The International Corpus of English in Hong Kong

PHILIP BOLT and KINGSLEY BOLTON

1. Introduction

The sociolinguistic realities of English in Hong Kong are unlike those of any other society in Asia, not least of all because Hong Kong is Britain's only surviving territorial colony of any economic or strategic significance. Although an estimated 97 per cent of the population is Chinese, and Cantonese is the principal local language, English has been an official language of government and law since 1841, and throughout the present century it has been accorded semi-official status as the major language in many secondary schools, colleges, and universities. In addition, English is also considered the dominant language of business in the larger companies.

Hong Kong is currently experiencing a period of transition in which the decolonialization of institutions from Britain is being given increasing priority. Partly as a result of this and of other economic and social forces, English and Chinese now have an increasingly complex coexistence in government, public administration, law, education, the business sector and the mass media. On 1 July 1997 Hong Kong will become a Special Administrative Region of the People's Republic of China. It is likely that after this date the *de facto* and *de jure* status of English will be changed, and that either Cantonese or Putonghua (Mandarin) will assume a greater proportion of some of the roles previously held by English.[1]

Considering the historical presence of English (of a quasi 'ESL' variety) and its public and official status, our motivations for undertaking the ICE work in Hong Kong were twofold. First, we were motivated by the opportunity to identify and analyse characteristics in the English used in the local setting across a broad range of contexts and text types. As such, it was anticipated that the ICE project would provide an excellent opportunity to carry out comparisons between the local data and data both from native-language countries on the one hand and from other ESL countries on the other. Secondly, our involvement in the ICE project was also motivated by the intention to consider the sociolinguistic background and context of English in use in Hong Kong, particularly as the local language situation has context; see, for example, Bacon-Shone and Bolton, 1995; Bolton and Kwok, 1990; Fu, 1987; Lord and Cheung, 1987; Luke and Richards, 1982; Richards and Luke, 1982; and So, 1987, 1992. In fact, over the f0ur-year research period our focus on these two concerns has tended to be reversed, in that we have been obliged, from the outset, to consider the sociolinguistic context of the data collection, the functions of English and the participant roles of English language users as much as the nature of the language used.

In Bolt (1994: 23) it was noted that 'Having now collected a substantial amount of data we will soon be embarking on its analysis with a fairly confident target of 75% to 80% completion of the ICE total of 1,000,000 words.' This prognosis was set against a brief description of the principal issues that we had confronted, although not necessarily resolved: who could be counted as a Hong Kong person, the uses of English and Chinese in Hong Kong, their relative distribution and quantity, the linguistic background to the education system, and the features of the range of English encountered. At the 1993 ICAME conference we were happy to endorse the decision to extend the data-collection period until the end of 1994. Indeed, we have had to use 1995 as well and will only finish collecting the last few collectable texts in early 1996. A positive result, however, of this lengthened collection period is that we will probably achieve almost 90 per cent of the ICE total: some 445 texts (although a small proportion will not have complete biographical data). Below we present and discuss the issues noted above in the context of our experience of identifying and collecting texts within the ICE categories. Section 2 considers the sociolinguistic context, language proficiency, and language use. Section 3 considers the experience of applying the ICE categories, essentially issues of sourcing and text inclusion. Section 4 offers a brief prognosis for the next stages of the project-tagging and parsing written text.

2. The sociolinguistic context and the quantification of language use

When we considered the ICE objective of collecting English language data and comparing the use of English in Hong Kong with its use in other societies, one basic assumption was that the particular range and quality of the Hong Kong data would be affected by a matrix of sociolinguistic relationships. These relationships include those between the ethnic and linguistic background; between the local educational system and the linguistic profile of the community; and between the linguistic backgrounds of local English speakers and salient features of the type of
English found in Hong Kong. These relationships are by no means totally deterministic. Coming from a particular ethnic and linguistic background and attending a certain type of school do not, for example, necessarily define the nature of the language used by a particular individual. However, it is felt that the parameters outlined above should serve to distinguish our data from that collected in, for example, India, Singapore, and the Philippines.

Members of the population that we are interested in have three characteristics. First, they have a Hong Kong Chinese background (possibly from the People's Republic 0f China but increasingly they have been born in Hong Kong). Secondly, their first language is Cantonese, or they would normally use Cantonese outside their home environment. Thirdly they have had the majority of their primary and secondary education in Hong Kong. These characteristics reflect the pattern of immigration into Hong Kong since 1945, the continued existence and use of minority Chinese dialects other than Cantonese (Bacon-Shone and Bolton, 1995), and the fact that, until relatively recently, the provision of tertiary education locally was very limited (a proportion of local people still go overseas for their education at all levels). The criteria for inclusion in the ICE-HK corpus exclude the local Hong Kong Indian community, the temporary Vietnamese community, the Philippine community, and expatriates from Europe, the United States, and other countries. We have endeavoured to keep to a minimum the number of those who have had secondary education in an English-speaking country, because such an experience is likely to affect the nature of the language they use. The impact of the educational system must be emphasized, as English is almost entirely a 'learnt' language in Hong Kong and the main context of such language learning is the public school system. The proportion of people who have contributed to the ICE collection and who fit centrally into this framework is over 95 per cent (a few biodata forms have yet to be returned and some will not be provided).

In Bolt (1994: 16) it was noted that two assumptions were made when work was begun. The first of these was 'given the use of English in Government, public administration, large business corporations, the judicial system, the media and communications industries and education, we assumed that there must exist a genuinely widespread use and comprehension of English.' Naturally, it was felt that the status, quantity, and quality of the English used within a population would help or hinder data collection and the resulting collection would reflect these factors.

In an ESL/EFL society such as Hong Kong, there is a range of complex and vexed issues relating to who 'knows' English, and knows to what degree of proficiency, and who uses English, for what purpose(s) and under what circumstances. Estimates based on various surveys conducted during the ICE period vary from 56.7 per cent (Bacon-Shone and Bolton, 1995), to 29.4 per cent *(Hong Kong Population Census,* 1991) to 20-30 per cent (Language Proficiency Perception Survey, 1994), to 12 per cent (Surry, 1994). Such a wide variation almost certainly reflects in part the attitude and methodology of the surveyors, and the authors of the most recent survey on perceptions of language proficiency are almost certainly correct when they conclude in a somewhat frustrated tone that perception 'is a two-way phenomenon: it depends on the perceived as much as on the perceiver' (Language Proficiency Perception Survey, 1994: 3). Surveys of language use should distinguish between the spoken and written language, resulting in a four-part matrix.

Table 14.1 *Knowledge of English in Hong Kong, 1983-1993 (self-reporting)*

Response	1983	1993
Not at all	33.1	17.4
Only a few sentences	23.5	21.7
A little	36.2	27.2
Quite well	4.7	26.6
Well	-	3.3
Very well	0.4	3.8

Note: The question asked was 'How well do you know English?'

Within this matrix, the most reliable category is spoken Chinese, which typically means Cantonese in this overwhelmingly Cantonese-speaking city, where we might assume little or no variation in ability. With the other three categories-written Chinese, spoken English, and written English-it is fair to assume that there is a wide range of abilities and proficiencies.

In the case of English, the lowest figure of 12 per cent, for which further details and methodological information are not available, tends to reflect thinking in parts of the business community, where the demand for English seems to have run ahead of supply. This in tum seems to reflect the rather rapid shift in industry from manufacturing to service. The upper estimate quoted above from Bacon-Shone and Bolton (1995) is based on self-reports of proficiency. A comparison of an earlier 1983 (Bolton and Luke, 1985) survey with the 1993 survey is given in Table 14.1 (from Bacon-Shone and Bolton, 1995: 27, table 25).

The problems of interpreting survey results based on self-reports are illustrated here. Categories such as 'a little' and 'quite well' are open to a wide range of interpretation. One clear result that does emerge over the ten-year period, however, is that there is a noticeable drop in the numbers of those claiming no knowledge of English at all. If we take the aggregate total for 'Quite well', 'Well', and 'Very well', we obtain a total percentage of 33.7, which broadly agrees with the census figure mentioned above.

The middle-range estimates of 29 and 20-30 per cent, representing respectively a population-wide census *(Hong Kong Population Census,* 1991) and a very focused group of 7,200 students, educators, graduates, employers, and parents (Language Proficiency Perception Survey, 1994) seem to give two further reasonable indications of the proportion of the Hong Kong population with a better than minimal knowledge of English. By conflating these latter two estimates with the total of 'better' ('Quite well', 'Well', and 'Very well') English users derived from Bacon-Shone and Bolton (1995), we are able to conclude that a figure of about 30-35 per cent of the population represents a realistic estimate of the proportion of educated English users in Hong Kong.

That said, this figure represents the maximum number of possible informants. When it came to the reality of data-collection a series of hurdles existed between identifying who used English and their inclusion in the ICE corpus. The first hurdle was actual, as opposed to simply potential, use of English within one of the ICE categories. The second was the opportunity of recording a text or obtaining copyright to it. The final hurdle was obtaining the permission of the speaker or writer and his or her relevant biographical information. Singularly and collectively, these hurdles have served to reduce the number of informants in practice, because people either elect to use Chinese rather than English, decline to be recorded (or decline to give permission to use material already broadcast), or tum out to have the wrong biographical detail.

3. The sourcing of ICE texts in Hong Kong

The second assumption we made was that 'this widespread usage [of English] would be reflected in a reasonably straightforward identification of text sources which would match the text categories required by the ICE project.' It was further noted that 'In making these assumptions we were partly right and partly wrong and certainly over-optimistic.' (Bolt, 1994). Throughout the data-collection process in Hong Kong, we encountered a number of rather delicate problems related to individual proficiencies in English and the reluctance of potential participants to co-operate fully in the ICE-HK project. Problems related to real or perceived proficiency were particularly· sensitive, since a major aspect of this study is the analysis and comparison of second or foreign language performance in English with the performance of native speakers. In addition, data-collection was not always helped by the constant public, and not necessarily well-informed, debates in the English language press and elsewhere, about the 'decline' in standards and the generally poor level of language proficiency achieved locally. [2]

Complementing the issue of proficiency in English is the parallel issue of the use of English in Hong Kong, i.e. who, of our target group, uses English, under what circumstances, and for what purposes? While recent studies (Bacon-Shone and Bolton, 1995) have recorded some intrusion of English into private life where no non-Cantonese speaking person is present, typically, the 'extended' use of English is found in the formal, public contexts, where, historically, the presence and significance of the non-Cantonese speaker has been greater. The important word here is 'extended' as we have not included 'intermittent' uses of English, as in the case of code-switching and code-mixing, which occur frequently in both spoken and written Chinese. There is undoubtedly a degree of status related to such uses, but the fragmented nature of the language, especially in written text, which tends mainly to involve names or short phrases, does not fit with our focus on the extended use of language. The ICE text types may be considered in relation to a number of areas of language use, all reflecting the productive, rather than the receptive, perspective and all embracing both spoken and written forms. These are shown in Table 14.2. Grouping the texts in this way permits an easier discussion of text sourcing in relation to broad areas of use and of the factors which make for easy or difficult text collection. Note that 'administrative and regulatory' appear in both governmental and commercial contexts.

Table 14.2 *ICE text types and data collection in Hong Kong*

Area of use	ICE text categories	Number of examples	
		ICE	Hong Kong
Educational	Class lessons, timed and untimed essays, and learned informational writing	80	80
Government and law	Parliamentary, legal cross-examinations and presentations, administrative and regulatory writing	35	35
Media/publishing	Broadcast interviews and discussions, broadcast news and talks, press reporting and editorials, skills and hobbies, creative writing, and popular informational writing	170	130
Corporate	Business transactions, administrative and regulatory writing, and business letters	30	30

Formal public	All speeches, demonstrations, commentaries, and administrative and regulatory writing	70	70
Private	Private conversations, distanced conversations and social letters	115	105
Total		500	450

3.1 *Educational*

The claimed medium of instruction in many secondary schools and all post-secondary institutions (with the notable exception of the Chinese (language) University of Hong Kong) is generally English, although this is by no means an absolute policy. The medium of instruction, like the language standard issue, has been the subject of much, still unresolved, discussion for a number of years. A recent study (Language Proficiency Perception Survey, 1994: 6-1) recorded the fact that Cantonese is in fact quite widely used within secondary and tertiary education, increasingly so in the latter case as the group size becomes smaller and the situation less formal. The relevant percentages for language use are given in Table 14.3.

Table 14.3. *Language use in Hong Kong colleges and universities (%)*

Mode	Cantonese	English	Putonghua	Cantonese and English	Other
Lectures	16	69.3	0.9	12.7	1.1
Small-group teaching	23.6	56.6	0.0	16.0	3.8
Tutorial (sic)	24.5	54.7	0.5	16.0	4.3

Nevertheless, such current shifts in the medium of instruction, if indeed they are current, have had little effect on our collection process, with the consequence that class lessons-which range from being close to monologues to being very interactive-and student essays have been collected in their entirety. The effect of this ease of identification and collection has been to make possible early comparative analyses of student writing in Hong Kong and the United Kingdom.[3] The final text category within education may require some explanation. While class lessons and student essays fit easily into a broad educational area, it should be noted that the 'informational learned' category is also included here, as all such texts that we have collected are written by academics, mainly for an academic audience for publication in journals.

A major issue to be aware of with most of the written data, and this is doubtless true for all ICE teams, is that published texts (including anything which is electronically produced), are almost certainly subject to editing. In particular, it must also be noted that there is a tendency to recycle text in different places, and, for us, it has been necessary to read with some attention to ensure that a particular text was both suitable and original.[4]

3 .2. *Government and Law*

The most visible governmental activities are the weekly session of the Legislative Council, its committee sessions, the open committee sessions of other agencies such as the Housing Authority, and the written ordinances and instruments issued by such agencies. Hong Kong is an executive-led government with only rudimentary political parties and little form of government

and opposition. Government tends to consist of a mixture of expatriate (mainly British) officials and local officials, while the members of the Legislative Council are, with a small number of notable exceptions, local people within our definition above. The lack of a government and-opposition framework means that the adversarial style of debate, which is typical of the British. parliament, is largely missing in the local legislature, a fact underlined by the very prominent, government-directed role of the President of the Council. Legislative Council sessions are broadcast, albeit only in part, so that this data was easily obtainable, but it is very much on the outer edge of the broad category of dialogue. To offset this, we have also obtained recordings of a number of committee and panel meetings where there is a greater amount of interaction.
A further complication with the full Legislative Council sessions is the use of Cantonese, even by those members who are fluent in English. In many cases this is a matter of conscious choice.

The use of English in the judicial system reflects the fact that English law is used and the administration of justice is based largely on English practice. In addition, a relatively large number of participants in the process are native speakers of English-judges, counsel, and jury members. A proficiency in English is a prerequisite for jury service. However, the existence of a substantial proportion of local judges and counsel, who were very helpful in the data-collection process, means that both legal categories have been successfully collected. However, while the legal presentation, as a form of monologue, probably compares with that of other ICE teams, the fact that witnesses and defendants are free to use English or Cantonese, and that most choose the latter, means that our cross-examination dialogues are mostly between the relevant counsel and an interpreter. Of some fifteen cases that we have had access to only one had a witness who spoke in English and that person declined to be included in our project.

3.3. *Media and Publishing*

The categories of broadcast texts-interviews, discussions, news, talks, and (usually) spontaneous commentaries-illustrate very well the range of issues attached to sourcing spoken English in Hong Kong. Sourcing printed texts illustrates the range of issues associated with the use of multiple editors and authors in the local print media. For the broadcast texts we were able to source, with increasing degrees of difficulty, the ten interviews, the twenty discussions, ten (out of twenty) broadcast news items, ten (out of twenty) talks (but with major qualifications below).
In addition, twenty commentaries, but none easily comparable to the prototypical commentary of the sports or ceremonial event, have been collected (below).

Interviews, ranging from short snippets in news programmes to longer in-depth sessions, were relatively easy to obtain. The interesting feature about such interviews, however, is that in almost all texts the interviewer is a native speaker of English and the interviewee a local person. Discussions proved extremely time consuming to complete, as there were very few sources (and there are even fewer now), and a number of people who appeared on the current affairs programme, 'Newsline', our major source for such discussions, were prominent people who also appeared in other contexts. Our other major source, Metro Radio, only broadcast two suitable programme series over a two-and-a-half year period. Again, however, with the exception of the male host of 'Newsline', the interviewer-interviewee divide is very similar to that found in the interviews. A further practical consideration was that in many discussions not only was the interviewer a native speaker of English (or the regular local presenter of 'Newsline', of whose speech we have restricted ourselves to the 2,000-word individual text amount), but one or more of the other discussants was a native speaker, resulting in a large measure of extra corpus text.

Television and radio news reading is typically anchored by expatriate presenters who are sometimes western, sometimes Asian, and sometimes overseas-born Chinese. The number of local (in our sense) news-readers is small, and this explains our collection of only half the targeted number of texts. In addition, these are somewhat fragmented texts, and therefore, in common with interviews and discussions, the number of subtexts is inflated in comparison with the comparable texts for those ICE teams where all participants are generally suitable for inclusion in the final corpus. Broadcast talks in which one person speaks uninterruptedly on a subject for any reasonable length of time are very few, 'Letter from Hong Kong' being, until 1995, the only example. This programme has tended to be dominated by expatriates, with only a small proportion of local speakers. In mid-1995 the format was changed so that a previously twelve to fifteen minute independent programme became a segment of an afternoon magazine programme. A small number of part-texts have been obtained from this source. To make up our total of ten texts we have included a number of narratives/voice-overs to current affairs type documentaries.

The range of printed and published material includes press reporting and editorials, popular informational writing, creative writing, and skills and hobbies. As with the broadcast texts, these five categories illustrate many of the relevant sourcing issues and the resultant state of collection. Press (newspaper and magazine or journal) reporting and editorials have both been fully collected. There are now three English-language newspapers and a number of the news reports are by-lined by local people. However, two interesting issues have emerged: first, the comparative use of local reporters in the Asian region in English-language newspaper publishing, and, secondly the extent and nature of editing of writers' copy which is undertaken.
One of the authors (Bolt) undertook a study of regional English-language newspapers over a three-day period and it yielded the comparative author information shown in Table 14.4. The far lower percentage of local writers in Hong Kong is very evident. In addition, local writers in

Table 14.4. *The use of local and foreign reporters in selected English-language dailies*

	Local writer	Foreign writer	Agency reports	Proportion of local writers (%)
South China Morning Post (Hong Kong)	154	292	108	25
Straits Times (Singapore)	203	36	200	46
New Straits Times (Malaysia)	358	55	142	64
Manila Bulletin (the Philippines)	410	3	81	82

Hong Kong tend to be concentrated in the local and China news sections, very few penetrating other kinds of news stories, non-news features, the arts and entertainment, or sport, while special features are invariably authored by expatriate writers.

The editor of the major local English-language newspaper noted recently that most copy needs more than just a little attention and that the worst copy needs rewriting from top to bottom, commenting that 'Even where the structure and broad style are more or less acceptable, grammatical errors and clumsy expression can demand heavy editing' (Braude, personal communication). Newspaper editorials have usually been written, until recently, by expatriates, or by writers who have had much of their education in an English-speaking country. However, a few local writers have recently begun to produce editorials and some of these have been

collected, although the biodata has not been made available. In addition to these, our editorials include an English-language editorialfeature from a daily Chinese language newspaper, some from a student publication and others from a number of magazines.

As with 'creative writing' and 'skills and hobbies' (below), sourcing texts for popular informational writing is not simply a case of browsing along the magazine racks or shelves of booksellers. Although a very wide range of magazines are available, very few are relevant because they are either imports, or are written by expatriates, or are in Chinese. A few that do have English along with Chinese articles employ translators to make a parallel text in English, and many of these translators are expatriate. As such, the sources of our popular writing are books, a small number of magazines and a small number of trade journals; for humanities, (only five) books; for technology, magazines and journals; and for the natural science category we have no texts at present. ·

In the area of creative writing we have identified only three or four published local writers, so we have decided to extend the interpretation of published to 'written to be published' in order to obtain a greater number of texts. This is an area in which there are few easily, publicly identifiable and available texts. Texts in the area of skills and hobbies, written by people having the right biodata, have not been sourced. Many magazines which have an English title or some English text tum out to be not suitable, as either the amount of English is too small, often restricted to names, or they have been authored by someone who is not a Hong Kong person.

3.4 *Corporate*

The relative degrees of difficulty in sourcing business transactions, administrative and regulatory writings, and business letters have provided interesting contrasts. Business transactions in the corporate and commercial world have been very difficult to obtain, issues of confidentiality being the usual barrier. As such, the notion of a business transaction has been interpreted very widely, covering the discussion and exchange of ideas rather than simply monetary issues. As a result of this, two thirds of our texts have been collected in the education sector-meetings of various degrees of formality, in the areas of research, faculty, and staff/management-with just four from the commercial sector. For written data, the experience has been much happier, such that both categories are almost complete.

3.5. *Formal Public*

Unscripted speeches, scripted, but not broadcast, speeches, demonstrations, and commentaries have been classified here as 'formal public' to reflect the fact that the general context of delivery is in front of an audience whose composition is usually not entirely known in advance. The qualification of 'usually' indicates that a number of our demonstrations have been recorded as part of an ongoing course of sessions, and a small number of speeches have been made to committees whose composition is known in advance.

The distinction between 'scripted' and 'unscripted' has not been easy to apply in practice as there are few occasions when the speaker fails to depart from a prepared text and few occasions where a speech is entirely unscripted, since speakers often resort to memorization or notes (including graphical aids). Of the largest category, unscripted speeches, many of our texts are from industry seminars, some from conferences, and some from organizations such as the Toastmasters' Club.

Demonstrations include computer software, cookery, nursing, and first aid demonstrations. The first fits squarely into the category of not knowing the composition of the audience in advance, whereas, for the latter three the ·audience was partly known as the session was part of a course. Of these, the cookery demonstration is an especially good example of the genuine use of English, as the audience consisted of Filipina 'amahs' (domestic servants).

Spontaneous commentaries represent an area where we have made our most significant departure from the normal meaning of ICE categories. Sports and ceremonial commentating reflects a sociolinguistic divide, with no local commentators performing in English for the major sport of horse-racing, the reasonably popular sport of tennis, or for minority sports or events, such as the dragon boat-racing. Interpreting widely both the individual terms 'spontaneous' and 'commentary' and their combined meaning, our spontaneous commentaries are six tourist guides, seven interpreters for the Legislative Council, and seven interpreters from the high court. Our motivation for using such texts is that the tourist guides are in fact reasonably flexible in what they talk about, even though they follow a fixed itinerary, and there is, in all cases, quite a lot of ad-libbing by the speakers. For the two interpreting situations, it is suggested that although there is no absolute spontaneity involved there is a high degree of having to respond to (linguistic) input in real time and, given the nature of virtually simultaneous interpretation, that this is in fact a form of commentary, if only on meaning and message. In the case of the high court, such interpretation is of vital importance and, on more than one occasion, its accuracy and the possible effects of inaccuracy have been raised. It has, however, proved impossible to obtain the biodata of the speakers for the interpretation, although it is highly unlikely that anyone other than a Cantonese native speaker would be able and likely to do this work. Nevertheless, our spontaneous commentaries as a whole may remain something of a rogue collection.

3.6. *Private*

Private conversations (75) and social letters constitute our collection in the private area. For a number of logistical reasons we have not been able to collect distanced (telephone) conversations; this is somewhat ironic given the immense local popularity of the device, especially the mobile variety. In the case of private conversations, as noted above, without the presence of a non-Cantonese speaker, Cantonese speakers tend to use Cantonese. This meant that to ensure that such conversations were as natural as is possible, given the fact that the speakers knew about the recording in advance, the presence of a non-Cantonese speaker was necessary. In fact this is the case with the majority of our spoken texts. Another important factor to note is that the vast majority of private conversations involve speakers aged between 20 and 30. A similar age-range exists for social letters, where we should also note that approximately half of such texts are email messages. This reflects the popularity of this channel of communication among students and the fact that email is only available in English on campus networks. A number of social letters have also been obtained from Hong Kong students studying abroad.

4. Processing the data – Tagging and parsing

What of the nature of the language found, our first motivation? We are not yet able to report global results because of the time it has taken us to complete the data collection. There remains, however, the question of how the data may be further processed, particularly of how it is to be tagged and parsed. The tagging stage is prior to the parsing stage and provides the input for

parsing, so much so that a number of ICE word-class categories reflect syntactic patternings, e.g. the various tags for 'it', and the range of transitivity options for verbs. Our problem can be seen from the following output from AUTASYS 3.0 on sentences taken from a business letter, student writing, and an editorial.

Sentence 1	
Regarding	V(cxtr,ingp)
the	ART(def)
mistake	N(com,sing)
caused	V(montr,edp)
by	PREP(ge)
our	PRON(poss,plu)
staff	N(com,sing)
who	PRON(rel)
was	AUX(pass,past)
overlooked	V(montr,edp)
and	CONJUNC(coord)
delayed	V(montr,edp)
the	ART(def)
check-in	N(com,sing)
for	PREP(ge)
the	ART(def)
guests	N(com,plu)
	PUNC(per)

Sentence 2	
Despite	N(com,sing)
in	PREP(ge):1/3
terms	PREP(ge):2/3
of	PREP(ge):3/3
property	N(com,plu):1/2
relations	N(com,plu):2/2
they	PRON(pers,sing)
might	AUX(modal,past)
have	V(montr,infin)
none	PRON(neg,sing)
,	PUNC(com)
they	PRON(pers,sing)
admire	V(montr,infin)
the	ART(def)
life	N(com,sing):1/2
style	N(com,sing):2/2
of	PREP(ge)
that	PRON(dem,sing)
of	PREP(ge)
the	ART(def)
upper	ADJ
class	N(com,sing)
and	CONJUNC(coord)
is	V(cop,pres)
in	PREP(ge)
attitude	N(com,sing)
conform	V(intr,pres)
to	PREP(ge)
conservative	ADJ
ideas	N(com,plu)
and	conjunc(coord)
resistance	N(com,sing)
to	PRTCL(to)
change	V(montr,infin)
	PUNC(per)

Sentence 3	
If	CONJUNC(subord)
a	ART(indef)
lecturer	N(com,sing)
finds	V(ditr,pres)
his	PRON(poss,sing)
superiors	N(com,plu)
involving	V(montr,ingp)
in	PREP(ge)
administrative	ADJ
indiscretions	N(com,plu)
and	CONJUNC(coord)
having	V(montr,ingp)
errors	N(com,plu)
of	PREP(ge)
judgment	N(com,sing)
,	PUNC(com)
it	CLEFTIT
is	V(intr,pres)
his	PRON(poss,sing)
duties	N(com,plu)
,	PUNC(com)
and	CONJUNC(coord)
rightly	ADV(ge)
so	CONJUNC(coord)
,	PUNC(com)
to	PRTCL(to)
report	V(montr,infin)
the	ART(def)
situation	N(com,sing)
to	PREP(ge)
someone	PRON(ass,sing)
higher	ADJ(comp)
up	ADV(phras)
in	PREP(ge)
the	ART(def)
bureaucratic	ADJ
hierarchy	N(com,sing)
before	PREP(ge)
more	PRON(quant)
damages	N(com,plu)
are	AUX(pass,pres)
done	V(montr,edp)
	PUNC(per)

The problem here is not that a few words are incorrectly *tagged-regarding* and *despite* in the first and second sentences respectively might better be labeled as prepositions-and the *it* in the main clause of the third sentence seems

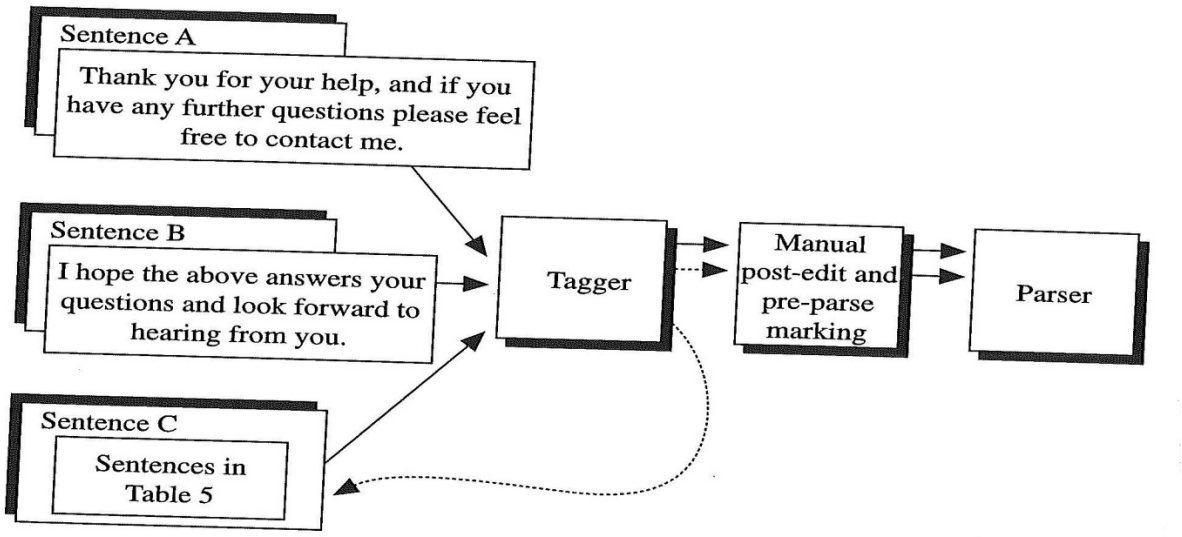

Fig 14.1 Output from AUTASYS 3.0 for standard and non-standard written text

uncomfortably classified as a CLEFT. Rather, the problem is that the range of tags available for local tag disambiguation may not provide the best guide for the parsing stage. As opposed to the local-typically three item context-of tagging programs, parsing involves the whole sentence. The constructions of interest in the sentence, phrases and clauses, might not be easily determinable on the basis of the wordclass information given in this output alone. This may be either because such information as provided by the tagger is not the· best guide to phrase and clause structure, but can be recovered manually, or because the raw data does not conform to the descriptions of 'standard' English grammar. There are three possible relations between the original sentence, the tagger, and the parser, as shown in Fig. 14.1. For Sentence A, the phrases and clauses of which conform to standard English grammar, the route to a parsed output is relatively straightforward, requiring minimal post-tagger marking, e.g. for clause boundaries, as shown by the directions of the arrows in Fig. 14.1. For Sentence B, again a well-formed sentence but one whose structure forces the tagger into minor mis-labelling, this mis-labelling can be corrected manually between the tagging and parsing stages. Again, the arrows in Fig. 14.1 indicate an onward progression to the post-tagger stage, where the broken line indicates that there is some mis-labelling of one or more word-forms which will require attention as evidenced below.

I	PRON (pers, sing)
hope	v (mont, pres)
the	ART (def)
above	ADJ
answers	N (com, plu)
your	PRON (poss)
questions	N (com, plu)
and	CONJUNC (coord)
look	v (intr, pres)
forward	ADV (ge)
to	PREP (ge)

hearing	N (com, sing)
from	PREP (ge)
you	PRON (pers)
.	PUNC (per)

However, for Sentence C, of which the three sentences displayed above are typical examples, it is not simply the case that some mis-labelling needs to be adjusted prior to parsing. Rather, not only do the original sentences not provide the right input to generate the level of tagging accuracy as produced for sentences A and B, but, more crucially, whatever tagging results are obtained, they cannot be altered sufficiently to provide a sensible input to the parser. It is not a problem with the tagger' s efficiency in these instances but with the original sentence, hence the direction of the even more broken and backward-looping arrow in Fig. 14.1. There would seem to be two possible strategies for coping with these sorts of sentences. Either a distinctive tag-set could be created for such sentences, which would incorporate the necessary additional, possibly first-language-specific information needed to produce some semblance of phrase and clause structure, or the sentences could be 'normalized' to bring them syntactically into line with the intention of the writer. Although this latter approach would be relatively simple in the case of sentence 3, neither option is attractive given the comparative objective of ICE.

Although other ICE teams may have syntactically odd· constructions, it is anticipated that up to 35 or 40 per cent of our data, including written texts, may have the kind of features shown in these sentences. The decisions regarding how to proceed, therefore, have far-reaching implications, especially as we wish for consistency of labeling and structure with other ICE teams. These issues will be the subject of discussion in the near future.

5. Conclusion

The complex and shifting realities of the current sociolinguistic context have had an obvious influence on the process and results of the data collection in Hong Kong. In some cases these have made for easy collection-the use of English in the education system, in government, in the judiciary, in large businesses, and an English-language press and publishing industry-have all made sourcing possible, if not always procedurally straightforward. In other cases, the relative paucity of English use has made for difficulties. In all cases, there has sometimes been the delicate link referred to above between the existence of a text and its author or speaker and its eventual inclusion in ICE-HK.

The dramatic spread of English through younger age-groups in Hong Kong has been commented on by Bacon-Shone and Bolton (1995), who explain this by reference to the huge expansion of secondary education in the 1970s and a similar expansion of university education in the late 1980s and early 1990s. On this they comment that at present 'more people than ever are speaking "good" English, and more people than ever are speaking "bad" English' (33), and we have noted the predominance of younger people in the 'private' category. At the same time, however, counter-tendencies can also be seen, particularly within the public domains of language use, and they have been evident in the data sourcing and collecting stages of the ICE-HK work, especially in the last two years. In particular, there has been a significant increase in the amount of Chinese used, alongside or instead of English. Within the government and the judiciary-which are actively promoting the future use of Chinese-in the meetings of many public bodies, and throughout education-the Education Department has for a long time been trying, without a great

deal of success, to persuade more schools to use Chinese as the medium of instruction-the use of Chinese is on the increase. Conversely, English-language teleyjsion and radio broadcasting seems to be declining rapidly in both scale and importance and we have noted above the compar. actively low, in regional terms, local input to the English-language press.' In this sense, it may be that our work coincides with a high-water mark in the use of English in Hong Kong and that it will represent the last large-scale exercise of this nature that achieves a near 90 per cent data collection rate across a wide range

This reflects the unique geopolitical context of Hong Kong with its scheduled reversion to Chinese sovereignty soon to be completed. With the change from being within the British Commonwealth to being an albeit special part of China it is difficult to imagine a decline in the use of Chinese at the end of the twentieth century, even given the complexities of the varieties of spoken Chinese. Yet English is fairly well-rooted in certain areas of public activity, especially those areas which have an international interface. Those people presently proficient in English are unlikely to lose this facility, even if there is less opportunity to exercise it. However, for those not especially proficient in English or those individuals (and organizations) contemplating language choice, the limitations of time, if not political correctness in its more literal sense, may begin to sharpen perceptions regarding language acquisition and use in the not too distant future.

Notes

1. Article 9 of the Basic Law, which serves as the proposed constitution for the new Special Administrative Region, gives some, albeit not entirely unambiguous, indication of this when it declares that: 'In addition to the Chinese language, English may also be used as an official language by the executive authorities, legislative and judicial organs of the Hong Kong Special Administrative Region' (Basic Law, 1991).
2. A recent issue (24 Feb. 1995) of one of the three English language newspapers *Eastern Express* is a not untypical example. There are three items relating to poor standards, one concerning the views of certain New Zealand schools who are worried about the influx of Asian non-English-speaking schoolchildren, one about the comments of a UK academic noting the increasing care which must be taken with language standards of non-maths and science applicants from Hong Kong, and a final letter to the editor containing a blistering attack on the Education Department and the proficiency of students emerging from the education system.
3. Two M.Phil. studies on discourse particles in broadcast interviews in ICE-GB and ICEHK, and another on student writing in the two corpora are underway.
4. In a recent case a senior academic admitted to using a former teacher's text in a textbook. What made matters worse was the somewhat cavalier fashion in which a more senior academic at the same institution dismissed the use of other people's work in a textbook as 'no big deal'.
5. Interestingly, we note that Pennington and Yue (1994) arrive independently at a similar conclusion on 'countertrends' in the use of languages in official and public domains.

References

BACON-SHONE, J. and BoLTON, K. (1995), 'Charting Multilingualism: Language Censuses and Language Surveys in Hong Kong', in M. C. PENNINGTON (ed.), 15-24.

Basic Law (1991), *The Basic Law of the Hong Kong Special Administrative Region of the People's Republic of China* (Hong Kong: Joint Publishing Co.).

BOLT, P. (1994), 'The International Corpus of English Project-the Hong Kong Experience', in U. FRIES, G. ToTTIE, and P. ScHNEIDER (eds.), *Creating and Using English Language Corpora: Papers from the Fourteenth International Conference on English Language Research on Computerized Corpora* (Amsterdam: Rodopi).

BOLTON, K. and KWOK, H. (1990), 'The Dynamics of The Hong Kong Accent: Social Identity and Sociolinguistic Description', *Journal of Asian Pacific Communication,* 111: 147-72.

--and LUKE, K. K. (1985), 'The Sociolinguistic Survey of Language in Hong Kong: The Background to Research and Methodological Considerations', *International Journal of the Sociology of Language,* 55: 41-56.

FU, G. S. (1987), 'The Hong Kong Bilingual', in R. LoRD and H. N. L. CHEUNG (eds.), 27-50. *Hong Kong Population Census: Summary Results* (1991) (Hong Kong: Census and Statistics Department).

Language Proficiency Perception. Survey (1994), Report prepared by the Education Commission Working Group on Language Proficiency, unpublished mimeo. (Hong Kong: Hong Kong Polytechnic University).

LORD, R. and CHEUNG, H. N. L. (1987), *Language Education in Hong Kong* (Hong Kong: The Chinese University Press).

LUKE, K. K. and RICHARDS, J. C. (1982), 'English in Hong Kong: Functions and Status', *English World-Wide,* 3: 47-63.

PENNINGTON, M. C. (1995) (ed.), *Language in Hong Kong at Century's End* (Hong Kong: Hong Kong University Press).

--and YUE, F. (1994), 'English and Chinese in Hong Kong: Pre-1997 Language Attitudes', *World Englishes, 1311:* 1-20.

RICHARDS, J. C. and LUKE, K. K. (1982), 'English in Hong Kong: Functions and Status', paper presented at 16th RELC Seminar, Singapore.

So, W. C. D. (1987), 'Searching for a Bilingual Exit', in R. LORD and H. N. L. CHEUNG (eds.), 249-68.

--(1992), 'Language-based Bifurcation of Secondary Schools in Hong Kong: Past, Present and Future', in *Into the Twenty First Century: Issues of Language Education in Hong Kong,* 69-95 (Hong Kong: Linguistic Society of Hong Kong).

SURRY, M. (1994), 'English not Spoken Here', *Window,* 1 Apr., 33-7.