A corpus-based sociolinguistic analysis of indefinite article use in London English

Costas Gabrielatos & Eivind Torgersen
(Lancaster University)

This paper reports on the analysis of the use of indefinite article forms (a/an) in front of vowel sounds in spoken London English, which formed a part of the completed project Analysis of spoken London English using corpus tools (funded by the British Academy). The study used the Linguistic Innovators Corpus (LIC), a 1.4 million word corpus comprising the transcribed and marked-up interview data from the Lancaster/Queen Mary ESRC-funded project, Linguistic innovators: the English of adolescents in London (Kerswill et al. 2008), as well as the Corpus of London Teenage English (COLT) (Stenström et al. 2002).

The research methodology combined approaches and techniques from sociolinguistics and corpus linguistics. Variables were examined individually and in cross-tabulations, using both manual/semi-automated and automated techniques (logistic regression analysis). The former analysis took account of the frequency of the a+vowel pattern relative to the number of opportunities for a choice between a or an (i.e. vowel-initial words preceded by the indefinite article) and the proportion of speakers who used the pattern.

The study examined both linguistic and sociolinguistic variables, but only the sociolinguistic variables yielded statistically significant results. This suggests that the linguistic variables play a minor role, if any at all, in the choice between a or an in front of a vowel sound. The sociolinguistic variables comprised the speakers’ sex, age, ethnicity and place of residence, as well as the ethnic make-up of the friendship networks. In particular the speakers’ ethnicity and place of residence, emerged as the strongest predictors of the use of a before vowels.

The comparative analysis of LIC and COLT showed an almost three-fold increase in the use of a before vowel-initial words by young speakers (19% and 8% respectively). Equally striking is the three-fold change in the proportion of young speakers who use the a+vowel pattern (58% and 20% respectively). More specifically, in LIC, the majority of speakers (52%) alternate between a and an, 43% use an+vowel only, and 5% use a+vowel only. In contrast, the vast majority of COLT speakers (85%) use only an+vowel, with a small minority (15%) alternating between a+vowel and an+vowel - no COLT speaker uses a+vowel only.

The indefinite article form a before vowels seems to have undergone a process of reallocation (Britain & Trudgill 1999) in which its sociolinguistic status has been realigned. While the form a in front of vowels earlier seemed to have been avoided, either because it was socially stigmatised or only formed a part of child language and L2 varieties, it is now frequently found among adolescent speakers in inner London. We argue that the indefinite article form a before vowels forms part of Multicultural London English (Kerswill et al. 2008), along with other phonological and grammatical features that have already been documented.

References


A corpus-based sociolinguistic analysis of indefinite article use in London English

Costas Gabrielatos & Eivind Torgersen
(Lancaster University)

Part of the project:
Analysis of spoken London English using corpus tools

PI: Eivind Torgersen
CI: Paul Kerswill
RA: Costas Gabrielatos
Consultant: Sebastian Hoffmann
Why study London English?

• London as the centre of linguistic innovation in British English
  – Diffusion of linguistic features from inner to outer London and beyond

• London as a multicultural city
  – High level of dialect and language contact
Lack of *a/an* alternation
in British English dialects

- Wright (1905:71):

  ‘very few dialects follow the rule of the literary language according to which *an* is used before a vowel’
Increase of *a*~*an* alternation with increased age in both L1 and L2.

Increase of *a*~*an* alternation for AAVE speakers in multi-ethnic friendship groups.

Decreased *a*~*an* alternation for white speakers in multi-ethnic friendship groups.
Tower Hamlets (Fox 2007)

- Tower Hamlets (London borough just South of Hackney)
  - *an+vowel*: 65%
  - *a+vowel*: 35%

- Use of *a+vowel*
  - Bangladeshi boys: 75%
  - Mixed race White/Afro-Caribbean boys: 35%
  - White boys: 15%

- Boys: 44%
- Girls: 5%
Research questions and hypotheses

- What variables are good predictors of use?
  - Linguistic
  - Sociolinguistic

- Predictions:
  - Non-Anglo boys in Hackney will be the highest users.
  - A person in a multicultural friendship group will use *a+vowel* more often.
# The Linguistic Innovators Corpus (LIC)

<table>
<thead>
<tr>
<th>Description</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data collected</td>
<td>2005</td>
</tr>
<tr>
<td>Data collection method</td>
<td>Sociolinguistic interviews</td>
</tr>
<tr>
<td>No. of speakers</td>
<td>118</td>
</tr>
<tr>
<td>Age</td>
<td>young=16-18, old=70+</td>
</tr>
<tr>
<td>Sex</td>
<td>female, male</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>Anglo, non-Anglo</td>
</tr>
<tr>
<td>Residence</td>
<td>Inner London (Hackney), Outer London (Havering)</td>
</tr>
<tr>
<td>Social class</td>
<td>Working class</td>
</tr>
</tbody>
</table>
Variables (1)

Linguistic variables

- Initial vowel sound (phoneme)
- Stress pattern (stress-bearing syllable)
- Number of syllables
Variables (2)

Sociolinguistic variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Values and number of speakers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>old=70+ (18)</td>
</tr>
<tr>
<td>Sex</td>
<td>female (53)</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>Anglo (77)</td>
</tr>
<tr>
<td>Residence</td>
<td>Hackney (58)</td>
</tr>
</tbody>
</table>
### Variables (3)

**Friendship network score**

<table>
<thead>
<tr>
<th>Score</th>
<th>Ethnicity composition</th>
<th>Speakers</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>all friends same ethnicity as self</td>
<td>16</td>
</tr>
<tr>
<td>2</td>
<td>up to 20% of a different ethnicity than self</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>21-40% of a different ethnicity than self</td>
<td>19</td>
</tr>
<tr>
<td>4</td>
<td>41-60% of a different ethnicity than self</td>
<td>26</td>
</tr>
<tr>
<td>5</td>
<td>61-80% of a different ethnicity than self</td>
<td>30</td>
</tr>
</tbody>
</table>
Methodology (1)

- Analysis and annotation of sorted concordances of *a* and *an*:
  - genuine instances of ‘indefinite article + vowel-initial token
  - socio-demographic details of *a/an*+vowel users

- Tabulation of speaker information:
  - user / non-user
  - tokens and types
  - variable values
## Methodology (2)

<table>
<thead>
<tr>
<th></th>
<th>Expression</th>
<th>Utility</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Density</strong></td>
<td>Frequency per 100 instances of indefinite article + vowel-initial word</td>
<td>Shows the relative frequency of <em>a+vowel</em> use, as opposed to <em>an+vowel</em>.</td>
</tr>
<tr>
<td><strong>Spread</strong></td>
<td>Number of users per 100 speakers</td>
<td>Shows the proportion of speakers using <em>a+vowel</em>.</td>
</tr>
</tbody>
</table>

Logistic regression analysis

- Variables considered individually and together
- Cross-tabulations
Indefinite article use: density

<table>
<thead>
<tr>
<th></th>
<th>Raw freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a$ + vowel</td>
<td>182</td>
</tr>
<tr>
<td>$an$ + vowel</td>
<td>1042</td>
</tr>
<tr>
<td>Total</td>
<td>1224</td>
</tr>
</tbody>
</table>

$an$ 85.1%

$a$ 14.9%
Indefinite article use: spread

<table>
<thead>
<tr>
<th>Category</th>
<th>Speakers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alternation ( (a \sim an+vowel) )</td>
<td>61</td>
</tr>
<tr>
<td>No alternation ( (a+vowel \text{ only}) )</td>
<td>6</td>
</tr>
<tr>
<td>No alternation ( (an+vowel \text{ only}) )</td>
<td>51</td>
</tr>
<tr>
<td>Total</td>
<td>118</td>
</tr>
</tbody>
</table>
**a+vowel: LIC-2 vs. COLT-2**

<table>
<thead>
<tr>
<th></th>
<th>Freq.</th>
<th>V-initial tokens</th>
<th>Users</th>
<th>Speakers</th>
<th>Density</th>
<th>Spread</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIC-2</td>
<td>170</td>
<td>907</td>
<td>58</td>
<td>100</td>
<td>18.7**</td>
<td>58.0*</td>
</tr>
<tr>
<td>COLT-2</td>
<td>9</td>
<td>119</td>
<td>3</td>
<td>15</td>
<td>7.6</td>
<td>20.0</td>
</tr>
</tbody>
</table>
Effect of linguistic variables

- Stress pattern
- Number of syllables

→ No effect on use of a+vowel
  - main effects
  - interactions

- Initial vowel sound (phoneme)

→ Only /aʊ/ predicts a+vowel (p<0.001)

→ No significant effect when regression model included all three linguistic variables.
Effect of sociolinguistic variables

- All individual variables have a significant effect on \( a+vowel \) use
  - Age = young
  - Sex = male
  - Ethnicity = non-Anglo
  - Residence = Hackney

- When interactions are included in the model
  - Ethnicity*Residence (non-Anglo*Hackney) emerges as a strong predictor.
Individual variables: density and spread

Density vs. Spread

Hackney
Havering
Young
Old
Anglo
Non-Anglo
Male
Female
Ethnicity*Residence: density and spread

Density

Spread

AA=Anglo, NA=non-Anglo, HK=Hackney, HV=Havering
Network score (1)

Average scores
- Non-Anglo=4.48; Anglo=2.75;
- Hackney=4.42; Havering=2.48

Prediction
- Correlation between network score and $a+vowel$ use

Why?
- NA and HK show higher preference for $a+vowel$
- High score networks by definition have high proportion of NA speakers

⇒ If so, AA and NA users of $a+vowel$ should have higher average scores than non-users.
Network score (2)

The bar chart compares the network scores of Anglo users and non-users versus Non-Anglo users and non-users. The scores range from 0 to 5.

- Non-Anglo users have a network score around 4.5.
- Non-Anglo non-users have a network score around 1.
- Anglo users have a network score around 3.
- Anglo non-users have a network score around 0.5.

The chart shows that Non-Anglo users have the highest network score, followed by Anglo users, Non-Anglo non-users, and Anglo non-users.
Network score (3)

- Correlation of network scores to usage preference probably symptomatic of …
  - the correlation between network score and ethnicity.
  - the very strong correlation between usage and ethnicity (non-Anglo).

- Overall, friendship network multi-ethnicity not a good predictor of a+vowel use.

However

- Network score good predictor of use among Anglo speakers.
Conclusions

- Strongest predictors of $a+vowel$:
  - Age (young)
  - Ethnicity*Residence (non-Anglo*Hackney)

- Reallocation of indefinite article form $a+vowel$ due to extensive dialect contact in inner London:
  - $a+vowel$ used in informal styles among young speakers

- $a+vowel$ a feature of Multicultural London English - along with several other phonological and grammatical features already documented
Appendix
## Annotation examples

<table>
<thead>
<tr>
<th>Category</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>a</em> is not an article</td>
<td><em>to have a a at least</em></td>
</tr>
<tr>
<td>Repetition</td>
<td><em>a a book</em></td>
</tr>
<tr>
<td>Self-correction</td>
<td><em>have you got a an extended family?</em></td>
</tr>
<tr>
<td></td>
<td><em>it would go like a like two years</em></td>
</tr>
<tr>
<td>Hesitation / fillers</td>
<td><em>a erm</em></td>
</tr>
<tr>
<td>Unfinished utterances</td>
<td><em>… you know what I mean it’s a. you might live in Clapton and you …</em></td>
</tr>
<tr>
<td>Backchannel interruption</td>
<td><em>Sue: mm what's /[Fatima: I liked my a]l your favourite subject at school</em></td>
</tr>
<tr>
<td>Humorous or metalinguistic uses of ‘<em>a + vowel’</em></td>
<td><em>a orse</em></td>
</tr>
<tr>
<td></td>
<td><em>… you can't say a aeroplane you have to say an aeroplane</em></td>
</tr>
</tbody>
</table>
Indefinite article use: types

<table>
<thead>
<tr>
<th>Type</th>
<th>Raw freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>a only</td>
<td>41</td>
</tr>
<tr>
<td>an only</td>
<td>253</td>
</tr>
<tr>
<td>Both</td>
<td>63</td>
</tr>
<tr>
<td>Total</td>
<td>357</td>
</tr>
</tbody>
</table>

- a only: 11.5% (41 of 357)
- an only: 70.9% (253 of 357)
- Both: 17.7% (63 of 357)