A corpus-based study of pragmatic markers in London English

EIVIND NESSA TORGERSEN, COSTAS GABRIELATOS, SEBASTIAN HOFFMANN, and SUSAN FOX*

Abstract

This article analyses the use of particular pragmatic markers in two corpora of spoken London English: the Linguistic Innovators Corpus (LIC) and the Corpus of London Teenage Language (COLT). We found variation according to sex, ethnicity and geographical location, with a different distribution for each pragmatic marker. The innovative pragmatic marker you get me was most frequent among male non-Anglo Hackney residents, indicating innovation in inner London. We argue that a number of pragmatic markers, most notably you get me, should be regarded as elements of Multicultural London English, along with other features that have already been documented.

Keywords: sociolinguistics, pragmatic markers, language and ethnicity, linguistic innovation, spoken language, Multicultural London English

1. London as centre for linguistic innovation

London has long been regarded as the centre of linguistic innovation in British English, particularly for phonological features (Wells 1982), but also for morpho-syntactic features (Nevalainen and Raumolin-Brunberg 2003). In line with this view, Cheshire at al. (2008a) found that adolescent speakers in an inner city area of London used a number of phonological, discourse and morpho-syntactic innovations. Phonological innovations include extreme fronting of the GOOSE vowel (Cheshire et al. 2008a) and un-shifting of PRICE, FACE and GOAT (Kerswill et al. 2008b). For discourse features there was innovation in the quotative system, with inner London young speakers using a previously undocumented quotative construction, this is (speaker), in personal narratives to introduce reported speech (Cheshire and Fox 2007a). Morpho-syntactic innovation was demonstrated in an analysis of was/were (Cheshire
which showed that the inner London young speakers did not use the expected pattern of was/were levelling typical of the outer-city young speakers, or of other varieties in the UK, such as in Reading (Cheshire 1982) or York (Tagliamonte 1998). The expected levelling pattern would be: was in positive polarity and weren’t in negative polarity; instead, we find levelling to was and wasn’t, respectively. Furthermore, Gabrielatos et al. (2010) examined the use of indefinite articles in London with similar results; the innovative forms had the highest use among the inner-city young speakers.

The speakers who seemed to be leading the changes had been born in London, but were of recent immigrant background. Multi-ethnic friendship networks also seemed to play a key role; the London data and the results of previous studies (Khan 2006; Fox 2007) have shown that networks are the conduit for the spread of ‘ethnic’ features to majority groups. The outcome of the innovations found in inner-London has led to a variety, or set of varieties, termed Multicultural London English (MLE) (Cheshire et al. 2008b; Kerswill et al. 2008a). MLE speakers seem to select variants from a ‘feature pool’ (Mufwene 2001) which contains variants ranging from non-standard to standard, for example in the use of was/were (Cheshire and Fox 2009) and relative pronouns (Cheshire and Fox 2007b), but also innovative features such as the quotative this is (speaker) mentioned above.

Previous studies have revealed interesting differences with respect to the sex of innovators. In Milton Keynes, some 90 kilometres north-west of London, the innovators were shown to be the adolescent females (Kerswill and Williams 2000), while in inner-London the leaders in change were the male speakers (Cheshire et al. 2008a; Gabrielatos et al. 2010; Kerswill et al. 2008a, 2008b). We assume that these differences are due to the high degree of language and dialect contact in London, which may accelerate language change. The types of linguistic innovation that can be observed in inner-city London – and their underlying motivations – may be different from those identified in previous studies. This is because London is a more complex speech community and for this reason it may be necessary to modify existing models of linguistic change that are based on a homogeneous speech community (Labov 2001). According to these models, young children replicate the grammar and phonologies of their parents in a process of transmission, with further changes – known as incrementation – as they get older (cf. Labov 2007). However, it is clear from the results of the London studies that speakers of recent immigrant backgrounds (we use the term ‘non-Anglo’ for these speakers) have a clear influence on the speech of those with an established London background (Anglos), which has, for example, been documented for changes in diphthongs (Kerswill et al. 2008b). This has an effect on the realisation of specific linguistic features. Anglo speakers acquire these new features as members of multicultural friendship groups. Kerswill and Torgersen (2009, forthcoming) show that dialect
contact is important in modelling causation and progress of language change, thus contradicting Labov (2001: 20), who rejects this in his model and instead focuses on “changes that emerge from within a linguistic system”. They argue that the lack of transmission of traditional vowel features is due to the large degree of in-migration to London. This is because the non-Anglo speakers bring with them influences from other languages or other varieties of English inherited from their parents and grandparents, and it is this multilingual influence that seems to be important in the process of language change. Kerswill and Torgersen also identify the male non-Anglo speakers as the leaders of change.

This article investigates whether a high-contact community will display the same patterns for discourse features. Tagliamonte and D’Arcy (2009) found that female speakers in Toronto were leading the changes in some morphosyntactic and discourse features, in line with Labov’s model. However, Gabrielatos et al.’s (2010) findings for indefinite articles suggest that we may find different patterns in London. The use of discourse features may also reveal linguistic developments at an early stage of the change process (Andersen 2001; Tagliamonte and D’Arcy 2004). Therefore, we will investigate the use of pragmatic markers (henceforth, PMs) as previous research has shown they are particularly frequent in spoken language (Holmes 1982, 1995; Stenström et al. 2002). We are interested in possible differences and variation in the use of PMs between speakers belonging to groups defined in terms of sex, ethnicity, place of residence and friendship network. We are especially interested in whether the use of certain PMs can be linked to linguistic innovation and, if so, who the linguistic innovators are, and where in London they are found.

2. Pragmatic markers and spoken interaction

According to Brinton (1996: 30), PMs are items whose function is to “express the relation or relevance of an utterance to the preceding utterance or to the context”. The main characteristics appear to be that they are found outside of, or are loosely attached to, the syntactic structure, and are a more or less open class of items often found in spoken language (Brinton 1996: 33–35). Different terms exist to cover more or less the same range of items (Schourup 1999: 228–230). Some researchers label them discourse markers and include items such as you know (e.g., Schiffrin 1987; Erman 2001), whereas Fraser (1996, 2006) is quite restrictive and only includes connectors such as and and so. PMs, as described by Andersen (2001: 22), focus on aspects of communication and speaker attitudes and evaluation. They facilitate pragmatic processes establishing what the speaker wishes to convey, i.e., how the listener should identify the intended meaning depending on the context. We regard the most important point to be that all of these items share a similar function in discourse.
For example, Schiffrin (1987: 326) concludes that discourse markers (as she
defines them) “function as contextual coordinates of talk in that they are used
to index an utterance to the local context in which the utterance was produced
and interpreted”. Erman’s discussion of the use of *you know* in adolescent con-
versations (2001) and in conversations between men and women (1992) ex-
plores the “potential participation transition” (2001: 317) between speaker and
hearer and between prior and upcoming content, and is therefore clearly fo-
cused on interaction. It is PMs sharing this function that we will focus on in
this article. We find that the overarching function of these PMs is to (appear to)
involve the interlocutor by (appearing to be) eliciting responses indicating that
the interlocutor agrees with, remembers, understands or follows the thread. In
this way they might not be PMs in the sense that they mark stages or changes
of direction in the discourse, but they do mark interactional attitudes; i.e., they
indicate speaker-sanctioned places in the discourse where the interlocutor can
comment (see examples 1–4).

1. Maria: well obvious . they they caught . no . my parents like .
sometimes I I like if I wanted to go raving right? I would not go jump
out of the window because . my parents will find out you get me my
mum goes to work right?

2. Brian: faces like and acting hard and all that and when they was like
when they was outside like they was calling their olders yeah so we was
calling our olders then they they start erm trying to move to us innit like
beat us up and that but they didn’t know that we had guns innit and so
the olders start busting at them and they they hit they start running and
then they stopped they hit one of them and then

3. Kate: he makes you laugh but he’s just annoying . if you know what I
mean but . . he’s just . he will come . behind me once yeah?

4. Interviewer: so what are you doing in the college at the moment?
Ahmed: I’m coming to see (name=Mark) and (name=Tina) innit
Mark: he’s wandering

3. Pragmatic markers in London English

Previous research on PMs in spoken London English has used data from the
*Corpus of London Teenage Language* (COLT). Stenström et al. (2002) investi-
gated the use of invariant tags, such as *yeah, innit, eh, right* and *ok*, quantita-
tively and qualitatively. The authors include frequency counts per thousand
words for groups of speakers according to age bands (such as preadolescents
and adults), but also include tag use for speakers whose social variables could
not be verified. Stenström et al. (2002) find differences between the older and
younger speakers in the data: there was an increase in the use of *right, innit* and
yeah from preadolescence until young adulthood, while adult speakers displayed a substantially lower use of these three tags. This is in line with Tottie and Hoffmann’s (2006: 304) claim that invariant tags are more frequently found in young people’s speech, probably because this age group uses fewer canonical tag questions than older speakers. Erman (2001) has examined you know in COLT and argues that it is undergoing a change in its meaning and function: young speakers use you know to ensure listener involvement and to link pieces of discourse together. Therefore, it functions as a “social as well as a metalinguistic monitor” (Erman 2001: 1356), and is oriented towards the activity of communicating. Erman did not examine sex differences, but Macaulay (2002) found that females used the PM you know more than males in two corpora of Scottish English.

Since COLT contains spoken data from the 1990s, it may no longer give a current picture of PM use in London English. In the present paper we therefore also make use of a more recently compiled corpus, the Linguistic Innovators Corpus (LIC), which contains spoken data collected in 2005. Social information is available for all speakers, who also represent comparable social groups, i.e., they are all working class. By comparing the young speakers in the two corpora we can also study change in real time (see also Gregersen and Barner-Rasmussen, this issue). We will examine the use of nine PMs; the invariant tags right, innit, ok and yeah studied by Stenström et al. (2002), you know examined by Erman (2001), the multi-word PMs you know what I mean, if you know what I mean and do you know what I’m saying and the PM you get me recently identified in Cheshire et al. (2008b). We hypothesise that there will be an increase in the use of innit, yeah and right in LIC as compared to COLT and that the male non-Anglo speakers will have a higher use of the emerging PM you get me than the male Anglo speakers, in the same way as the male non-Anglo speakers have the highest use of innovative vowel features (Kerswill et al. 2008b) and use of indefinite article a before vowel-initial words (Gabrielatos et al. 2010). In general, we expect a higher use of innovative PMs in inner-city than in outer-city locations (see Section 4.1 below). While an increase in the use of some PMs is predicted by Labov’s model (Labov 2001), because of the process of incrementation, the model does not predict how new forms are acquired, i.e., how innovative forms spread. We believe that examining the use of PMs in London English will shed further light on the process of linguistic innovation and contribute to models of linguistic change.

4. Corpora and methodology

The combination of quantitative and qualitative sociolinguistic analysis of PMs in spoken language has not been very common among traditional
sociolinguists, probably because the high frequency of PMs requires extensive manual analysis. However, corpus linguists started including sociolinguistic data in their studies about twenty years ago (for overviews see Baker 2010; McEnery and Wilson 2001: 115–117) – although such studies still represent a minority of work in both corpus linguistics and sociolinguistics. Corpus linguistic tools enable researchers to automatically extract tokens from a corpus (irrespective of its size), together with other encoded information about the speaker, for further manual analysis (see Section 4.2). In addition, corpus linguistics and sociolinguistics approaches are compatible and complementary (see Gabrielatos et al. 2010: 304).

4.1 Corpora

This study uses two spoken corpora, both collected in London: the Linguistic Innovators Corpus (LIC) (Gabrielatos et al. 2010) and the Corpus of London Teenage Language (COLT) (Stenström et al. 2002). LIC contains data collected for the project Linguistic Innovators: the English of adolescents in London (Cheshire et al. 2008b). More specifically, LIC comprises 1.4 million words of transcribed sociolinguistic interviews with 100 working-class adolescents (16–19 years old) and 18 working-class elderly speakers (aged 70 or above) in two boroughs, Hackney and Havering. Hackney is an inner London borough in the traditional East End, while the borough of Havering is in the east, and was formerly part of Essex (see Figure 1). The ethnicity of young LIC speakers was self-defined, and speakers were assigned to one of two groups: Anglos and non-Anglos. The group of ‘Anglo’ speakers was homogeneous overall, as it comprised speakers who self-identified as ‘white British’, and who were from families with relatively deep local roots. The speakers in the ‘non-Anglo’ group were children or grandchildren of immigrants, almost all from developing countries. Speakers were also asked to name their closest friends, whether they were male or female, whether they knew them from the college they attended, and their ethnic background. Each speaker was then assigned a score of 1–5 depending on the ethnic makeup of his/her friendship network:

1 = all friends same ethnicity as self
2 = up to 20 percent of a different ethnicity than self
3 = 21–40 percent of a different ethnicity than self
4 = 41–60 percent of a different ethnicity than self
5 = 61–80 percent of a different ethnicity than self

The network score analysis revealed that for Anglo adolescents in Hackney a score of 3 was the minimum, whereas for those in Havering 3 was the maximum.
COLT was compiled with the aim of recording the vernacular of London teenagers from different social backgrounds. Informants (called ‘recruits’) were given tape recorders and recorded themselves in conversations with friends at home and in the street and also at school both inside and outside of the classroom. All recordings were made in 1993 and the transcribed corpus contains approximately 500,000 words. The recruits were 13–17 years of age, with two thirds being 13–14. There are in total 30 sets of recordings (i.e., from 30 recruits). Even though there were twice as many male as female recruits, the number of words in COLT is about equal for both sexes. COLT recruits represented three different social groups, and came from three boroughs in inner London (Hackney, Tower Hamlets, Camden), one borough in outer London (Barnet) and from nearby Hertfordshire (see Figure 1). There is little information about the ethnicity of the recruits – there is only a broad classification of white and ethnic minority speakers.

When contrasting the features of LIC and COLT it becomes evident that the two corpora are not fully comparable:

a) COLT contains the speech of the friends, family and teachers of the recruits (Stenström et al. 2002: 3, 19), with no consistent encoding of their age, sex, ethnicity and residence.

b) The majority of COLT recruits are slightly younger than the LIC young speakers (average ages are 14 and 17, respectively).
E. N. Torgersen, C. Gabrielatos, S. Hoffmann, and S. Fox

Table 1. LIC-2 and COLT-2: Comparison of features

<table>
<thead>
<tr>
<th>Feature</th>
<th>LIC-2</th>
<th>COLT-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of words</td>
<td>865,369</td>
<td>123,500</td>
</tr>
<tr>
<td>No. of speakers</td>
<td>100</td>
<td>19</td>
</tr>
<tr>
<td>Data collection period</td>
<td>2005</td>
<td>1993</td>
</tr>
<tr>
<td>Data collection method</td>
<td>Sociolinguistic interviews</td>
<td>Self-recordings</td>
</tr>
<tr>
<td>Age</td>
<td>16–18 (average 17)</td>
<td>12–17 (average 14)</td>
</tr>
<tr>
<td>Sex</td>
<td>female; male</td>
<td>female; male</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>Anglo; non-Anglo</td>
<td>White; ethnic minority</td>
</tr>
<tr>
<td>Residence</td>
<td>Inner London (Hackney)</td>
<td>Inner London (Hackney, Tower Hamlets, Camden)</td>
</tr>
<tr>
<td></td>
<td>Outer London (Havering)</td>
<td>Outer London (Barnet)</td>
</tr>
<tr>
<td>Social class</td>
<td>Working class</td>
<td>Working class</td>
</tr>
</tbody>
</table>

All LIC speakers are working class, whereas COLT speakers belong to both working- and middle-class social groups.

Their data-collection methods differ: sociolinguistic interviews (LIC) and self-recordings (COLT). Interviews, being more structured than self-recordings, may have elicited more standard English features, as well as a higher number of checking PMs than the unstructured conversations in COLT.

In order to make the results of comparisons between LIC and COLT more reliable, two sub-corpora were created. The COLT sub-corpus (COLT-2) comprises only the turns of the 19 working class recruits whose age, sex, ethnicity and borough of residence were known. The LIC sub-corpus (LIC-2) contains only the turns of the young speakers, while the contributions of the interviewers were discarded. The resulting sub-corpora are thus largely, albeit not fully, comparable (see Table 1; for a discussion on corpus comparability, see Gregersen and Barner-Rasmussen, this issue). Therefore, conclusions based on the results of the comparison should be understood to carry the caveat ‘to the extent that the two corpora are comparable’.

4.2 Methodology

The methodology used in this study adapts the approach used in Gabrielatos and Torgersen (2009) and Gabrielatos et al. (2010), which, in turn, was informed by the approaches used in Berglund (2000), McEnery et al. (2000), McEnery and Xiao (2004), and Rayson et al. (1997). Concordances of candidate word-forms, with a co-text of 150 characters on either side of the node, were extracted automatically, together with information on age, sex, ethnicity, place of residence and friendship network score. Concordances were then ex-
amined manually to filter out instances that did not function as pragmatic markers. When the co-text did not offer enough clues as to the function, the issue was resolved by examining the recordings for prosodic clues. In the examples (5)–(9) below, bold font indicates instances of PM use, and strike-through indicates instances excluded from the analysis, because they were not pragmatic markers.

(5) Ellie: I get panicky with one of my mum’s friends cos she’s only got .
    she’s got a baby **right** and he’s erm eight months and I won’t do it yet .
    she wants me to babysit but I won’t do it yet cos he’s too little for me .
(6) Amjad: it’s a quick way to make money but it’s not **right**
(7) Chris: I don’t look like I can talk like that .
    Interviewer: but do you think white and black people talk differently then? .
    Chris: **yeah** cos like me cos I’ve been brought up around my granddad
    **yeah**? . he’s always talked to me with his Caribbean accent . because . it
    was like a lot of it .
(8) Jess: didn’t help you really that much . and where I was quite a loud
    child . they didn’t really like me some of them so . they kinda branded
    me really . but . I did I thought it was going to be a problem **you know**
    to come to college
    Interviewer: mmm
(9) Dale: shut up I used to live in ah I was born in Harold hospital **yeah**? .
    Kieran: Harold Wood hospital?
    Dale: Harold Wood **yeah** . if you know where that is . .

In examining the extent of use of the PMs in focus, two metrics are used. The first is the number of occurrences per million words. Although this is usually termed *normalised frequency*, it will be referred to here as *frequency* for reasons of simplicity (the term *raw frequency* will be used to refer to the actual frequency of a PM in the corpus or sub-corpora). The use of normalised frequencies facilitates informal comparisons between occurrences of a feature in corpora of uneven sizes (as is the case with LIC-2 and COLT-2) – of course, any calculation of the statistical significance of the differences is carried out with raw frequencies, taking into account the size of the sub-corpora. However, the high frequency of a PM may be due to its recurrent use by a small number of speakers in a group (see also Gries 2008: 404). Therefore, the analysis also takes into account the proportion of speakers in a group who use a given PM, even if only once, which is referred to as *spread* (Gabrielatos and Torgersen 2009; Gabrielatos et al. 2010). The two metrics are summarised in Table 2. Furthermore, we will use the term *speakers* when referring to all those whose speech is included in the two sub-corpora, whereas the term *users* will refer to those who employ a particular PM (those who do not will be referred to as *non-
users). The interaction of frequency and spread for each PM will be referred to as its extent of use. The use of multiple regression analysis was not possible, as it would have to take into account all other PMs, which was not feasible within the constraints of the project.

For two of the candidate word-forms (right and yeah), LIC contained a very large number of instances; therefore random samples of 2,000 concordance lines were examined instead and the normalised frequencies given thus represent extrapolated figures. For example, the 2,000 random instances of yeah examined in this study represent 10.7% of its 18,693 instances in LIC; therefore, calculations of normalised frequencies were based on 10.7% of the word count of LIC (and the relevant sub-corpora). It must also be clarified that the sub-corpora mentioned here were not compiled and stored separately, but were created on the fly by means of regular expressions.

When comparing the two corpora, only differences that are above the 95% confidence threshold (established through the log-likelihood statistic) will be taken into account. Differences registering \( p \leq 0.05 \) will be considered statistically significant, whereas those registering \( p \leq 0.01 \) will be deemed to have high statistical significance. When differences are not significant, the frequency or spread of the groups involved will be treated as comparable. Sometimes spread differences are fairly high (e.g., 50%), but due to the very low number of speakers involved, they are not statistically significant. In such cases, the differences will be commented on, but spread values will be treated as comparable in the analysis. Because in the present study it is rare for spread differences to establish statistical significance, those that do are treated as strong indicators of diachronic changes in the extent of use.

5. Analysis

Our analysis will focus on three aspects. First, we examine short-term diachronic developments in the extent of use of the nine PMs through comparisons between LIC-2 (2005) and COLT-2 (1993) in order to ascertain the exis-
A corpus-based study of pragmatic markers in London English

5.1 Diachronic comparisons

Two PMs have comparable levels of use in the two corpora (innit and if you know what I mean), four have a higher level of use in COLT-2 (ok, right, yeah and you know), and three have a higher level of use in LIC-2 (you get me, (do) (you) know what I mean and (do) (you) know what I’m saying)3 (Table 3). Bold font indicates that both frequency and spread differences are significant. Underline and Roman indicate that only the frequency difference is significant: underline indicates a difference of at least 100%; Roman indicates a difference clearly below 100%.

Two PMs, innit and if you know what I mean, show no significant difference in either frequency or spread in COLT-2 and LIC-2 (Table 4, Figure 2). It must be noted that if you know what I mean has 50% lower frequency in LIC-2, but its frequency in either corpus is too low for that difference to reach statistical significance (LL = 0.37). On the other hand, innit has comparable frequency in the two corpora (though slightly lower in LIC-2), and almost identical spread.
Init is one of the three most widely used PMs among those examined (together with yeah and you know), but it is the only one that has not shown significant decline in use. Arguably, init can be seen as an established PM within young London speakers, a view supported by its very high spread (84%) in both corpora – the highest of all PMs examined.

Four of the PMs examined show significant decline in use (Table 5, Figure 3). Two (yeah and you know) are among the three most widely used in both corpora, and two (ok and right) among the least frequent. The most significant decrease in use is shown by ok: LIC-2 speakers display an almost twenty times
lower frequency and almost five times lower spread, with both differences being highly significant. Also substantial is the decrease in the use of *right*: its frequency is three times lower in LIC-2 (and the difference is highly significant), but its spread is comparable in the two corpora. The other two PMs (*yeah* and *you know*) have comparable spread in the two corpora, and smaller frequency decrease in LIC-2. For *you know*, the frequency difference is about 33\% (\(p < 0.01\)), whereas for *yeah* it is only 20\% (\(p < 0.05\)). In the case of *you know*, there is a small, but not significant, increase in spread. The results seem to suggest that *ok* and *right* are becoming less popular among young London speakers. However, for *yeah* and *you know* it can be argued that, despite their lower frequency in LIC-2, they can still be considered as established PMs within the population in focus. The reasons are that the frequency decrease is relatively small, and their spread is high in both corpora (at least two-thirds of the speakers use them).

Only three PMs show more extensive use in LIC-2: *you get me*, *(do) (you) know what I mean* and *(do) you know what I’m saying* (Table 6, Figure 4). All three have significantly higher frequency in LIC-2. Spread is also substantially greater; however, due to the small number of users in the two corpora, the spread difference is significant only for *(do) (you) know what I mean*. In COLT-2, none of the three is used by more than one speaker, and in LIC-2, none is used by more than one-third of speakers. These observations suggest that,
although their use became much more prominent during the period 1993–2005, they still have not been widely adopted by young London speakers.

5.2 Use by LIC-2 speakers

This section will first examine the frequency and spread of the nine PMs among LIC-2 speakers in terms of three social factors: sex, ethnicity and place of residence. The examination will be carried out in two complementary ways, by examining both the social attributes of speakers preferring each PM and the PMs favoured by individual groups of speakers. It will then examine the influence of friendship networks on the extent of use of the emerging PM, you get me.

### Table 6. Increasing PMs: Extent of use in LIC-2 and COLT-2

<table>
<thead>
<tr>
<th>PM</th>
<th>Corpus</th>
<th>Raw freq.</th>
<th>Corpus size</th>
<th>Freq./mil.</th>
<th>Users</th>
<th>Speakers</th>
<th>Spread</th>
</tr>
</thead>
<tbody>
<tr>
<td>know . . . mean</td>
<td>LIC-2</td>
<td>289</td>
<td>865,369</td>
<td><strong>334.0</strong></td>
<td>32</td>
<td>100</td>
<td>*32.0</td>
</tr>
<tr>
<td></td>
<td>COLT-2</td>
<td>13</td>
<td>123,500</td>
<td>105.3</td>
<td>1</td>
<td>19</td>
<td>5.3</td>
</tr>
<tr>
<td>you get me</td>
<td>LIC-2</td>
<td>144</td>
<td>865,369</td>
<td><strong>166.4</strong></td>
<td>17</td>
<td>100</td>
<td>17.0</td>
</tr>
<tr>
<td></td>
<td>COLT-2</td>
<td>11</td>
<td>123,500</td>
<td>8.1</td>
<td>1</td>
<td>19</td>
<td>5.3</td>
</tr>
<tr>
<td>know . . . saying</td>
<td>LIC-2</td>
<td>23</td>
<td>865,369</td>
<td>*26.6</td>
<td>4</td>
<td>100</td>
<td>4.0</td>
</tr>
<tr>
<td></td>
<td>COLT-2</td>
<td>–</td>
<td>123,500</td>
<td>–</td>
<td>–</td>
<td>19</td>
<td>–</td>
</tr>
</tbody>
</table>

Figure 4. PMs with a greater extent of use in LIC-2

although their use became much more prominent during the period 1993–2005, they still have not been widely adopted by young London speakers.

5.2 Use by LIC-2 speakers

This section will first examine the frequency and spread of the nine PMs among LIC-2 speakers in terms of three social factors: sex, ethnicity and place of residence. The examination will be carried out in two complementary ways, by examining both the social attributes of speakers preferring each PM and the PMs favoured by individual groups of speakers. It will then examine the influence of friendship networks on the extent of use of the emerging PM, you get me.
5.2.1 Extent of use of pragmatic markers by social factors

The extent of use of PMs seems to be affected by different combinations of social factors, and to varying degrees (Table 7). Arguably, the number of social factors affecting the use of a PM, and the degree to which they do, are helpful indicators of the extent to which the use a PM can be considered to be typical of speakers belonging to particular social groups.

The analysis of LIC-2 strengthens the conclusion that *innit* is an established PM among young London speakers. No differences were established for speaker residence, spread is almost identical in terms of sex and ethnicity, and frequency differences in terms of either sex or ethnicity are fairly low (70% and 50%, respectively). This is contrary to the increase we had expected; instead, its use seems to have stabilised. Combined with its comparable frequency and spread in COLT-2, the above suggests that the extent of use of *innit* is no longer characteristic of a particular group of speakers. Another interesting case is *you know*, which seems to have become an established PM among young London speakers. Although it is favoured more by the older speakers in LIC (Torgersen and Gabrielatos 2009), and has seen a moderate, but significant, decline in frequency (33%) among young speakers in the 12 years separating COLT-2 from LIC-2. It also shows a small (albeit non-significant) spread increase in LIC-2, and is a high-frequency PM among young speakers in both corpora. Therefore, the possibility cannot be dismissed that the above changes in frequency and spread are fluctuations in the use of an established PM, rather than indications of a continuing decline in use. The other declining PM, *yeah*, is more characteristic of male speakers (in both Hackney and Havering) and Havering residents in general. Taken together, the findings for *innit* and *yeah* are in line with Labov’s model: the incrementation function has reached the maximum and is now stabilising.

The emergent PM *you get me* is characterised much more strongly by the speakers’ ethnicity and residence, rather than their sex. Non-Anglo speakers (in both Hackney and Havering) have almost nine times higher frequency

<table>
<thead>
<tr>
<th>LIC-2</th>
<th>Sex</th>
<th>Ethnicity</th>
<th>Residence</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>innit</em></td>
<td>Male</td>
<td>Non-Anglo</td>
<td>---</td>
</tr>
<tr>
<td><em>ok</em></td>
<td>Female</td>
<td>Non-Anglo</td>
<td>---</td>
</tr>
<tr>
<td><em>right</em></td>
<td>Male</td>
<td>Non-Anglo</td>
<td>Hackney</td>
</tr>
<tr>
<td><em>(do) you know what I mean</em></td>
<td>Female</td>
<td>Anglo</td>
<td>Havering</td>
</tr>
<tr>
<td><em>(do) you know what I’m saying</em></td>
<td>Female</td>
<td>Anglo</td>
<td>Havering</td>
</tr>
<tr>
<td><em>if you know what I mean</em></td>
<td>---</td>
<td>---</td>
<td>Havering</td>
</tr>
<tr>
<td><em>yeah</em></td>
<td>Male</td>
<td>---</td>
<td>Havering</td>
</tr>
<tr>
<td><em>you know</em></td>
<td>---</td>
<td>Non-Anglo</td>
<td>Hackney</td>
</tr>
<tr>
<td><em>you get me</em></td>
<td>Male</td>
<td>Non-Anglo</td>
<td>Hackney</td>
</tr>
</tbody>
</table>
(p < 0.01) and three times higher spread (p < 0.05) than Anglo speakers. The same difference (and significance) was observed for Hackney speakers when compared to Havering speakers. In contrast, male speakers (regardless of ethnicity and residence) have only 50% higher frequency (p < 0.05) than female speakers – and almost identical spread. Only one other PM, ok, seems to be as strongly associated with particular social groups: female speakers have twelve times higher frequency and six times higher spread (both p < 0.01) than males, and non-Anglo speakers have six times higher frequency (p < 0.01) and four times higher spread (p < 0.05) than Anglo speakers.

The cognate multi-word PMs, (do) (you) know what I mean, (do) you know what I’m saying and if you know what I mean, seem to be characteristic of Havering residents, and the former two are also characteristic of female speakers and Anglo speakers. If you know what I mean appears to be the PM that is least associated with particular groups of speakers, as it only characterises Havering residents.

5.2.2 Focus on social factors
When we consider PM use in terms of sex, preferences seem to be polarised: four PMs are preferred by males, three by females, and only two show comparable levels of use by male and female speakers (Table 8). However, female speakers seem to have stronger preferences than male speakers; alternatively, it may be that male speakers tend to avoid what they may perceive as ‘female’ PMs. PMs characterising female speakers show highly significant frequency differences, and, in one case (ok), also a highly significant spread difference – whereas for PMs characterising male speakers differences are significant at the 5% level. Male speakers appear to prefer the emerging PM, you get me, and the high-frequency PMs, innit and yeah.

Differences in terms of ethnicity are less polarised, with the majority of PMs (five) favoured by non-Anglo speakers, two favoured by Anglo speakers, and two showing comparable use (Table 9). Similarly to male speakers, non-Anglo speakers tend to prefer the emerging PM you get me as well as OK, right, innit and you know. Anglo speakers tend to prefer the low-use PMs (do) (you) know what I mean and (do) you know what I’m saying.

Table 8. Extent of use of PMs according to sex

<table>
<thead>
<tr>
<th>Male</th>
<th>Comparable use</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>– you get me</td>
<td>– you know</td>
<td>– ok</td>
</tr>
<tr>
<td>– innit</td>
<td>– if you know what I mean</td>
<td>– (do) (you) know what I mean</td>
</tr>
<tr>
<td>– right</td>
<td></td>
<td>– (do) you know what I’m saying</td>
</tr>
<tr>
<td>– yeah</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Residence, too, shows polarised preferences (Table 10). We should also note the extensive overlap between PMs favoured by Hackney residents and non-Anglo speakers, on the one hand, and between Havering residents and Anglo speakers, on the other. Hackney residents, like non-Anglo speakers, tend to prefer the emerging PM *you get me*, the high-use PM *you know* and *right*. Havering residents, like Anglo speakers, tend to prefer two of the low-use multi-word PMs, *(do) (you) know what I mean* and *(do) you know what I’m saying*.

5.2.3 Influence of friendship networks on the extent of use of *you get me*

The friendship network score would seem to be a reliable predictor of the use of *you get me*, as the average score of users is almost 50% higher than that of non-users, and more than one-third above the average of all young speakers (Figure 5).

However, a more detailed picture emerges when network scores are conflated into two groups: high (4–5) and low (1–3). If network score is a dependable predictor of the use of *you get me*, then we would expect to find significantly more users with high scores, and, conversely, significantly more non-users with low scores. However, as Table 11 shows, only the former is the case. When looking at users (comparisons along columns in Table 11), there are 17 users with high scores, as opposed to none with low scores (*p* < 0.01), whereas among non-users, the number of high-score and low-score speakers is comparable (the small difference is not significant). This suggests that although users of *you get me* can be expected to belong to highly multi-ethnic networks, no safe predictions can be made regarding the networks of non-users. When
looking at scores (comparisons along rows in Table 11), although there are more non-users than users with a low score, there are also more non-users than users with a high score (in both cases, \( p < 0.01 \)). This suggests that, for an emerging PM like you get me, a high network score is not a dependable predictor of use, whereas a low network score is a dependable predictor of non-use.

Further support for treating apparent usage effects of networks with caution comes from the comparison of the network scores of users and non-users with reference to their ethnicity, residence and sex. If network scores indexed the use of you get me, then we would expect users of this PM – irrespective of their ethnicity, residence or sex – to consistently have higher scores than non-users. As Figures 6–7 clearly show, this is not the case for ethnicity and residence. Anglo and non-Anglo users have comparable average scores with non-Anglo non-users (Figure 6), and Havering users have a comparable score with Hackney non-users (Figure 7). Only in the case of sex is the network score of users higher than that of non-users (Figure 8).

What is more, network scores do not seem to consistently predict usage even within the same sociolinguistic group. Non-Anglo users and non-users have comparable scores (Figure 6), as have Hackney users and non-users (Figure 7). It is only male/female users that have higher scores than the corresponding non-users (Figure 8). What seems to be the case, therefore, is that network
scores are, largely, a by-product of ethnicity, as, by definition, and due to the way that ethnicity has been coded in the corpus, multi-ethnic (i.e., high-scoring) networks comprise mainly non-Anglo speakers. This can also explain the comparable scores of Hackney users and non-users, as almost three-quarters of Hackney users of you get me are non-Anglos. In the case of sex, too, ethnicity can explain the much higher scores of male and female users, as they are predominantly non-Anglo (91% and 83%, respectively). Similar conclusions were drawn in Gabrielatos et al. (2010) regarding the utility of network scores in predicting the use of indefinite article a before vowel-initial words.
6. Discussion and conclusion

Our analysis shows that the use of different PMs seems to be affected by different combinations of social factors, and the strength of each social factor varies – however, there are converging indications that ethnicity plays the most important role. Many of the already documented differences in the use of phonetic and grammatical features are linked to the ethnic composition of networks, and we have shown that the use of *you get me* is also influenced by the ethnic composition of a speaker’s friendship network for the Anglo speakers. Arguably, the number of social factors affecting the use of a particular PM, and the degree to which they do, are helpful indicators of the extent to which the use of a PM can be considered to be characteristic of a group of speakers. The comparison of COLT-2 and LIC-2 shows developments over time, with an increase in use of some PMs (*you get me*, *(do) (you) know what I mean* and *(do) you know what I’m saying*) and decline in others (*ok*, *right*, *yeah* and *you know*), while there are small differences for *innit*. There are also usage differences between inner and outer city: the emerging PM *you get me* is more frequent in the inner city borough of Hackney. We can probably trace these developments back at least 25 years and follow their development from emergence in inner-city areas to their establishment in outer areas of the city. Non-Anglo speakers in inner-city areas of London appear to be the innovators.

Reports from the 1980s indicate an innovative use of *you know what I mean* as a PM among teenagers of Afro-Caribbean background and Anglo teenagers in areas of London with high dialect contact (Hewitt 1986; Sebba and Tate 1986). Although *you know what I mean* continues to be used by non-Anglo speakers in Hackney, the most frequent users are now Anglo speakers in Haver-
Our data show a development in the use of PMs in the 1990s (COLT-2) and the 2000s (LIC-2), demonstrating that linguistic innovation is continuing in London. Sebba (1993: 71) argues that you know what I mean as a PM in ethnic minority speech “has gone from being an agreement-seeker to being a marker of agreement in conversation” and that it is confined to speakers under the age of about 30. Its use in this way originated in Jamaican Creole and it was later documented in British Jamaican Creoles – further spreading and innovating in London English spoken by black speakers and thence to white speakers. Before it was documented in London Jamaican, it was not reported in British English. At the time, it was found among speakers of all backgrounds, but mostly in areas with high levels of dialect contact. This use is referred to as a tag by Sebba and Tate (1986), who also examined the use of you know as an agreement marker. You know was used in a similar way by the same groups of speakers. In a list of Creole features in London English (in the 1980s), Hewitt (1986) states that the use of innit as invariant tag is one of the most frequent forms of Jamaican Creole found in the speech of white adolescents in London. The influence was reported to be strongest in high-contact areas and occurred particularly in ethnically mixed conversation, but also in conversation between Anglo teenagers (Hewitt 1986: 128). Hewitt also comments on the use of you know what I mean as an agreement-marker response and says it is “a very recent idiomatic innovation, and one which appears to be developed within the London English of black adolescents but derived from a Caribbean source” (1986: 133). The development of innit may be parallel to that of you know what I mean. Andersen (2001: 110) argues that innit has developed in the same way: in high-contact communities and with non-Anglo speakers and adolescents in general in the lead. This can be seen in the innovative use of non-paradigmatic innit (i.e., where there is no grammatical relationship between innit and the clause it is referring to) and the use of innit as what he refers to as an invariant follow-up. This shows that “invariant tags thrive in multilingual environments”, (Andersen 2001: 112). The use of invariant tags of the innit-type is also found in several varieties of English outside the UK. Algeo (1988: 174), discussing the development of innit, states that invariant tags are more frequent in Englishes with a high degree of contact with other languages. This is exactly that situation which we find in inner London: LIC was collected in Hackney, where 50% of the population is non-Anglo, of whom many do not have English as their first language and are recent immigrants to the city. Today, there is little difference between Anglo and non-Anglo speakers in the extent of innit use. Innit is now an established PM – with the proviso that this investigation has not looked at potential new functional uses. You know what I mean appears to have spread to Havering and you get me may be its innovative replacement in Hackney. The following entry in the ‘Urban Dictionary’ may indicate this (there are also entries for do u get me, ygm and get me):
“Used by urban ‘yoots’ as an alternative to ‘you know what I mean’. Generally an indication that the person delivering the phrase should be the subject of police interrogation as it is likely that they will be the latest contributor to Britain’s knife culture.

TV Interviewer: Does anyone you know carry a knife?

yoot: Yeah man, all da kids is carrying knives, innit. For protection, you get me?”

Taken together, these trends regarding the use of *innit* and *you know what I mean* show some degree of diffusion of PMs. One important caveat is that it may just be that the forms themselves are diffusing, and that some of the innovative functions are not. Why are the forms spreading? The linguistic and most likely the socio-demographic situation described by Hewitt (1986) and Sebba (1993) is somewhat different today. Third generation Afro-Caribbeans living in London may have less knowledge of Creole forms, and there is also more variation in ethnic backgrounds in the area: Afro-Caribbeans are currently not the majority immigrant group. Certainly, the LIC data reveal very little of the type of code-switching and crossing described by Hewitt (1986) and Rampton (1995) and many speakers appear to use exclusively one variety, namely Multicultural London English (MLE) (Kerswill et al. 2008a). The highest use of MLE is in the inner city area among the non-Anglo population. Examining the change patterns for discourse features, we get some information on how the forms are spreading. There is innovation in the centre, but there is also limited spread within the locality. Innovation is in the non-Anglo group, which we can observe for *you get me*, and the use most clearly indexes ethnic identity. *You get me* may just be an innovation in MLE alongside other phonological and grammatical features; its spread is a product of contact. It is first associated with male non-Anglos and thereafter with multi-cultural friendship groups where it spreads to Anglo speakers with a high friendship network score. As it is an emerging PM, it is currently not used by some Hackney speakers with high network scores, regardless of ethnic background. We do not know yet if it will spread to outer London: it was used by only two non-Anglos in Havering. However, once an emerging PM has become established, the comparison between COLT-2 and LIC-2 suggests that it takes around 20 years for it to achieve neutral status, i.e., comparable frequency of use in terms of sex, ethnicity and locality.

As a whole, the developments show that changes and innovation in the use of PMs are often due to dialect contact. In our case, it seems to be caused by contact between speakers of different ethnicities, and in fact ethnicity is one of the most important factors in change in London English (Cheshire and Fox 2009; Fox 2007; Gabrielatos et al. 2010; Kerswill et al. 2008b). PMs that were emerging in the 1980s are today established across groups of speakers, with no large differences in use between them. It is therefore unsurprising that we find no large differences between Hackney and Havering. For some PMs, such as *you know what I mean*, it is today Havering that has the highest use, and they
are used by female Anglo speakers, while Hackney speakers, in particular the male non-Anglo speakers, instead choose other PMs, mainly you get me. In sum, the results indicate that young people, ethnic minorities, an urban environment, and dialect contact are of great importance in language change, a finding that can feed into an explanatory model of language variation and change in high-contact speech communities. We also believe that our method has allowed for a more detailed and efficient exploration of large corpora of spoken data and has demonstrated how different corpora can be made comparable – using fairly simple techniques.

Bionotes

Eivind Nessa Torgersen is Associate Professor of English Language at Sør-Trøndelag University College in Trondheim, Norway. He was previously a Senior Research Associate at Lancaster University, UK, working on two ESRC-funded projects on London English. His research interests are variation and change in London English and sociophonetics. E-mail: <eivind.n.torgersen@hist.no>

Costas Gabrielatos is a Senior Research Associate at Lancaster University, UK. His research interests are conditionals and modality in English, and, more generally, the use and development of corpus-based approaches in various areas of (applied) linguistics. E-mail: <c.gabrielatos@lancaster.ac.uk>

Sebastian Hoffmann is Associate Professor of English Linguistics at the University of Trier, Germany. His research predominantly focuses on the application of usage-based approaches to the study of language. He is particularly interested in syntactic change, aspects of fixedness and the interplay between corpus data and language theory. E-mail: <hoffmann@uni-trier.de>

Susan Fox is a Research Fellow at Queen Mary, University of London, UK. Her research reflects her interest in language variation and change and the social processes that bring about language change, particularly in a multicultural urban environment such as London. E-mail: <s.p.fox@qmul.ac.uk>

Notes

* We are grateful to the British Academy for funding the project Analysis of spoken London English using corpus tools (SG–47692), of which the work reported here forms a part. Likewise, we thank the Economic and Social Research Council (RES–000–23–0680) for funding the project Linguistic innovators: The English of adolescents in London, from which the data for LIC was taken. We also thank audiences at the Corpus Research Group meeting at Lancaster University in March 2009, the Corpus Linguistics Conference in Liverpool in July 2009, CLAVIER in Modena in November 2009, the guest editors, two anonymous reviewers,
Paul Baker and Paul Kerswill for valuable input. We are grateful to Heike Pichler for checking and correcting the *init* tokens in the original LIC transcriptions.

1. In tables, significant and highly significant results are marked with one or two asterisks respectively (* or **) preceding the higher frequency/spread value – two asterisks are also used for \( p \) values of lower than 0.01. Calculations were carried out using Paul Rayson’s online log–likelihood calculator: <http://ucrel.lancs.ac.uk/llwizard.html>.

2. The terms ‘significant’, ‘significantly’, and ‘significance’ will only be used in this article to refer to statistical significance.

3. Parentheses indicate that a word may be elided. For example, (do) (you) know what I mean represents three instantiations: do you know what I mean, you know what I mean, know what I mean.

4. Dotted lines (---) in cells indicate no significant difference in either frequency or spread.


**References**


Kerswill, Paul & Eivind Torgersen. forthcoming. Modeling dialect innovation, transmission and diffusion in high-contact speech communities: an answer to Labov.


