for carrier word was therefore included in all mixed-effects models that will be reported on. Sum coding was used for all these analyses so that main effects and interactions (instead of *simple* effects and interactions) could be identified. For the vocalic variables, the set of main predictors entered into the maximal model was: style, age group, gender, social class, frequency, vowel duration, place of articulation (preceding sound), manner of articulation (preceding sound), place of articulation (following sound), and manner of articulation (following sound). Style is the independent variable I am most interested in as the presence or absence of style shifting is taken as an indicator of salience (cf. Chapter 4). It is quite possible (and actually expected given the main hypothesis of this study) that style differences can be present in one group but lacking in another (or be present in all groups, but not to the same extent). To test for this (and other, sociolinguistically meaningful combinations), all two-way interactions of style, age group, gender, and class were included as well, along with the two three-way interactions of style, age group, and one of the other social variables gender and social class. Interactions of the phonetic-phonological factors were not considered, as these predictors are not of primary interest in this study, and adding their interactions would have unduly inflated the models.

Model structure for the two consonantal variables velar nasal plus and /k/-lenition was identical as far as the social predictors are concerned (both in terms of main effects and interactions). Frequency of the carrier word was also included, but the set of phonetic and phonological predictors had to be different. Firstly, vowel duration is not applicable to consonants (plus the timing domain is already included in the dependent variable – proportional duration of friction), so this factor was not relevant for the mixed-effects regression models that were fit to the /ŋ(g)/ and /k/ measurements. Secondly, the phonological context had been restricted to intervocalic and word-final occurrences from the start, so it was considered unnecessary to enter information in the same way as it had been done for the two vocalic variables (i.e. 'spread out' over four different independent variables). Instead, phonological environment was summarised in a single predictor ("Environment" in the spreadsheet), which was to code whether the measurement had been taken in an intervocalic context (within a word) or at the end of a word. The second context was further divided with respect to whether the measurement was followed by silence (pre-pausal), or by another word, in which case the type of the first sound in the following word (vowel, affricate, liquid…) was coded.

All statistical test were performed using the R software (R Core Team 2015). Mixed linear effects models were computed with the help of 'lmerTest' (Kuznetsova et al. 2015), an R package which builds on 'lme4' (Bates et al. 2015), but adds p-values calculated on the basis of F statistics, with degrees of freedom derived from Satterthwaite's approximation. Sum coding, instead of R's default treatment coding, was used for all regressions. Model selection was based on AIC scores and F-tests comparing nested models. Calculating a simple goodness-of-fit measure is not a straightforward task in the context of mixed-effects models. As a rough (!) equivalent of the R^2 value known from linear regression models this book reports the R^2 of a linear model that regresses the observed values on the fitted ones from the linear mixed-effects model (cf. r-sig-mixed-models mailing list 2015). Models were checked for collinearity using the kappa.mer and vif.mer functions written by Austin Frank[4].

[4] code downloadable from hlplab.wordpress.com/2011/02/24/diagnosing-collinearity-in-lme4/

Sound change, priming, salience

This volume investigates the realisation and perception of four phonological variables in Liverpool English (Scouse), with a special focus on their sociolinguistic salience. Younger speakers' speech is found to be more local, but only for the two salient variables in the sample (NURSE-SQUARE and /k/ lenition), which appear to carry considerable amounts of covert prestige. Local variants of non-salient happY-tensing and velar nasal plus, on the other hand, are actually found to be receding, so at least to a certain extent Scouse also seems to be participating in regional dialect levelling.

The importance of salience is also obvious in the perception data, with only the two highly salient stereotypes generating robust effects in a social priming experiment (albeit in the unexpected direction). These results indicate that the investigated variables differ measurably not only in their use in production, but also in terms of how central they are to mental sociolinguistic representations of Scouse. They also tell us more about the way we process, store, and (re-)use sociolinguistic variation in perception. By defining likely contexts for significant priming effects they might finally even help in coming up with a more elaborate 'theory of priming' in the realm of sociophonetics.

ISBN 978-3-96110-119-1

9 783961 101191