Spoken Interaction with Computers in a Native or Non-Native Language — Same or Different?

Nils Dahlbäck[2], Seema Swamy[1], Clifford Nass[1], Fredrik Arvidsson[2], Jörgen Skågeby[2]

[1]Dept. of Communication
Stanford University
Stanford CA 95305
USA

{nass, sseema}@stanford.edu

[2]Dept. of Computer and Information Science
Linköping University
SE-581 83 Linköping, Sweden

{nilda, frear, jorsk}@ida.liu.se

Abstract: This study examines the effects of interacting with voice interfaces in an ingroup or an outgroup accent, for both native and non-native but competent English speakers. In a balanced, between-subjects experiment, (N = 96), ingroup and outgroup participants were randomly paired with one of two types of computer speech outputs: 1) computer speech output accent which matched the participants' accent or 2) computer speech output accent which mismatched with the participants' accent. The content of the output was identical in all the conditions. Participants' matched with accents similar to their own showed strong similarity-attraction effects. The matched users 1) disclosed socially undesirable behaviors they engage in, to a much larger extent, 2) found the interviewer to be endowed with more socially rich attributes, 3) perceived the interviewer to be more sociable. In short, similarity of accent is more important than 'correctness'of the accent when interacting with a computer. We discuss implications of these results for HCI design.

Keywords: Cross-cultural communication, speech interfaces, native and foreign accents, similarity-attraction effect.

1 Introduction

International and cross-cultural issues are becoming increasingly important for the HCI community. It is becoming abundantly clear that it is not enough to simply translate the text of e.g. a website into another language. Cultural differences are more than linguistic differences. The work of Hofstede (1997) is often taken as a starting point for this (for an introduction to the HCI-issues here and the work of Hofstede see e.g. Marcus & Gould, 2000). Most of the current work is concerned with analyzing how to translate and adapt interfaces to different cultures.

But there is also another aspect to the internationalization process. More and more people are interacting with computers using a language that is not their mother tongue. Software is often first developed and released for English speaking users, and for especially advanced tools and technology, there exists only an English version. In other cases users do not want to wait for the tools to be translated into their language. And in multi-national organizations, both commercial and others, English is often the official language for communication. Consequently, large numbers of non-native English speakers use English versions of the software. But will the fact that these users are proficient enough to understand and express themselves in the English used in these situations make the cultural differences go away? That is the issue addressed in the work presented here.

Language is not only used for communicating propositional content. Linguists and other language researchers describe many functions of language and linguistic communication. A basic distinction is often made between the expression of 'content', and the function of expressing social relations and personal attitudes, by Brown and Yule (1983) called *transactional* and *interactional* respectively.

In most foreign language teaching the focus is on the communication of content, and in many countries the pupils reach an impressive level of competence in transactional language use. But not necessarily so for the interactional aspects, especially since this is more closely related to the cultural differences described by Hofstede and others, and therefore more difficult to teach. Furthermore, much of the interactional function is communicated using intonation and other so-called suprasegmental aspects of utterances that are closely related to the accent of the speaker. And it is well known that non-native patterns of speech production is almost inevitable in the speech of late second language (L2) learners (Flege, Munro & Mackay, 1997) or non-native learners learning the native language in non-native countries. Even individuals that learn a second language as early as 3.1 years of age may have a foreign accent (Long, 1990).

Taken together, the facts above suggest that the issues of cross-culture communication using a non-native language become especially critical for interfaces using spoken input and output. We have therefore focused this study on the use of spoken interaction with computers

1.2 Ingroup and outgroup accents

An accent acts as a cue identifying a speaker's group membership (Ladegaard, 1998). We form an impression of speakers based on their speech patterns and pronunciations (Giles, Williams, Mackie & Rosselli, 1995). A basic distinction is between an ingroup or an outgroup accent, where an outgroup accent is an accent that sounds distinctly dissimilar to the native speaker's pronunciation norms (Munro & Derwing, 1995). We evaluate people as superior, inferior, fun, or unattractive, based on their accents and stereotypes about their nationalities (Cargile & Giles, 1997). Stereotypes are mental concepts, pictures in our heads which govern the process of perception (Lippman, 1922), whether or not they represent social realities. There often exists a tension between the inherent value and the imposed norm of stereotypes, i.e., between the objective social reality and the constructed reality about people belonging to a particular nationality.

The effects of foreign and other non-standard language accents have been studied by researchers in intercultural communication. The so-called 'speaker evaluation paradigm' (Ryan, Giles & Sebastian, 1982) has studied the attitudes towards speakers with different accent. The evaluative reactions to speakers is here seen as a two-step process – 1) listeners identify of the social group membership from cues such as accent, age, race, and gender and 2) derive specific stereotypes about the speaker (Ryan, 1983). Two examples of this are Giles et al's (1995) work on the reactions to anglo- and hispanic-american accented speakers by US participants, and Ladegaard's (1998) work on the perception of British, American, and Australian language and culture in Denmark. The general finding from these and other studies is that prestige accented speakers are upgraded on traits of socioeconomic success relative to non-standard-accent speakers.

Even when the content of the speech is entirely understood, research shows that listeners discriminate against the speakers with foreign accent, display irritation, and exhibit prejudice in other ways (Anisfeld, Bogo & Lambert, 1962, Brennan & Brennan, 1981).

In addition, a combination of similarity-attraction theory and proximate source orientation (Nass & Lee, 2000), purports that when you match the accent of the user with the accent of the speech output, the user's reactions should affect not only to the voice but also to the speaker of the voice and the interaction itself.

What the research within this paradigm, however, does not address, are the actual effects, if any, that these attitudes have on the actual interaction. Does the interaction change merely because the accent of the dialogue partner or system is not the same as the speaker's? In the present study we have attempted to address both the attitude and the behavioral aspect issues involved.

If changes occur, in which direction will they be? There are two possible hypotheses here, pointing in opposite directions. The work on speaker evaluation described above suggests that speakers will prefer to interact with speakers with a native, and hence 'correct' accent. There is, however, an alternative hypothesis, based on the so-called similarity-attraction effect. As humans we prefer to interact with personalities that resemble our own (Byrne & Nelson, 1965). This is true also for computer interfaces (Nass & Lee, 2000, Ibister & Nass, in press). If we hypothesize that the similarity attraction effect also is true for other attributes than personality, this suggests that speakers will prefer to interact with a speaker with an accent similar to their own.

Which one of these hypotheses is the correct one is not only of theoretical interest. It has obvious practical implications for the HCI-community. If a native or 'correct' accent is the most positively valued accent for both native and non-native users of the software, software providers can with confidence use English only software also for that large and growing user population that has a good command of English as a second language. But if native and non-native users react differently to this accent, the issues of how to make software available to an international audience becomes even more important than what has hitherto been thought to have been the case.

In the present study we have addressed these questions by having both native and non-native users interact with a computer system using spoken input and output, where the computer's language also has either a native or a non-native accent, similar to the non-native users' accent.

2 Experiment

2.1 Method

The experiment was a full-factorial 2 (participants' nationality: American and Swedish) x 2 (computer speech output: American accent and Swedish accent) balanced, between-participants design. For each of the nationalities, a male voice and a female voice of a graduate student in each country was recorded. All recordings from both countries were in English.

The Swedish speakers of English were all native Swedish speakers with Swedish speaking parents. They had taken English as their first foreign language in school, which they had taken for approximately 8 years. They used English as a professional second language, both written and oral. Their pronunciation was closer to US English than e.g. British or Australian, but they had a noticeable Swedish accent. This was not, however, so strong that it impeded the understanding of their speech. (Samples of the voices can be found at http://www.ida.liu.se/~nilda/voices.html.)

Gender of output was balanced across condition, so that half of the participants in each condition heard a male voice and half heard a female voice. Similarly, all conditions were balanced for gender of participants.

To administer the experiment, we used the CSLU Toolkit for presenting the voices and recording and recognizing participant responses. (For more information on the CSLU Toolkit, see http://cslu.cse.ogi.edu/toolkit/.)

2.2 Participants

Participants were 96 Caucasian students; 48 American students enrolled in an undergraduate class in a U.S. university and 48 Swedish students enrolled in undergraduate classes in a Swedish University.

All of the U.S. students were native speakers of English. All the Swedish participants were born in Sweden by Swedish speaking parents. They had taken English as their first foreign language in school, which they had studied for 7-9 years, with approximately 2-3 lessons a week. We excluded all participants who had lived for longer periods in an English speaking country as a child or as an exchange student, but shorter visits, for instance on holiday, were accepted.

The Swedish participants could all use English in casual conversations, and most of them were used to using textbooks and other literature in English in their university studies.

Since we did not want the participants to focus on the fact that the study was interested in effects of the fact that they had to interact with a computer in English, we screened the participants only *after* the participation in the experiment. 17 participants were excluded because they did not match these criteria, and are therefore not included in the figures presented here.

The ages of the participants ranged from 18 years to 24 years. The US participants received course credit for participating in the experiment, while the Swedish participants were paid two movie tickets for their participation. The participants were randomly assigned to the four conditions, with gender of participant and gender of voice balanced across conditions.

2.3 Procedure

Upon arrival, participants were required to read a consent form and sign it. Participants were assured that the information submitted was confidential. Each participant was fitted with a hands-free microphone to respond to the questions posed by the computer. The voice from the computer was audible through two speakers placed on either side of the computer. The experimenter demonstrated the manner of interaction by walking each participant through the practice round. The practice round consisted of three questions similar to the actual interaction.

The first round asked questions derived from the BIDR-Impression Management (IM) subscale (Kroner, & Weekes, 1996). The original 20 BIDR items were first-person statements, for example, "I sometimes tell lies if I have to". To suit the interview nature of this study, the items were adapted to "Do you" or "Have you" questions, such as "Do you sometimes tell lies if you have to?" “Do you gossip about other people’s business?” “Do you ever take things that don’t belong to you?” “Do you ever read sexy books or magazines?” The computer screen displayed the numbers 1-7 on separate buttons, with anchors of “not true” and “true,” respectively (consistent with the original BIDR-IM).

To respond to this scale, the participants spoke the number corresponding to their response for each of the questions. The number spoken by participants in response to a question asked by the computer lit up yellow as a confirmation of the participants’ input.

The second round asked questions from Moon’s (1998) nine self-disclosure questions. A few example questions include: “What do you dislike about your appearance?” “What are some of the things that make you furious?” and “What characteristics of your friend really bother you?” These are open-ended questions. Participants dictated their answers to the system; although they believed that the system recognized their responses, in fact they were simply recorded for later analysis. The system continued recording until the participant stopped answering the question.

Participants were allowed to listen to all questions as many times as they wished by saying the word “repeat.” After responding to a question, participants could either proceed to the following question by saying “next” or go back to the previous question to change their response by saying “back.” English was used for both questions and answers for all participants, both American and Swedish.

After the participant had answered all the questions, a screen with a message thanking the participant for their participation popped-up and the participant was instructed to complete a paper-and-pencil questionnaire placed on the other side of the room. After the questionnaire was completed, the experimenter thanked and debriefed the participants.

2.4 Measures

The BIDR-IM is constructed so that higher scores indicate less self-disclosure. The 20 items were organized into a single index (Cronbach’s alpha = .81).

A second measure for self-disclosure was based on the number of words spoken in response to Moon’s (1998) self-disclosure questions. The index was highly reliable (Cronbach’s alpha = .72).

The closed-ended paper-and-pencil questionnaire was divided into four sections. The two sections asked how well a list of adjectives described the interviewer and the interview, and the participant’s feelings during the interaction. Each adjective was associated with a 10-point Likert-type scale anchored by “Describes Very Poorly” (=1) and “Describes Very Well” (=10).

Based on theory and factor analysis, we created three indices from the questionnaire. All indices were very reliable

Social Richness index was comprised of eleven items describing the interviewer: warm, lively, vivid, personal, accessible, sensitive, immediate, emotional, friendly, responsive, and sociable (Cronbach’s alpha = .86).

Liking of the interview was an index of five items: friendly, likable, nice, and pleasant (alpha = .85).

Sociableness was an index of five items describing the interviewer: friendly, likable, outgoing, pleasant, sociable, and warm (alpha = .85).

We asked all the Swedish participants whether there were any words they did not understand in the questions they were asked. For only 8 of the 1920 (0.4 %) questions asked were the Swedish participant uncertain of the meaning of a word, and only one participant asked for more than one word.

3 Results

All results are based on a full-factorial ANOVA. Controls for gender of participant and gender of voice did not have substantive effects on our conclusions, so they are not reported here.

3.1 Manipulation Check

All participants correctly identified the computer output accent as belonging to their nationality (ingroup) or a different nationality (outgroup). They did not, however, in all cases identify the outgroup accent correctly.

3.2 Balanced Inventory of Desirable Responding

There was a significant cross-over interaction effect with respect to the BIDR-IM, $F(1, 92) = 3.93$, $\eta^2 = .04$, $p = .05$, consistent with similarity-attraction (see Figure 1). Participants provided more honest responses to their ingroup accent as compared to the outgroup accent.

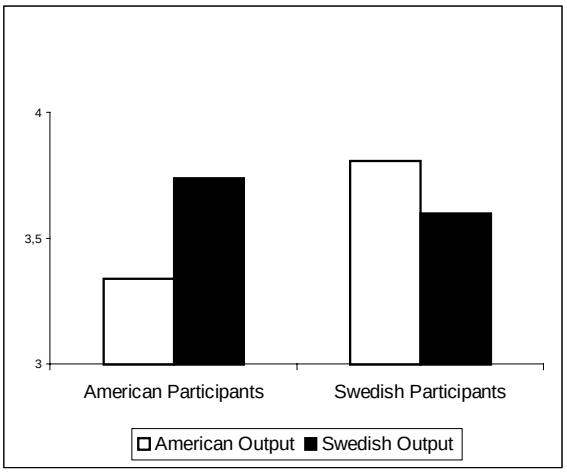

Figure 1: Balanced Inventory of Desirable Responding.

3.3 Moon's Self-Disclosure Items

There was no interaction between participant nationality and output nationality, $F(1,92) = 2.15$, $\eta^2 = .02$, $p > .14$ (see Figure 2). Americans did exhibit similarity-attraction, $F(1,46) = 4.5$, $\eta^2 = .09$, $p < .04$, but there was no effect for the Swedish participants, $F(1,46) = 1.90$, $\eta^2 = .04$, $p > .17$. The American participants spoke significantly more to the interface than did Swedish participants, $F(1,92) = 114.1$, $\eta^2 = .554$, $p < .001$. There was a significant effect for nationality of output, $F(1,92) = 6.19$, $\eta^2 = .063$, $p < .02$, but this finding was an artifact of the interaction.

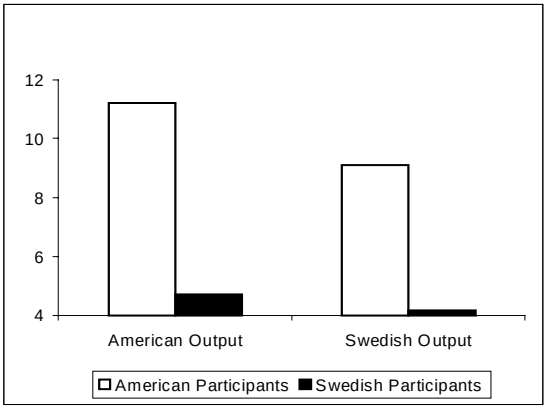

Figure 2: Word Count on Open-Ended Items.

3.4 Social Richness

There was a significant cross-over interaction with respect to social richness, $F(1, 92) = 4.70$, $\eta^2 = .03$, $p < .05$, consistent with similarity attraction (see Figure 3). U.S. participants found the U.S. voices to be more socially rich, while Swedish participants found the Swedish voices to be more socially rich. There were no main effects for participant or voice nationality.

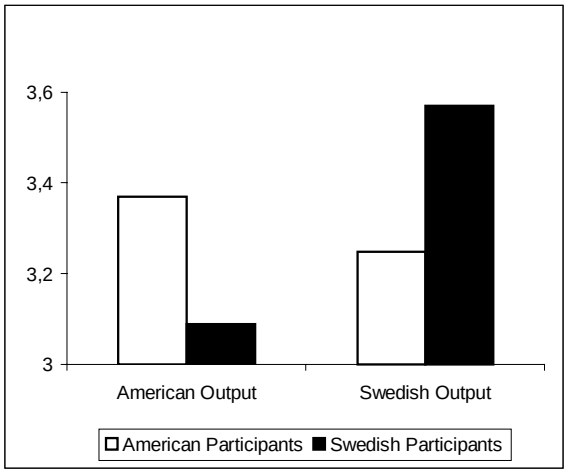

Figure 3: Social Richness

3.5 Liking of the interview

There was a significant cross-over interaction effect with respect to liking the interview, $F(1, 92) = 9.25$, $\eta^2 = .07$, $p < .01$ (see Figure 4). Consistent with the previous results, Swedish participants liked the inter view more when it was in a Swedish accent while

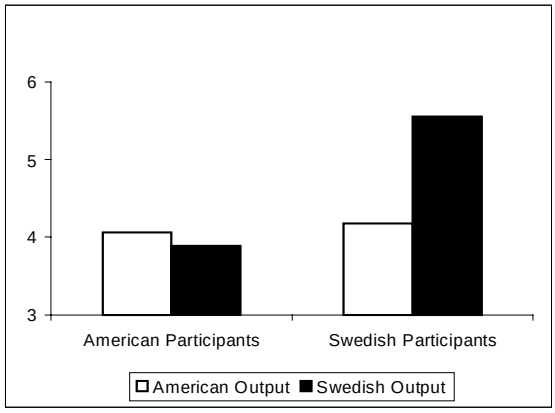

Figure 4: Liking of the Interview

U.S. participants liked the interview more when presented in a U.S. accent. There were no main effects.

3.5 Sociableness of the interviewer

There was a significant cross-over interaction with respect to sociableness of the interviewer, $F(1, 92) = 17.58$, $\eta^2 = .16$, $p < .001$ (see Figure 5). Once again, the results were consistent with similarity-attraction theory, as U.S. participants found the U.S. voice more sociable, while Swedish participants found the Swedish voice to be more sociable.

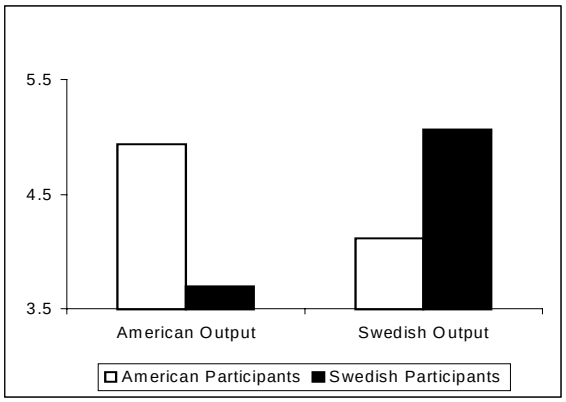

Figure 5: Sociableness of the Interviewer

4 Discussion

The central question addressed in this work is whether computer users prefer interacting with spoken interfaces using a native and correct accent, or whether they prefer computers with an accent similar to their own. Our results clearly show an ingroup preference. While the content of the output was identical in all the conditions, when participants' were matched with accents similar to their own they: 1) disclosed socially undesirable behaviors they engage in, to a much larger extent, 2) found the interviewer to be endowed with more socially rich attributes, 3) perceived the interviewer to be more sociable

Similarity-attraction theory provides a clear explanation for the interaction effects for the indices BIDR, Lombard's social richness scale, likability, and perception of sociability of the interaction where participants respond positively to speech output in an accent similar to themselves. This is supported by similar findings in a study, which found that participants are positively influenced by ingroup members (Nass & Lee, 2000).

That American users should prefer systems with an American accent is perhaps not all that surprising. But our results also suggest that the many computer users around the world that interact with computers in English, despite not being native English speaker, will prefer their own ingroup accent, even if a native accent has higher prestige and status. Whether this is true also for other cultural groups than those studied here, remains of course a question for further work.

The Swedish participants in our study did not experience any problems in using the system, or in performing the task. In fact, not one single Swedish participant expressed any surprise when they found that the study required them to interact with the system in English instead of Swedish. They have all done this often before, and are used to the situation.

But even if they experienced no problems in using the system, or performing the task, they did not react to the system like native English speakers. Our results therefore clearly indicate that issues of cross-cultural communication and issues of internationalization of software will remain, also for those non-native English speakers that can perform reasonably well in transactional aspects of communication.

What are then the consequences of our results for the HCI community, and especially for designers of spoken interfaces? When addressing the needs for native English speaking users the implications seems rather straightforward. Designers should use the accent that is the most widely accepted as the standard accent within the nation. Unless, of course, they in specific instances, for games and other forms of entertainment, actually want to create an outgroup reactions in the users. In this case our results show that also relatively small changes can create effects in users.

When, on the other hand, we come to the consequences of our results for all those users of English-based speech software, that are not native English speakers, it is less clear what the recommendations should be. The closest parallel to the situation in English speaking countries would be to suggest that the software should speak with the accent of the users. But creating English with a Swedish accent for Swedish users, a French accent for French users, etc., sounds like a rather odd recommendation.

In some cases, especially when using synthesized speech, and when no native language synthesis exists, this might of course actually be the best solution. But clearly more research is needed

to clarify the issues involved here. We have here focused on English, since English is becoming a *lingua franca* in much international work. And we have studied users from two countries that are similar in many respects. It is interesting to note that despite these similarities in culture, we find similarity-attraction effects of accent here. While one single study like the one presented here can not illuminate all the issues involved, the work presented here indicates that cross-cultural issues in spoken interfaces is an important area of future research.

Acknowledgements

This research was supported in part by SITI (The Swedish Research Institute for Information Technology) through its e-Society program, and the Swedish Graduate School for Human-Machine Interaction.

References

Anisfeld, M., Bogo, N., & Lambert, W. (1962). Evaluational reactions to accented English speech. *Journal of Abnormal Social Psychology 69*, 89-97.

Brennan, E. & Brennan, J. (1981). Measurements of accent and attitude toward Mexican American speech. *Journal of Psycholinguistic Research 10*, 487-501.

Brown, G. & Yule, G. (1983) *Discourse analysis.* Cambridge University Press, NY.

Byrne, D & Nelson, D. (1965) Attraction as a linear function of proportion of positive reinforcements, *Journal of Personality and Social Psychology Bulletin*, 4, 240-243.

Cargile, A.C. & Giles, H. (1997). Understanding language attitudes: Exploring listener affect and identity. *Language and Communication 17,* 3, 195-217.

Flege, J., Munro, M., & Mackay, I. (1997). Factors affecting strength of perceived foreign accent in a second language. *Journal of Acoustical Society of America 97* (5), 3125-3134.

Giles, H., Williams, A., Mackie, D.M., & Rosselli, F. (1995). Reactions to Anglo- and Hispanic-American-accented speakers: Affect, identity, persuasion, and the English-only controversy. *Language and Communication 15,* 2, 107-120.

Hofstede, G. (1997) *Cultures and Organizations: Software of the Mind*, McGraw-Hill, New York.

Ibister, K. & Nass, C. Consistency of personality in interactive characters: verbal cues, non-verbal cues, and user characteristics. *International Journal of Human-Computer interaction*, in press.

Kroner, D. G. & Weekes, J. R. (1996), Balanced inventory of desirable responding: Factor structure, reliability, and validity with an offender sample. Personality and Individual Differences, 21(3), 323-333.

Ladegaard, H.J. (1998). National stereotypes and language attitudes: The perception of British, American and Australian language and culture in Denmark. *Language and Communication 18*, 251-274.

Lippmann, W. (1922) *Public Opinion*. The Free Press, NY.

Long, M. (1990). Maturational constraints on language development. *Study of Second Language Acquisition 12*, 251-285.

Marcus, A. & Gould, E. W. (2000), Crosscurrents: Cultural Dimensions and Global Web User-Interface Design. *Interactions* VII (4), 32-46.

Moon, Y. (1998) Intimate self-disclosure exchanges: Using computers to build reciprocal relationships with consumers. *Working Paper for Harvard Business School*, 1998

Munro, M.J. & Derwing, T.M. (1995), Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech 38*, 3, 289-306.

Nass, C. & Lee, K.M. (2000). Does Computer-Generated Speech Manifest Personality? An Experimental Test of Similarity-Attraction. *Proceedings of CHI '00*, (Amsterdam, April 2000), ACM Press, 329-336.

Ryan, E. B. (1983). Social psychological mechanisms underlying native speaker evaluations of non-native speaker. *Studies in Second Language Acquisition 5* (2), 148-159.

Ryan, E. B., Giles, H. and Sebastian, R. J. (1982). An integrative perspective for the study of attitudes toward language variation. In Ryan. E. B. and Giles. H. (Eds). *Attitudes Toward Language: Social and Applied Contexts*. Arnold, London.