The Monash Corpus of Spoken Australian English
Julie Bradshaw, Kate Burridge and Michael Clyne
Monash University

Abstract
This paper takes stock of findings based on the Monash Corpus of Australian English. In 1996–97 members of the (then) Monash University Department of Linguistics embarked on the collection of a corpus in Victoria to facilitate the study of variation in phonology, morphosyntax, lexis and discourse patterns. The largest part of the corpus was based on data from Year 10 students in ten schools selected according to socioeconomic status of locality and type of school (state, Catholic, independent including Greek Orthodox and Jewish; co-educational and single-sex, boys and girls). The data comprises two conversations per student with a stranger (including some citation reading), and two self-taped conversations, one with (usually) three generations of their family and one with same-age friends. The corpus has been used for research by colleagues and graduate students from LaTrobe, Melbourne, and Monash Universities. It has enabled some hitherto unidentified syntactic features of Australian English to be recognized (concord, articles, relative clauses). It has drawn attention to intergenerational change in certain vowels, to developments in /t/ tapping and glottalization, most especially in informal settings, to onset glottalization, and to the emergence and disappearance of ethnolects and the identification of their features. It has also been employed for studies of discourse quotatives, including comparisons with American, British and Canadian English. As yet, the corpus remains underutilized. For example, phonological analysis has concentrated on the interview data, and much could still be done on situational variation, particularly in families of migrant background. There is also scope for a new round of recordings to make the project a longitudinal one.

1. Introduction
The purpose of this paper is:
1. to describe the Monash corpus of Australian English and discuss what has been done with it so far,
2. to discuss which parts of the corpus have been underused, what value they could provide,
3. to describe how the project could be extended.

We have been encouraged by international interest in the project and this has prompted us to make information about the corpus more readily available.

The project began in 1996–97 and was initiated for the following reasons:
1. There had been very little research on Australian English (post-Mitchell and Delbridge) carried out in Melbourne or other parts of Victoria.
2. It was to be a departmental project enabling each staff member to work on aspects of the corpus relevant to their area(s) of specialization and at the same time for all to work together.
3. It could stimulate students’ interest in undertaking research on Australian English.
4. A multidimensional data collection method could be trialled.
1.1. Data Collection

The main corpus was collected in metropolitan Melbourne. The main participants (55 Year 10 students) were selected from a range of schools (schools of different socio-economic status (based on fees and areas of Melbourne), boys’, girls’ and co-educational schools, state, Catholic, independent schools, traditional and low fee Anglican/Protestant schools, one Jewish and one Greek Orthodox school (the latter two in order to explore the presence of ethnolects). Each participant’s socio-economic status was also identified based on the mean house price of their suburb of residence. There were 30 female and 25 male core participants. All of them had received all or almost all of their schooling in Melbourne. Two recordings were made of conversations between a research assistant and a participant aged 15–16. In one session the participant was asked to read words and sentences, some of which contained sociolinguistic variables. The interviewers were for the most part postgraduate students only a little older than the participants. In addition, the participant recorded conversations with about five friends of the same age and with family members.

In addition, pilot data collection on rural Australian English took place in eight families in the Victorian Western District – four in each of two research areas which had been settled in the mid-19th century, one from Ireland, and the other from eastern Germany.

The recordings, which are of varying quality, were made with cassette recorders and more recently copied on to CDs. These recordings have since been transcribed.

2. How the corpus has been used so far

The Monash Corpus has been used as the basis of research into aspects of Australian English phonology, syntax, discourse, ethnolects and historical change (mostly grammatical). We will briefly summarize the findings of some of the studies to date. The data has been utilized by staff of the Linguistics Program at Monash (some now at other universities) and by graduate and honours students of Monash, Melbourne and LaTrobe Universities. We can touch only briefly on some of the studies in this paper.

<table>
<thead>
<tr>
<th>Sub-discipline</th>
<th>Topic</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phonology</td>
<td>The variable realizations of consonant /t/ (plosive, tapped, fricated or glottalled/glottalised)</td>
<td>Tollfree (2001)</td>
</tr>
<tr>
<td></td>
<td>Vowel variation in six Australian English items (beat, boot, getup, hair, fear, tour)</td>
<td>Fricker (2007)</td>
</tr>
<tr>
<td></td>
<td>Intonational variation in Australian English</td>
<td>Fletcher and Loakes (2006a; 2006b)</td>
</tr>
<tr>
<td></td>
<td>Vowel variation in TRAP, DRESS and NORTH</td>
<td>Clyne, Fletcher, Loakes, Tollfree (2000)</td>
</tr>
</tbody>
</table>
2.1. Phonology

There have been a number of studies investigating different phonological aspects of Australian English. The following are the most significant studies to emerge.

2.1.1 Tollfree (2001)

Tollfree uses the Monash Corpus data, together with some urban speech data she collected in 1993–4 to describe the variable realizations of consonant /t/ (plosive, tapped, fricated or glottalised/glottalled¹), using both auditory and instrumental acoustic analysis. She analyses data from two socioeconomic groups (lower and middle), in two speaking styles (word and sentence reading style and conversational style) and in four phonetic contexts: intervocalic medial, intervocalic final, pre-consonantal final and pre-pausal final (p. 56).

She finds that among the 15–16 year olds, glottalised variants and the tap are competing in intervocalic final position (e.g. *lot of*, *get out*), with no evidence of the fricated or fricative variants found in older speakers. She suggests that fricated/fricative forms are age-related, and are disappearing from Australian English (p. 59).

The effect of speaking style is observed most strongly among the middle socio-economic group speakers, who shift from 100% glottalised forms in pre-consonantal final contexts (e.g. *pelt with,* *paint that*) in conversational style to 41% in formal style (p. 59). This group also shows a preference for tapping rather than glottalisation in formal style in intervocalic final position (p. 60), suggesting the former is the more acceptable variant. She suggests that this may reflect the prestige of American English, a claim she says is supported by evidence of categorical use of the tap by many speakers in certain lexical items (e.g. *theoretical, data, automatic, city, attitude,*...
beauty). By contrast she finds glottalisation of /p/ and /k/ in certain lexical items in the data: occasionally of /p/ in couple, and near-categorically of /k/ in like among 15 and 16 year olds (p. 52).

2.1.2 Clyne, Fletcher, Loakes & Tollfree (2000)

These researchers examine data from 42 adolescents from the corpus (21 male and 21 female) and track the variants for the vowels in the DRESS, TRAP and NORTH sets. They test the vowels against the following variables: (1) speaker sex; (2) government versus non-government school (indicative of socio-economic class); (3) single sex versus co-educational school; (4) male single sex versus female single sex school. They report that the simple distinctions government versus non-government school or co-educational versus single sex are not significant predictors of variation. Findings reveal considerable variation away from the standard variants previously identified, especially for TRAP and NORTH. A number of general trends emerged: (1) lowering of TRAP and DRESS; (2) lowering of TRAP and DRESS significantly more in females than males, especially for students of single sex non-government (fee-paying) schools in middle class areas; and (3) slight centralizing of NORTH.

They suggest a follow-up study that involves sub-corpora (e.g. cross-generational discussions and discussions involving same-age friends) to determine the degree of accommodation to interlocutor.

2.1.3 Fletcher and Loakes (2006a & b)

Both studies confirm that uptalk (high rising terminals) is a characteristic intonational tune of adolescents in southeastern Australia. Fletcher and Loakes (2006a) examine the speech of ten females from the corpus in order to establish whether or not regional variation occurs in the intonational tunes of urban versus rural adolescents. Comparing two groups of Melbourne speakers (Mount Waverley in the eastern suburbs) and two groups from the Western District (around Tarrington 300 km west of Melbourne), they analyse 15 minutes of conversational data for each group. Both rural and urban speakers use uptalk to varying degrees, although the findings show marginally more high rises associated with declaratives in the speech of the urban teenagers compared to their rural counterparts. Many of these statement high rises are turn-internal and have a continuative discourse function; i.e. that of a ‘floor-holder’.

Following on from this last study, Fletcher and Loakes (2006b) examine intonational patterns in the conversational data of 17 female participants from Melbourne and the areas. They observe a high incidence of rising tunes, although statement high rises are relatively scarce (13% in the rural data, 11% in the urban data), though this is considerably more than the 1.6% reported by Horvath (1985). As before, almost all of these appear within a turn with the function of floor-holding. Fletcher and Loakes suggest that a more important characteristic of adolescent speech is the use of mid-level non-falling tunes to signal that the discourse is on-going.
2.1.4 Fricker (2007)

Fricker uses Monash corpus data to conduct a socio-phonetic study of vowel variation among Melbourne adolescents. Using formant analysis, he identifies the variant vowel forms for six items \(\text{beat, boot, getup, hair, fear, tour}\). The tokens were selected from the Monash corpus word list and sentence reading data so he describes formal style only. His focus is on the correlation of forms with sex and social class. He compares the findings with Sydney corpus data from the Sydney-based ANDOSL corpus, supplied by Cox and Palethorpe, and examines the applicability of Mitchell and Delbridge’s framework to the Melbourne data.

In general, the data shows only small differences between Melbourne and Sydney speakers. Mainly these consist of greater retraction and/or raising in Melbourne (and occasional lowering). However these differences do not apply across all sociolects. (p. 108). He finds that the sociolectal continuum described by Mitchell and Delbridge with its categorization into cultivated, general and broad can be applied to the 1996 Melbourne data, though the lectal range is less. As has been suggested elsewhere, Fricker’s findings indicate that there has been a narrowing of the sociolectal range, with more speakers using General variants rather than the Cultivated or Broad ends of the continuum. Mitchell and Delbridge (1947, 1965) found 34% of speakers using Broad, 55% General and 11% Cultivated. In the tokens that Fricker analyses from the Monash Corpus, he finds 12.5% Broad, 69.6% General and 17.4% Cultivated.

Analysis of speech by gender does not show the expected clustering of males at the Broad end of the continuum, and females at the Cultivated end (pp. 114–5). On four of the six vowels studied, male speakers use more cultivated than broad variants (p. 127).

Fricker uncovers a relationship between sociolect and the socioeconomic status of the participants’ home suburbs and of their schools (p. 118). Participants from higher SES schools produced fewer Broad tokens, those from lower SES schools produced more, but students from medium SES schools produced more than either. Students with a mismatch between SES of suburb and school showed evidence of hypercorrection: students from medium SES suburbs attending higher SES schools produced very few Broad tokens, while students from higher SES suburbs attending medium SES schools produced a high number of Broad tokens. He attributes this to upgrading or downgrading their speech to fit in with their peers (p. 121).

2.2. Morphosyntax

The corpus has been very underutilized, especially when it comes to grammatical aspects. The few studies which have been done are discussed below.

2.2.1 Newbrook (1998)

Newbrook has made passing comments on the Monash Corpus, but appears to base most of his articles on Australian English grammar on the ACE corpus (= Australian Corpus of English) held at Macquarie University. In this particular piece, he draws on interview data (51 recordings) and conversational data (7 conversations) from 39 of the Monash corpus participants to report on
the incidence of variable grammatical and lexical features. He notes a number of features including the omission of pronouns and/or auxiliary or copula verbs (e.g. Ø got to go to the U2 concert), absence of definite or indefinite pronoun (e.g. he’s Ø very good batsman), the use of there’s with plural complements (e.g. there’s a few people at school).

2.2.2 Taylor (2001)
In her honours thesis, Jo Taylor explores modal usage in the corpus, in particular the grammaticalization of four quasi-modals gonna, gotta, hafta and better (especially interesting are the rarely investigated gotta and better). She bases the study on interviews with 15 male and 15 female subjects and reports that in terms of their linguistic behaviour gonna, gotta, hafta are well advanced along the path of grammaticalization in spoken Australian English (increasing subjectification, omission of auxiliary etc.). A number of interesting side issues emerge from this thesis. Taylor reports: (1) differences between gonna and will; (2) the appearance of betu (where better is remodelled to fall in line with the other modals want to, have to etc); (3) zero use of shall and rare appearance of must (replaced by gonna and hafta).

2.2.3 Salasinski (2006)
Caroline Salasinski’s work is another honours thesis. Her investigation of possessive constructions is largely historical — the development of inflectional –s versus prepositional of. However, she draws on the Monash corpus to complete the picture of the current-day situation and to hypothesize on the future of these two markers. For her data, she selects two students (male and female) from each of the schools and analyses their interviews for constructions of possession. Not much can be concluded on the basis of such a small sample. Nonetheless, she confirms what other researchers have claimed; namely, faced with these two constructions, speakers choose between them on the basis of animacy and subjectivity (subjective animate modifiers showing preference for the inflected genitive) and also the semantics of individual nouns.

2.3. Discourse
A number of studies of Australian English discourse have drawn on the Monash corpus, in particular work by Newbrook, Norrby and Winter, and Winter. The following is a selection.

2.3.1 Winter (2001a) and Winter (2002a)
Winter (2001a) draws on the corpus data to illustrate the use of cultural icons in adolescent boys’ negotiation of masculine identities – in this case drawing on the cartoon Beetlejuice to nickname a boy Pondscum, after a spoof James Bond character James Pondscum.

Winter (2002a) uses the Monash corpus data to examine the discourse quotatives (say, go, be like and zero quotative), in order to compare with other varieties of English, particularly in relation to the innovation be like. Data is drawn from 30 of the sociolinguistic interviews: those in which the participants performed the voices of others in their talk. Though the other forms (say, go, zero-quot.) are more common, she does find evidence of some use of be like, generally in third person singular form (e.g. yesterday there was this kid’s brother he was like, ‘ah bugger
it’). However be like seems to function differently from the pattern documented in American, Canadian or British English, often closing the account and expressing the speaker’s or institutional representative’s attitudinal positions about the events narrated (2002a:20).

2.3.2 Winter and Norrby (2000); Norrby and Winter (2001)

Both studies focus on the functions and meanings of discourse extenders, or set marking tags such as and stuff and (in Swedish) å så “and so”. Such set marking tags indicate that the preceding NP (or other constituent) is to be interpreted as an example of a more general set, or as a specific case to which other referents may be added by interlocutors’ inference.

(1) but it has become sort of quite popular with now too right (. I think [(.) with soul sort of soul and such]. [Winter & Norrby 2000:2]

(2) I go out with my friends and stuff … Just, I’m not into .. big clubs and [stuff] [Norrby & Winter 2001:1]

(3) … and that’s good and he’s good because you know we can talk to each other about stuff and that … whereas other people you know, tell someone else or something … [Norrby & Winter 2001:1]

The studies draw on two corpora — the Monash corpus (48 interviews with adolescents in urban Melbourne) and a roughly comparable Swedish corpus (14 recorded informal group interviews with senior high school students) as the basis for the analysis of the use by English and Swedish-speaking adolescents.

The most frequent set marking tags in the Australian English corpus emerged as or some(t/f)hin(g/k), and stuff, and and that. Overall they are used equally by girls and boys: both use or something to the same extent, but boys make greater use than girls of and that while girls use and stuff more. One function appears to be to establish group relations and common ground, drawing on positive (inclusion) and negative (hedging) politeness. They function to signal next turn topic-completion, and may mark youth identity. The researchers note that “the adolescents use the extenders as a resource to affiliate with the interviewer, with present and absent peers and to the topic under discussion” (Winter & Norrby 2000:2), allowing participant co-construction of the relevant implied meaning.

The proposed longitudinal study of the corpus participants (see below) would provide the possibility of verification of the claim that these forms function as youth identity markers. Replication of the data collection with adolescents would provide evidence of the stability or transience of youth language across different cohorts.

2.4. Ethnolectal Variation

Ethnolects are varieties of a language which mark speakers as coming from families or communities which originally spoke another language. Often ethnolects are employed within an in-group by those who use mainstream Australian English with the wider community. The ethnolect generally takes on the symbolic function earlier vested in the community language –
especially the solidarity and identity functions. Ethnolects may be marked by lexical, phonological and syntactic transference. This part of the project built on earlier work at Monash (e.g. Clyne 1970, 1981). The data resulted in Clyne, Eisikovits and Tollfree (2001, 2002), focusing on three ethnolects: Greek Australian English in Melbourne, Jewish/Yiddish Australian English in Melbourne, and German/Lutheran Australian English in the Western District. Fletcher and Loakes (2006a, b) included data from English in the originally German-settled area of the Western District.

In each case, the corpus has been supplemented by additional material:
- Greek ethnolect data supplemented by recordings within one large extended family;
- Jewish/Yiddish ethnolect data supplemented by recordings collected among students of a Chassidic Ultra-Orthodox school, in which Yiddish and Hebrew as well as English were mediums of instruction; and
- German/Lutheran ethnolect data supplemented by an English segment in German interviews with community members one generation earlier (though not necessarily in the same family) than the oldest generation recorded in the Monash corpus data.

The studies find that the Greek-Australian ethnolect (GAE) is strongest in the family setting. There is variation in the number and degree of ethnically marked variants, between speakers and between data sets, because of social networks and desire and need to identify with the in-group. There is a range of features from salient to peripheral but still contributing to the totality of the ethnolect. Some of the phonological features are shared with the other young people but are overrepresented in GAE, e.g. the lowering of /e/, [a] for /ə/ as in soccer, stronger final element [ia] in year, here; and changes as in the mainstream: final /k/ glottalization, tapping of /t/, velarized and vocalized /l/.

The specific features are voicing of /p/ and /s/ in medial position (properly, baseball), heavy aspiration of /k/, and the stopping of fricatives as in the, them. In family communication, Greek lexical items relating to food, entertainment, trips to Greece and family reminiscences are most readily transferred into English. Living in a ‘Greek neighbourhood’ increased the ethnolectal features. They were reduced in exogamous families, where family members were raised in a remote country town or attended an Anglican/Protestant school.

It is difficult but possible to differentiate between Yiddish-based and general Jewish ethnolects. The former contained phonological features such as uvular [ɾ] and [ɔ] addition [θinɡə], lexical features such as Yiddish expressions (Zay gezund ‘Be healthy’, Shmekh the luft ‘Smell the air’), kinship (Bubbe ‘Grandmother’) and food (Kneidlach ‘dumplings’) and some grammatical ones: direct object deletion after have, get, make, as in I don’t want to have. The Jewish ethnolect transfers Hebrew words for religious festivals and observances (chasene ‘wedding’, mitzvah tans ‘wedding dance’, pesach ‘passover’, daven ‘to pray’ (probably via Yiddish), but may also contain some Yiddish-derived features (e.g. shmus, shlep and the grammatical transfers mentioned above).
The Western district data showed the disappearance of the German/Lutheran ethnolect over three generations. Contrary to anecdotal information, we found no evidence of a continued Irish ethnolect. German was the main language of community institutions, including church, and one of the mediums of instruction at school, in the Tarrington-Tabor areas until the First World War, and many people over 60 could still use it until the early 1970s and beyond (Clyne 1968, 1981, Kipp 1980, 2002). Phonological features included devoicing of /d/, /z/, /b/ and /g/, velarization of /r/; the pronunciation of pastor [pastә] and Lutheran (short [ʊ]), semantic transference of German noch, schon and verschieden in yet, already and different (e.g. Our family has been here for three generations already but we have some German books yet, That happened different times already), and syntactic transfers such as deletion of indefinite article before an occupation and use of definite article before language names (She was teacher; He doesn’t speak the German), to of relationship (He is a father to Jim Nagorcka).

The grandparent generation of the Monash Australian English corpus was characterized by variation between monophthongs and diphthongs, /e/ for /æ/, semantic transference of prepositions and of expressions such as come with, bring with (He brought his congregation with), already, yet and different. By the middle generation, only yet, bring with, to of relationship, and the pronunciation of pastor and Lutheran remained of the ethnolect. The English of the younger generation was indistinguishable from that of other Australians. To different informants, the ethnolect signified three aspects of identity—German ancestry, Lutheran affiliation and regional origin or a combination of these.

Across the studies, ethnolects are generally a product of language shift, whether in progress or completed. The continued presence of ethnolectal features is associated with the identity-rich domains of family and religion and a habitat with demographic clusterings, reinforced by denominational schools. Variation corresponds to a continuum of ‘in-groupness’.

3. Where to from here?

There is no doubt that the current corpus has been underutilized. As yet, no one has considered the corpus with a view to studying lexical variation. Much more could be done on aspects to do with rural versus urban differences, grammatical and phonological features. Moreover, research so far has been based overwhelmingly on the interview data Comparisons with data from interactions with family and friends might yield interesting findings. Furthermore, data from twelve children from non-English-speaking families, especially in two mainstream schools of high migrant density, a state school in the outer south-east and a Catholic one in the north-west, has not as yet been analysed for ethnolectal features.

There are two tasks that could be carried out (subject to funding) to make the current corpus more attractive to potential researchers:

- The recordings were made with cassette recorders (and have since been transferred to CDs). They vary in quality. Fricker (2007) was able to clean up the recordings for his study of vowel variation. This remains be done to the rest of the corpus.
The corpus needs to be made more accessible to potential researchers — just as Macquarie University (Linguistic Department) has various password protected linguistic corpora available on the website, the same could be set up for this corpus. This would also make the data more user-friendly, being searchable etc.). This could perhaps be an outcome of the proposed Australian National Corpus, a current initiative of the Australian Linguistics Society.

There are two possible ways of following up this study.

• We could track down the original speakers and record them again, after 12 years or so. A follow-up study of these adolescent participants would allow insights into the development of the various linguistic features to see whether they continue in the participants’ adult discourse (especially interesting with respect to the ethnolectal features, but also generally to see what of the age-related variation might signal changes in progress).
• We could replicate the study (ensuring as much as possible the comparability of informants). Such an update of the corpus would enable researchers to assess the linguistic features previously examined but some 12 years down the track. For example, we could observe the progress of the vowel variants and also grammatical features and discourse markers such as like (and perhaps compare findings with recent London youth corpus data showing a new quotative form this is X).

In this way the corpus could be the basis for continuing sociolinguistic research on Australian English, especially in the Victorian context. We hope that it may contribute towards the proposed Australian National Corpus.

1 Tollfree (2001:52) uses the term glottalisation to cover the range of pre- and post-glottalised forms and canonical glottal stop. “The acoustic results from the current survey … identify no lone glottal stops, but instead a range of glottalised variants, involving a glottal gesture in addition to an alveolar closure gesture for plosive [t]” (2001:51).
References


Winter, J. 2002b. 'I s'pose we've had some pretty higgledy-piggledy times too': Women's experience narratives about marriage and motherhood'. Paper presented at the Paper presented to The XV World Congress of Sociology July 2002, Brisbane.