Published in final edited form as:
Lang Speech. 2010 ; 53(Pt 4): 510–540.

The Wildcat Corpus of Native- and Foreign-Accented English: Communicative Efficiency across Conversational Dyads with Varying Language Alignment Profiles

Kristin J. Van Engen, **Melissa Baese-Berk**, **Rachel E. Baker**, **Arim Choi**, **Midam Kim**, and **Ann R. Bradlow**
Department of Linguistics, Northwestern University, Evanston, Illinois, USA

Abstract

This paper describes the development of the Wildcat Corpus of native- and foreign-accented English, a corpus containing scripted and spontaneous speech recordings from 24 native speakers of American English and 52 non-native speakers of English. The core element of this corpus is a set of spontaneous speech recordings, for which a new method of eliciting dialogue-based, laboratory-quality speech recordings was developed (the Diapix task). Dialogues between two native speakers of English, between two non-native speakers of English (with either shared or different L1s), and between one native and one non-native speaker of English are included and analyzed in terms of general measures of communicative efficiency. The overall finding was that pairs of native talkers were most efficient, followed by mixed native/non-native pairs and non-native pairs with shared L1. Non-native pairs with different L1s were least efficient. These results support the hypothesis that successful speech communication depends both on the alignment of talkers to the target language and on the alignment of talkers to one another in terms of native language background.

Keywords

Diapix task; foreign-accented speech; spontaneous speech; task-oriented dialogue; type-token ratio

1 Background

Many conversations across the globe today take place between interlocutors who do not share a mother tongue, or who, for various socio-political reasons, communicate in a language other than their shared mother tongue. In the case of English, non-native speakers have come to outnumber native speakers, and interactions in English increasingly do not include native speakers at all (Graddol, 1997, 2006; Jenkins, 2000). From the perspective of research on spoken language processing, this globalization of English imposes an expanded level of complexity onto the fundamental issue of the lack of invariance in the mapping of speech signals to their cognitive-linguistic representations. In addition to handling variability that arises from local phonetic context-, individual talker-, and dialect-related sources, listening to foreign-accented English involves handling variability that arises from interactions between the sound structures of English and that of the talker's native language. Furthermore, the spoken English of individual foreign-accented speakers tends to exhibit

Corresponding author: Kristin J. Van Engen, Northwestern University, Department of Linguistics, 2016 Sheridan Road, Evanston, IL 60208-4090, USA, k-van@northwestern.edu.

Reprints and permission: sagepub.co.uk/journalsPermissions.nav

less internal consistency than native-accented English (e.g., see Jongman & Wade, 2007, and Wade, Jongman, & Sereno, 2007, for demonstrations of greater vowel category production variability by non-native compared to native English talkers), and may be relatively unstable over time due to the influence of increasing experience with the sound structure of English (i.e., changing target language proficiency). In an English speech community that includes multi-lingual speakers for whom English is not the first language (L1), an added complication may be the fact that high proficiency talkers from various native language backgrounds may share many target language speech patterns with each other, while continuing to share other target language features with low proficiency talkers with whom they share native language background. Thus, it is possible that an English-based communication network within a multi-lingual setting presents a more complex and dynamic system than most current theories and models of speech perception and production have typically taken into account.

Current theories of speech perception have focused primarily on accounting for variability due to local phonetic context, talker-specific characteristics, and to some extent, sociolinguistic and stylistic factors. For example, in exemplar theories of speech perception (e.g., Goldinger, 1996; Johnson, 1997; Pierrehumbert, 2002), the cognitive representation of speech involves highly detailed encoding of incoming speech signals and multiple levels of categorization/labeling to support instance-contingent speech recognition and production. Fine phonetic details that are indicative of particular talker-, dialect- and/or style-based category labels can be brought to bear on the task of word recognition via the multi-layered, inter-connected, exemplar-based cognitive representation of speech by imposing lexical access biases. To the extent that foreign-accent based categories resemble these other levels of categorization, they can quite easily be incorporated into exemplar models of speech perception (e.g. Goldinger, 1996; Johnson, 1997; Pierrehumbert, 2002). However, as noted above, the high degree of variability associated with foreign-accented speech (e.g., Jongman & Wade, 2007, in combination with its highly dynamic nature, will probably require some elaboration of the mechanisms inherent in current exemplar models of speech perception. Similarly, for models of speech perception that rely on processes of normalization across various sources of variability to access abstractly defined linguistic categories (see Johnson, 2005, for a review of theories of speaker normalization in speech perception), the challenge of foreign-accented speech lies in its inherent variability and instability. These aspects of foreign-accented speech create a type of “moving target” for the development of efficient and effective normalization schemes.

The challenges posed by the globalization of English for theories and models of speech perception and production also offer a potentially groundbreaking opportunity for linking individual-level mechanisms to population-level phenomena, such as contact-induced sound change. By documenting and ultimately delineating the mechanisms underlying individual native English speakers’ adaptations to non-native speech, we stand to gain crucial insight into the changes that the population of English speakers as a whole is undergoing as a result of the contemporary multi-lingual context. With these issues in mind, the goal of the present project was to create an extensive database of native- and foreign-accented English—the Wildcat Corpus of Native and Foreign Accented English.[1] This corpus could then be used to investigate how spoken language processing responds to the particular challenges presented by the globalization of English. The ultimate goal of our long-term research agenda, of which the corpus is an important part, is to articulate a model of speech communication that integrates individual-level speech perception and production mechanisms with population-level, contact-induced sound change. As a necessary first step, we have developed the

[1] Like the developers of the Buckeye Speech Corpus at The Ohio State University, we have used Northwestern University’s mascot—the wildcat—in naming this corpus.

Wildcat Corpus in the hope that it will serve as a resource for establishing a sound, empirical base for consequent theoretical advances.

2 Corpus design

2.1 Alignment between conversational partners

The central hypothesis that underlies the design of the Wildcat Corpus is that successful speech communication in a global context depends on two independent aspects of “talker–listener alignment” (Costa, Pickering, & Sorace, 2008; Pickering & Garrod, 2004, 2006). First, successful speech communication depends on each of the conversation participants’ knowledge of the target language, that is, on the individuals’ levels of target language proficiency. Native speakers of the target language are, of course, well-aligned to the structure of the target language; they all share sufficiently similar long-term memory representations of the target language’s phonetic, phonological, lexical, syntactic, semantic and prosodic structures to facilitate relatively effortless transmission of linguistic propositions across the speech chain. In contrast, in the initial stages of target language acquisition, non-native speakers have varying degrees of misalignment to the target language, depending on the extent of the differences between the native and target language structures. For example, speakers of languages with large and small vowel inventories will be relatively well and poorly matched with English, respectively, in terms of this particular aspect of sound structure.

The second relevant dimension of alignment is the “match” or “mismatch” of the conversation participants in terms of native language background. Two native speakers are well-aligned on this dimension, as are two non-native speakers from the same native language background (e.g., two Korean-accented speakers of English).[2] However, a Korean-accented speaker of English and a Spanish-accented speaker of English are misaligned along this dimension, as are a native English speaker and a Turkish-accented speaker of English. (Note, however, that the degree of misalignment between two non-native talkers from different native language backgrounds will vary depending on the relationship between the sound structures of their native languages and the degree of experience/proficiency of each of the talkers with respect to the target language.)

Based on these two dimensions of alignment, namely alignment of each talker to the target language and alignment of the talkers’ native language backgrounds, we can define several different types of conversations (see Table 1). The Native+Native (N-N) type consists of conversations between two native talkers where each talker is aligned to the target language and the two talkers are aligned with each other; the matched Non-Native+Non-Native (NN1-NN1) type involves two non-native talkers (misaligned to the target language) from the same native language background (aligned on the native language background dimension); conversations between two non-native talkers from different native language backgrounds are of the mismatched Non-Native+Non-Native (NN1-NN2) type (both misaligned to the target language and misaligned along the native language background dimension); the Native+Non-Native (N-NN) type involves a native and a non-native talker, who are aligned and misaligned, respectively, to the target language (therefore in an intermediate row in Table 1), and are misaligned along the native language background dimension. The one unfilled cell in Table 1 (marked with an X) is the cell where both talkers are aligned to the target language but they are misaligned along the native language background dimension. A possible example of this logical type would be two native talkers from different dialect groups. In order to allow for analyses of the effects of these two dimensions of alignment

[2] At least at the very initial stages of target language acquisition before individual differences in target language experience and proficiency have come into play.

between conversational dyads, the spontaneous speech recordings in the Wildcat Corpus involve dialogues between each of the four dyad types shown in Table 1. Because all of the native English speakers in the corpus are speakers of General American English, conversations of the type marked X in Table 1 are not included.

2.2 Scripted and spontaneous speech materials

A second major design principle of the Wildcat Corpus (in addition to the arrangement of conversational dyads as shown in Table 1 and described above) is the inclusion of both spontaneous and scripted speech recordings. Previous work has documented cases of talker adaptation to the listener and listener adaptation to the talker under essentially de-contextualized, monologue conditions (but see Pardo, 2006, for work on phonetic convergence between native-speaking English partners in a dialogue, and Krauss & Pardo, 2004, for additional discussion). For example, work in our own and other laboratories has focused on the production and perception of clear versus conversational speech (see Uchanski, 2005, and Smiljanic & Bradlow, 2009, for reviews of the broad clear speech research agenda). This speaking style variation (clear vs. conversational speech) is a prime example of talker adaptation to the communicative needs of a listener in a compromised auditory setting (e.g., due to a hearing loss or the presence of background noise), and it has been shown to be a fairly effective means of enhancing speech intelligibility for non-native listeners (Bradlow & Bent, 2002; Smiljanic & Bradlow, 2007). However, this work has typically involved presenting pre-recorded, isolated words or sentences to listeners, and therefore provides limited insight into speaking style variations in real-world dialogue situations. One recent study (Ryan, 2007), which compared clear speech in reading and in dialogues, showed that read speech materials may, in fact, underestimate the extent of clear speech features used in natural clear speech.

Similarly, there is a substantial literature showing listeners’ perceptual adaptation to various forms of “deviant” speech, including foreign-accented speech. For example, perceptual adaptation has been demonstrated in response to speech produced by talkers with hearing impairments (McGarr, 1983), to computer synthesized speech (Greenspan, Nusbaum, & Pisoni, 1988; Schwab, Nusbaum, & Pisoni, 1985), to time-compressed speech (e.g., Dupoux & Green, 1997; Pallier, Sebastian-Gallés, Dupoux, Christophe, & Mehler, 1998), and to noise-vocoded speech, i.e. a signal manipulation that simulates the input to an electrical hearing device such as a cochlear implant (Davis, Johnsrude, Hervais-Adelman, Taylor, & McGettigan, 2005). As for the particular case of listener adaptation to foreign-accented speech, previous work in our group and by others has provided strong evidence for adaptation to a specific, single foreign-accented talker (i.e., talker-dependent adaptation, e.g., Bradlow & Bent, 2008; Clarke & Garrett, 2004) and adaptation to an accent as it extends over a group of non-native talkers from the same native language background (i.e., talker-independent adaptation, e.g., Bradlow & Bent, 2008; Weil, 2001). However, as in the cases of demonstrations of talker adaptation discussed above, these studies of listener adaptation to the talker all involved de-contextualized communicative situations in the sense that the listeners were presented with pre-recorded speech in essentially monologue situations. These studies have therefore provided little information about talker–listener adaptation strategies that extend across a dialogue. That is, they provide no insight into how talker- and listener-related adaptation strategies affect each other over the course of a dialogue with multiple turns for each participant. The design of the Wildcat Corpus therefore places a major emphasis on the collection of dialogues, while still providing materials for experiments with sufficiently controlled stimuli to facilitate valid acoustic and perceptual measurements.

For the purposes of this corpus, we developed a new dialogue elicitation procedure, the Diapix task,[3] in which the two conversation partners must work together to find differences between two highly similar pictures. Each participant can see only one picture. This task (described in detail below) incorporated several anticipated benefits (for our purposes) over other dialogue elicitation tasks such as the Map Task (Anderson et al., 1991) in which one participant is typically the “Giver” of instructions while the other is the “Receiver”. First, our task was designed to elicit a wide range of utterance types (questions, answers, declarative descriptions, exclamations, etc.) rather than primarily instructions (imperatives). It was also designed to provide a more even balance of speech production by each of the participants since the two participants have equal roles in the “game” and each has some information the other does not have. Because our aim was to compare different types of interlocutor pairings (as opposed to different types of individual talkers), this symmetrical task was advantageous. Finally, the Diapix task retains a major benefit of the Map Task, namely the ability to include experimenter-determined target items that can be used for relatively controlled subsequent acoustic analysis. Nevertheless, we would like to emphasize that the Map Task remains an excellent option for dialogue studies with other goals, and indeed, the Diapix task was inspired by the Map Task.

The Wildcat Corpus shares some features with existing databases,[4] but its most important distinctive feature is the inclusion of both scripted recordings and task-based dialogues between participants who have been paired in a principled way (N-N, N-NN, NN1-NN1, and NN1-NN2, as shown in Table 1). These pairings and the common conversational task performed by each pair provide a unique tool for investigating the wide range of communicative situations in which speakers of English may be engaged across the globe. By including high-quality recordings of both scripted speech and spontaneous dialogues between the different combinations of native and non-native speakers of English, the Wildcat Corpus aims to fill a gap in the resources available for empirically-based studies of speech communication in a global context. The ultimate goal of this project is to make theoretical advances in our understanding of the linguistic consequences of the globalization of English, particularly with respect to the process of contact-induced sound change.

3 The Wildcat Corpus of native and foreign-accented English[5]

3.1 Talkers

The corpus contains scripted and unscripted (i.e., spontaneous) speech samples from each of 76 talkers: 24 native speakers of English and 52 non-native speakers of English. The native language backgrounds of all participants are summarized in Table 2.

The native English speakers (12 males, 12 females) ranged in age from 18 to 33 years old (average = 20.5); the non-native speakers (32 males, 20 females) ranged from 22 to 34 years (average = 25.8). Native speakers were recruited by word of mouth and through advertisements posted on the Northwestern University campus. Four of the native English speakers reported learning an additional language in the home before age 5 (Krio, Amharic, Creole, Gujarati). In each case, however, the participant was raised in an otherwise English-speaking environment (Nairobi, Kenya; St. Paul, Minnesota; and the Chicago suburbs (2)) and had no detectable foreign accent in English. Most of the non-native speakers were

[3] The name “Diapix” blends two key components of the task: “dialogue” and “pictures”. The basic idea behind the Diapix task is attributed to Valerie Hazan of the Department of Phonetics and Linguistics at University College London. The Diapix recordings in the Wildcat Corpus represent our implementation of this basic idea.

[4] Other databases that include foreign-accented English: The Translanguage English Database Speech Corpus, ICSI Meeting Speech Corpus, the ISL Meeting Speech Part 1, N4 NATO Native and Non-Native Speech Corpus, CSLU: Foreign Accented English Corpus Release 1.2, the Speech Accent Archive (George Mason University), Jilka et al. (2008), the Learning Prosody in a Foreign Language corpus (LeaP, University of Bielefeld), the Voice Interactive Language Training System (VILTS) corpus (Speech Technology and Research Laboratory at SRI International), English as a Lingua Franca in Academic Settings (ELFA, Mauranen, 2003), the Vienna Oxford International Corpus of English (VOICE).

[5] Inquiries regarding the use of Wildcat Corpus materials or recordings may be directed to abradlow@northwestern.edu

recruited from the Northwestern University International Summer Institute (ISI) (*n* = 40), an intensive English language and acculturation program for incoming Ph.D. students at the university. The program takes place during the month before the start of the academic year, so most of the students had been at Northwestern for a month or less. Twelve of the native Korean speakers were recruited from the broader Northwestern community by word of mouth or posters. These participants had been in the U.S. for no more than 3 years. The Northwestern University Graduate School requires TOEFL scores of at least 600 for the paper-based test, 250 for the computer-based test, and 100 for the internet-based test, so all graduate student participants had achieved standardized English test scores at these levels or higher.[6] All speakers received payment for their participation.

3.2 Materials and procedures

Each participant participated with one other talker in the Diapix task (an interactive, goal-oriented task described in detail below) and then read a set of scripted English materials. All recordings took place during a single session, which lasted approximately one hour.[7]

3.2.1 Scripted materials—The purpose of including scripted materials in the database was twofold: (1) to allow for well-controlled acoustic analyses and (2) to provide standardized stimuli for perception experiments. Accordingly, the set of scripted materials includes utterances of various lengths, ranging from isolated words to sentences and paragraphs. For each of these material types, items commonly found in comparable speech databases from other laboratories have been included to facilitate comparison across databases. For example, the “Stella” passage from the Speech Accent Archive (Weinberger, n.d.) and the “North Wind and The Sun” (as in the IPA Handbook; IPA, 1999) were included as paragraph-length utterances. In addition, items developed at Northwestern for use with non-native English speakers (e.g., high and low predictability sentences from Bradlow & Alexander, 2007, a second-mention reduction passage from Baker & Bradlow, 2009) were included. Further information about the scripted portion of the corpus will be made available online and discussed in subsequent reports.

3.2.2 Unscripted materials: the Diapix task—Spontaneous speech, produced in the context of dialogues between two cooperating speakers, was elicited from the participants using the novel Diapix task. The Diapix task is a spot-the-difference game involving a pair of pictures and a pair of participants. The two pictures within a pair represent the same general scene, but there are 10 differences between the two pictures. Of these, 6 items are present in one picture but absent in the other (3 “missing” items from each picture), and 4 items are slightly different in the two pictures (“change items”, e.g., in terms of color or some other detail). (See images in Appendix.) Each participant is given one version of the scene. Participants wear head-mounted microphones and are seated back-to-back in a large recording booth such that they can easily hear each other’s speech but cannot see the other person’s picture. They are instructed to work together (not in competition) to find the differences between their two pictures. They are also instructed to indicate the identification of a difference by marking (with a circle, sketch, or note) the relevant portion of the picture.

[6] Standardized test score minima are unavailable for four participants (1 Chinese, 3 Korean), as they were partners of graduate students.

[7] In addition to English recordings, all Korean participants also read scripted materials in Korean. Their recording sessions, therefore, took approximately 1.5 hours. Eight additional Korean participants (not included in the present total), also read the scripted materials in both languages and performed the Diapix task in Korean (= 4 Korean Diapix conversations). The purpose of collecting these materials was to develop a parallel (though smaller in scale) corpus that could be used as a basis for comparing speech patterns within individuals in their native and non-native languages. This comparison is important for disentangling the effects of native language transfer from phenomena that are specific to second-language production in general or to non-linguistic, cultural factors. We leave discussion of the Korean recordings for a subsequent report.

Each of the 76 speakers in the corpus participated in one Diapix recording session. This limit was imposed to control for task familiarity across all pairs. Participants were paired so that we could investigate conversations in which the partners were both native speakers of the target language (N-N: 8 pairs); both non-native speakers of the target language (English), but sharing native language background (NN1-NN1: 11 pairs); both non-native speakers of the target language (English), with different native language backgrounds (NN1-NN2: 11 pairs); and mixed native and non-native English-speaking pairs (N-NN: 8 pairs). These pairings are summarized in Table 3, with participants listed by their native languages.

Each Diapix recording session began with a familiarization task, in which each participant was given a pair of pictures that contained 8 differences. The familiarization images were taken from a published, photographic version of a similar find-the-difference task in a popular, local magazine. The participants were given approximately 3 minutes to identify as many differences as possible. It was then explained that, for the speech recording, they would be working with their partner on a similar task in which each person has just one of the two pictures. Participants were instructed that there were 10 differences in the experimental task, and that they should try to identify them as quickly and accurately as possible. They were reminded that the task was cooperative and that they must use only English in their conversations.

Participants were recorded in a large sound-treated booth in the Northwestern University Phonetics Laboratory. They wore AKG C420 headset microphones, and their conversations were recorded in stereo using a Marantz PMD 670 flash recorder. If the participants had not completed the task within 20 minutes, the experimenter(s) ended the session. We determined that participants who had reached the 20-minute point were generally becoming quite frustrated with the task and were having difficulty continuing the conversation. In a few cases (4), participants were allowed to continue their conversations slightly beyond 20 minutes in order to avoid an awkward interruption by the experimenter, but all analyses presented here will include only the first 20 minutes of any Diapix session that was not completed in less than 20 minutes.

The Diapix conversations were transcribed orthographically by trained research assistants. These transcriptions were automatically aligned to the sound files, and the aligned files were hand-corrected to ensure that the boundaries between speech and non-speech or silence were accurate. Detailed information about the conventions and software employed for these purposes will be made available online.

3.2.3 Accent ratings—As a means of assessing the relative proficiencies of the set of non-native talkers, accent ratings were obtained for each of the non-native talkers in the corpus. We note here that this accent rating test was necessarily conducted after all recordings (scripted and Diapix) had been collected, and could therefore only be used for post-hoc analyses of the role of proficiency (an issue that we take up in the discussion section), and not as a means of selecting particular dyad pairings for the Diapix task. Fifty native speakers of American English (undergraduate students at Northwestern University) listened to the recordings of each non-native talker reading a scripted paragraph (the “Stella” passage). The listeners rated each speaker on a scale of 1 (no foreign accent) to 9 (very strong foreign accent). Recordings from a subset of the native talkers ($N = 13$) were included in this test to provide the full range of accentedness to the listeners. The average rating for native speakers of English was 1.27 (range: 1.04 to 1.67) and for non-native speakers was 6.35 (range: 3.10 to 8.31).

4 Corpus analysis: communicative efficiency

The specific question we sought to address with these analyses was: How is speech communication efficiency affected by (a) the interlocutors' alignment to the language of communication (native vs. non-native) and (b) the interlocutors' alignment to each another (shared/matched vs. unshared/mismatched native language)?

Following Marshall, Freed, and Phillips (1997) in their work on severe aphasics, we understand communicative efficiency to refer generally to the completeness, clarity, and speed with which information is exchanged in a conversation. In order to measure and compare the efficiency of the Diapix conversations, we cast this general conception of communicative efficiency in terms of two primary measures: (1) time to complete the task, and (2) word type-to-token ratio.

The task was designed with the intent that all participants would be able to identify all (or at least most) of the differences, thus controlling the relevant information exchanged in all conversations. Participants were also instructed to complete the task as quickly and accurately as possible. Given these elements of control, task completion time can be used as a simple, gross measure of communicative efficiency: relatively short and long task durations indicate relatively high and low communicative efficiency, respectively, in the Diapix task.

Word type-to-token ratio provides additional, duration-independent, information about communicative efficiency by offering a window into the participants' active vocabularies during the Diapix task. As for task completion time, the validity of type-to-token ratio as an efficiency measure is reliant on the constraints inherent to the Diapix paradigm. Here, the number of types (different words) is constrained by the limited number of items on the page, the nature of the task, and the relative simplicity of the scene. Because number of types and conversation time are constrained, type-to-token ratio measures efficiency within the Diapix task: a relatively large ratio indicates few repetitions of a wide range of unique words, that is, efficient use of an effective vocabulary set.

It is crucial that these measures be interpreted strictly within the context of the Diapix task. In real-life conversations, long durations and high amounts of word repetition may, in some cases, be indicative of efficiency (rather than inefficiency) in communication. For example, it may be the case that people who are able to carry on longer, more complex conversations with one another are also more efficient in their communication. Or, with respect to type-to-token ratio, interlocutors who are well-aligned to one another may use the same words to refer to items and concepts, which would lower their type-to-token ratio.

In keeping with our central hypothesis, we predicted that the interlocutors' alignment with the target language and with each other would contribute to communicative efficiency (as laid out in Table 1). We therefore predicted that N-N pairs would have the fastest task durations and highest word type-to-token ratios, that NN1-NN2 pairs would have the slowest task durations and lowest word type-to-token ratios, and that the N-NN and NN1-NN1 pairs would have intermediate task durations and word type-to-token ratios. We were unsure whether the N-NN and NN1-NN1 pairs would differ in terms of task completion times and word type-to-token ratios although, anecdotally, there seems to be a general belief that the presence of a native speaker is not particularly helpful, leading us to predict that the NN1-NN1 pair type would show greater efficiency (faster task completion times and higher word type-to-token ratios) than the N-NN pair type (e.g., see suggestions by Costa et al., 2008). For example, consider this quote from a recent article in the *Financial Times* ("One language fits all", by Henry Hitchings, May 3/May 4, 2008):

> As English increasingly becomes the language of business, native speakers feel, quite understandably, that they are at an advantage. But, discussion often goes more smoothly when the native speakers leave the room…. The people who see themselves as facilitators are, in reality, obstacles. (p.17)

5 Results

5.1 Task success

In order to make meaningful comparisons of communicative efficiency across participant pair types (NN, N-NN, NN1-NN1, NN1-NN2), it was necessary to control both the goal of the conversational task (identify the same 10 differences) and participants' ability to achieve that goal. To that end, one of the aims of the Diapix task design was to create a dialogue-based task in which all participants would be able to achieve a high level of success. Data from the corpus indicate that, indeed, all pairs understood the task and all pair types were quite successful in identifying the differences in the Diapix scenes; the median score was 10 (100%) for all pair types.[8] A Kruskal-Wallis one-way analysis of variance by ranks shows no significant differences between pair types (Kruskal-Wallis chi-squared = 2.83, *df* = 3, p-value = 0.420). This non-parametric test was used because the accuracy data are not normally distributed and because of the relatively small number of data points: 8 N-N pairs, 8 N-NN pairs, 11 NN1-NN1 pairs, 11 NN1-NN2 pairs. Task accuracy scores of less than 10 occurred where participants simply did not identify a difference (or did not do it within 20 minutes); where they identified as different something that was not (e.g. use of different color labels for what was actually the same color); or where they interpreted a single difference as two separate differences. For example, one of the two images depicts a beehive with bees flying around it, whereas the other image has no hive or bees. This was intended as a single difference (hive/no hive), but a small number of pairs counted the hive and the bees as two separate differences. Such a decision may lead a pair to miss another difference. Overall, we felt confident that all pair types were able to complete the Diapix task with a high degree of success.

5.2 Task completion time

The amount of time that participants took to complete the Diapix task ranged from 5.34 minutes to 20 minutes (limit imposed by the experimenters as discussed above). Data by pair type are presented in Figure 1.

A Kruskal-Wallis one-way analysis of variance by ranks showed a significant effect of group (Kruskal-Wallis chi-squared = 12.44, *df* = 3, p-value = .00601), and Wilcoxon rank sums tests (unpaired) show that N-N pairs were significantly faster at the Diapix task than N-NN pairs (*W* = 11, p-value = .0281), NN1-NN1 pairs (*W* = 8, p-value = .00177), and NN1-NN2 pairs (*W* = 7, p-value = .00257). No other comparisons between pair types were significant.

Despite the lack of statistically significant differences among the other groups' means, several interesting patterns emerge in this data. First, the variance was much greater for the three groups that involved NN talkers compared with the N-N group: at least one pair within each of these groups performed the task at or below the median N-N task duration, and at least one pair in each of these groups had to be cut off at 20 minutes. Within this wide range of durations, the medians were similar for the N-NN group and the NN1-NN1 group (12.40 minutes and 11.21 minutes respectively), but the median for the NN1-NN2 group was much higher (18.77 minutes). This median is greater than 75% of the NN1-NN1 pairs' task

[8] Means and standard deviations: N-N: 9.88 (.35); N-NN: 9.38 (1.06); NN1-NN1: 9.45 (.69); NN1-NN2: 9.09 (1.51).

durations, suggesting that, without the 20-minute limit, this group would likely have taken significantly longer than the others. (Summary statistics are shown in Table 4.)

In order to investigate whether the differences in task duration were primarily due to time spent in silence versus actual speaking time, analysis of total speech duration by pair type was also examined. Speech duration was calculated for each talker by excluding silences greater than or equal to 500 ms (determined heuristically based on initial auditory and visual inspections of the dialogue recordings), as well as non-speech sounds such as laughter, breaths, sighs, and other noises. The total speech duration for a given conversation, then, is the sum of the two interlocutors' individual durations. As shown in Figure 2, the pattern across the four pair types is similar for total task time and for total speech duration.

A Kruskal-Wallis one-way analysis of variance by ranks showed a significant effect of group (Kruskal-Wallis chi-squared = 12.9234, *df* = 3, p-value = .00481), and Wilcoxon rank sums tests (unpaired) show that N-N pairs spent significantly less time speaking than NN1-NN1 pairs (*W* = 5, p-value < .001) and NN1-NN2 pairs (*W* = 8, p-value = .00177). No other comparisons between pair types were statistically significant.

5.3 Balance of speech

In order to verify whether the Diapix task generally elicits a relatively equal amount of speech from two interlocutors, the balance of interlocutors' speech was also analyzed. Such analysis also allows us to determine whether different pair types exhibited different patterns with respect to balance of speech. This is particularly relevant, for example, in the case of N-NN partners, where one might expect that the native speaker of the target language would speak more than the non-native, perhaps to hurry the task along or to provide clarification to the non-native partner.

The balance of speech across the two talkers within conversations was measured by calculating the ratio of the two partners' total speech durations within a conversation. The partner with the shorter overall speech duration was arbitrarily entered as the numerator, so that a ratio of 1.0 represents perfect balance *and* is the highest possible ratio value. Boxplots showing the ratios by pair type are presented in Figure 3.

As expected based on the structure of the Diapix task, partners in most conditions were relatively balanced in terms of speech duration, and a Kruskal-Wallis analysis showed no significant differences between the groups. Balance was also measured in terms of the number of speech intervals per person (intervals of speech bounded by silences greater than or equal to 500 ms), and the balance ratios were even greater (i.e., closer to 1). However, as shown in Figure 3, all pair types did have at least one pair with a duration "balance ratio" of less than .5, meaning one partner contributed twice as much speech (measured by duration) as compared to the other.

With respect to particular pair types, we found that N-NN pairs were not less balanced than the other groups. As shown in Figure 4, the N partner spoke less than the NN partner in 4 out of 8 conversations.

Another notable observation regarding balance of speech by pair type is that the N-N pairs were (numerically) the least balanced of all groups (see Figure 3), perhaps suggesting that an efficient strategy (i.e., one that leads to rapid completion of the task) involves the spontaneous adoption of a leadership role by one of the pair members.

5.4 Word type-to-token ratios

Even though the Diapix instructions told participants to complete the task as quickly and accurately as possible, many pairs did not seem to hurry, and several pairs continued discussion after they had identified all 10 differences, simply because they were not keeping careful track of the number they had identified. In general, considerations of rapport, as well as general anxiety about speaking English in a laboratory setting or performing a strange task, seemed to outweigh the instruction to work quickly. We therefore deemed it especially important to measure communicative efficiency in a non-time-related manner, namely with word type-to-token ratio.

This ratio (the number of unique words to the number of total words spoken) serves as an indicator of the efficiency of effective vocabulary use. It was calculated for each individual who participated in the Diapix task and for each Diapix conversation (over the entire pool of words spoken by both participants). For the purpose of the present analysis, types were defined strictly based on orthographic strings, such that singular and plural forms of a given noun were counted as separate types, as were different tenses of a given verb, provided they are spelled differently (i.e., different verb tenses with identical spelling would count as instances of a single type). Type-to-token ratios for individuals and conversations were compared across pair types.

It should be noted first that the number of types in the conversations was similar across all pair types, as shown in Figure 5 below. As seen above (Figure 1), the durations of these conversations did differ significantly, at least insofar as the N-N conversations were significantly shorter than all of the others. The similar number of types across conversations indicates that, over a wide range of conversation durations, similar numbers of unique words were used. This also provides further evidence that all pair types had the opportunity to discuss all of the items in the picture.

The type-to-token ratios, however, did show different patterns across pair types, as illustrated in Figures 6 and 7 below. Note first that the higher overall type-to-token ratios for individuals as compared to pairs reflect the repetition of words by the two talkers in a conversation. If each partner uses the word “bench” once, they have an unrepeated type in each of their individual type-to-token ratio calculations, but the conversation contains a repetition.

With respect to type-to-token ratio calculated over conversations (Figure 6), a Kruskal-Wallis one-way analysis of variance by ranks showed a significant effect of group, Kruskal-Wallis chi-squared = 19.6043, $df = 3$, p-value < .001, and Wilcoxon rank sums tests (unpaired) show that N-N pairs differ significantly from N-NN pairs, $W = 56$, p-value = .0104, NN1-NN1 pairs, $W = 88$, p-value < .001, and NN1-NN2 pairs, $W = 88$, p-value < .001. In addition, N-NN pairs differed from NN1-NN2 pairs, $W = 70$, p-value = .0328.

The effect of group was also significant when the type-to-token ratios were calculated for individual talkers, Kruskal-Wallis chi-squared = 34.0358, $df = 4$, p-value < .001. Wilcoxon rank sum tests on these data showed that talkers in the N-N condition differed from those in the N-NN condition, $W = 219$, p-value < .001, the NN1-NN1 condition, $W = 343$, p-value < .001, and the NN1-NN2 condition, $W = 340$, p-value < .001. In addition, talkers in the N-NN condition again differed from those in the NN1-NN2 condition, $W = 256$, p-value = .0174.

Wilcoxon tests were also performed to compare the native talkers in N-N pairs to those in N-NN pairs. The number of types produced by these groups of speakers did not differ significantly, but interestingly, the native English speakers in these two conditions differed significantly in terms of type-to-token ratio, $W = 113$, p-value = .00166,[9] such that native English speakers paired with other native speakers had higher ratios than those who were paired with non-native talkers. This difference shows that native speakers employ a greater amount of repetition in their interactions with non-native speakers. A reduced vocabulary size could also be involved in lower type-to-token ratios, but the lack of difference in the number of types (see Figure 4) shows that this difference is driven primarily by higher repetition. This greater repetition may be produced by native speakers in an effort to be especially clear for non-native partners, to respond to non-natives' questions or confusion, or both.

Non-natives in the N-NN condition were also compared to non-natives in the two other conditions, and were found to have significantly higher type-to-token ratios than the non-natives in the NN1-NN2 condition, $W = 134$, p-value = .0308, though there was no significant difference between the ratios of non-native talkers with native partners and non-native talkers with matched non-native partners, $W = 120$, p-value = 0.142.

5.5 Accentedness ratings

The accentedness ratings (averaged over the 50 listeners) for each of the non-native speakers were analyzed to determine whether differences in accentedness (an aspect of speaking proficiency) might account for differences between pair type groups in the Diapix task. Mann-Whitney tests showed no significant differences between the groups of NN speakers divided by Diapix pair type (means: N-NN = 6.00; NN1-NN1 = 6.51; NN1-NN2 = 6.33).

To further investigate the role of non-native accent in the Diapix data, Spearman rank correlations were used to test for correlations between accent ratings and our two primary dependent measures: task duration and type-to-token ratio. Correlations for pairs were tested using the rating of the more accented partner, the rating of the less accented partner, and the average accent rating of the two partners. No accent measure was significantly correlated with task duration within any pair type or across the groups that included non-native talkers. Furthermore, no accent measure was significantly correlated with a pair type-to-token ratio (or the number of types or tokens) within any pair type. The average accentedness of pairs did significantly correlate with pair type-to-token ratio, $p = .013$, rho = –.451, across the groups such that pairs with a lower average accent rating had higher type-to-token ratios. This pattern was unsurprising given that pair averages from the N-NN group were necessarily lowered by the native partners' low accent ratings. Indeed, when the N-NN pairs were omitted from the analysis, the correlation was not significant. Similarly, the accent ratings of the less accented partners also correlated with task duration across the groups, $p = .028$, rho =−.402, but the ratings of the more accented partners did not. (The less accented partners were always the native speakers in the N-NN condition.) Correlations between individual speakers' accent ratings and individual type-to-token ratios were also tested for the full set of non-native talkers, showing no significant correlations.

5.6 Qualitative analysis of Diapix conversations

In addition to the quantitative analyses presented above, a single experimenter (KVE) listened to all of the Diapix conversations in order to ascertain their task strategies. It was predicted that, in the absence of instructions with regard to strategy, participants might adopt one of a small number of strategy types. For example, we noticed in pilot sessions that pairs tended to approach the task either with a question-and-answer strategy (i.e., participants interviewing one another about the picture's contents) or with a describe-and-respond

[9] The native talkers in the N-N condition also differed significantly from the NNs in the N-NN condition ($W = 106$, p-value = 0.00875).

strategy (one participant describing his/her picture in an orderly way, the other interjecting where differences are detected). It was also predicted that the clear emergence of a leader might be a characteristic of some Diapix conversations, and that this may vary across the different pair types.

The qualitative assessment of all of the Diapix conversations in the corpus, however, did not lend itself to these simple categorizations. Pairs generally used a combination of strategies in the task, and while a single leader could be identified in a few conversations, it was most common for leadership behavior to move between partners during the course of a conversation. This observation highlights a design feature of the Diapix task, since one goal of the task was to equalize the roles of the participants.

Two other characteristics of the conversations did emerge as observable qualitative dimensions. First, it was noted that N-N pairs often began the task by explicitly discussing some aspect of their strategy (e.g., “Alright, how about we go from left to right and describe what’s going on in the scene?”; “Uh, how do you want to do this? Left to right?”). Each pair in the corpus was classified, therefore, as “strategic” or “non-strategic”, depending on whether they began the task with such an utterance. The second qualitative dimension of categorization was whether or not a pair was spatially systematic in the order of the items they discussed. A pair was considered to be spatially systematic if their conversations generally proceeded from left to right, right to left, clockwise, or counterclockwise with respect to the items in the scene.[10] The proportion of pairs displaying these two characteristics is shown by pair type in Figure 8.

The most striking result of this analysis was that N-N pairs were highly likely to begin the task by discussing strategy, whereas all other pair types were quite unlikely to do so. In terms of spatial systematicity, the differences are less marked, though N-N pairs tended to be most likely to proceed systematically, followed by the N-NN pairs, and then the NN-NN pairs of both types.

It is not entirely clear why N-N pairs, and no other type, showed such a strong tendency to discuss strategy at the outset of the Diapix task. This may be a characteristic of N-N conversations; that is, native speakers of the target language may focus more on task strategy since they are able to speak the target language with ease. This pattern may also be due to social or cultural factors. For example, Americans may be more familiar with tasks of this type and/or more comfortable giving instructions to unfamiliar interlocutors. Future analyses of N-N pairs in other target languages will allow us to begin to disentangle native speaker effects from social/cultural effects.

The process of qualitative analysis also allowed us to begin to understand why particular conversations may have been quite long or short. In several of the very long conversations, pairs identified all but one difference and could not locate the last one. They ended up returning to items they had already discussed or dwelling in great detail on irrelevant aspects of the scene. In general, the ability to hone in on the appropriate level of detail for completing the task appeared to have a great impact on task duration—many of the pairs with long durations were unnecessarily detailed in their discussions of items in the scene.

[10] The reliability of these classifications was checked independently by a second rater. The raters were 100% reliable on judgments of whether pairs were strategic, but there were some discrepancies (10/38 conversations) in their judgments of spatial systematicity. These discrepancies generally arose where pairs displayed a degree of spatial systematicity for part, but not all, of the conversation. In these cases, the two raters reviewed the conversations and settled on a consensus judgment. When a pair generally discussed items in a spatially systematic order but missed some differences such that they had to jump around the page to identify them, a judgment of “systematic” was given. If the differences missed were so many that the conversation was predominantly not spatially systematic, or if systematicity could not be observed until well after the conversation had begun, a judgment of “non-systematic” was given.

6 Summary and discussion

The analyses of communicative efficiency in Diapix tasks across conversational dyads with varying degrees of linguistic alignment are generally consistent with the hypothesis that successful speech communication in a global context depends on (a) alignment of the talkers to the target language, and (b) alignment of the talkers to each other in terms of native language background. Specifically, the Diapix data showed:

1. Talker pairs involving two native English speakers (well-aligned to the target language, and well-aligned native language backgrounds) were most efficient at the English Diapix task. That is, the N-N pairs performed the task in the shortest amount of time and had the greatest word type-to-token ratios.

2. Talker pairs involving one native and one non-native English speaker (mixed alignment to the target language, and misaligned/mismatched native language backgrounds) were no more efficient, in terms of task completion time and word type-to-token ratio, than pairs involving two non-native speakers from the same native language background (misaligned/mismatched to the target language, but well aligned/matched native language backgrounds).

3. Talker pairs involving two non-native speakers from different native language backgrounds (both misaligned/mismatched to the target language, and misaligned/mismatched native language backgrounds) tended toward the low-efficiency end of the scale for both task duration and word type-to-token ratio. However, with the relatively small sample size within each pair type, this numerical pattern was not statistically significant.

4. Talker pairs involving two native speakers discussed task strategy far more often than other pair types, and were also most likely to discuss the items in the scene in a spatially systematic manner. N-NN pairs were slightly more likely than the NN-NN pairs to proceed systematically.

This general pattern of variation in communicative efficiency according to the two dimensions of talker–listener alignment—alignment of each talker to the target language and alignment of the two talkers' native language backgrounds to each other—extends previous work on talker–listener alignment to situations involving non-native speakers in more ecologically valid communicative settings. In particular, these data from the Diapix task, a dialogue-based spontaneous speech task, are consistent with previous findings of an intelligibility benefit of foreign-accented speech for non-native listeners relative to native listeners (the so called "interlanguage speech intelligibility benefit" reported in Bent & Bradlow, 2003; Bent, Bradlow, & Smith, 2008; Hayes-Harb, Smith, Bent, & Bradlow, 2008; Imai, Walley, & Flege, 2005; Major, Fitzmaurice, Bunta, & Balasubramanian, 2002; Munro, Derwing, & Morton, 2006; Smith & Rafiqzad, 1979; Stibbard & Lee, 2006; van Wijngaarden, 2001; van Wijngaarden, Steeneken, & Houtgast, 2002). While the previous studies examined intelligibility (in terms of word recognition accuracy), the present study examined overall communicative efficiency (in terms of task completion time and word type-to-token ratio). These two measurement parameters are not identical (intelligibility versus communicative efficiency), but the pattern we observe in the present study suggests that they may be related. In particular, the fact that the pairs in which both speakers are non-natives (NN-NN pairs) appear to be just as efficient on the Diapix task as pairs that include one native speaker (N-NN pairs), suggests that there may be no disadvantage in communicative efficiency for NN-NN pairs relative to N-NN pairs in much the same way as there appears to be no disadvantage in intelligibility for non-native listeners (relative to native listeners) when presented with non-native speech (as demonstrated in the above-referenced studies).

An important finding of the above-referenced studies of intelligibility of native and non-native speech for native and non-native listeners is that the “interlanguage speech intelligibility benefit” is mediated by the non-native talker’s proficiency in the target language. (For extensive discussion of this issue see Stibbard & Lee, 2006, Hayes-Harb et al., 2008, and references cited in these papers.) In the design of the Wildcat Corpus, proficiency was effectively reduced to a binary distinction along the target language alignment dimension: native and non-native talkers were coded as matched/aligned and mis-matched/mis-aligned along this dimension, respectively. The English proficiency of the non-native speakers in this corpus was generally quite high as measured by standard tests such as the TOEFL. However, within this population, there is still a considerable amount of variability in terms of speaking proficiency. Controlling for this factor in the present corpus proved to be extremely difficult from a logistical point of view. Furthermore, because only a small number of non-native participants reported their TOEFL or other proficiency scores, we were unable to enter speaking proficiency score into the analyses.[11]

We were able to begin to address this issue within the present data set by investigating whether differences in accentedness (an aspect of proficiency) across the non-native speakers in the different groups might account for differences that were observed. Given that there were no significant differences between the groups in terms of accentedness, it appears, at least preliminarily, that the communicative efficiency differences that were observed across pair types were not driven by differences in the proficiencies of the speakers in these groups. However, a more fine-grained approach to investigating the effects of alignment to the target language would involve careful proficiency-matching and/or grouping. For example, NN1-NN1 pairs comprised of high proficiency talkers could be compared to NN1-NN1 pairs comprised of low proficiency talkers, or NN1-NN1 pairs of mixed proficiency could be compared to NN1-NN2 pairs of mixed proficiency. Such comparisons would enhance our current data by further delineating the role of alignment to the target language.

A second important feature of the Diapix portion of the Wildcat Corpus was the arrangement of conversational dyads along the native language background dimension. As with the target language alignment dimension, this dimension of variation across talker pairs was expressed as a binary distinction: talkers within a pair were coded as either matched/aligned (same native language background) or mis-matched/mis-aligned (different native language background) along this dimension. However, in reality, this dimension is also one with grades of variation: languages can be typologically more or less similar to each other rather than categorically the same or different. And, we have every reason to expect that these grades of language similarity will have a significant impact on the degree of mutual intelligibility between talkers from different native language backgrounds. For example, we can expect that two talkers from two different native language backgrounds that both have processes of final consonant devoicing will be “closer” along the native language background dimension than two talkers from two different native language backgrounds that differ in this regard. In order to make principled predictions about the mutual intelligibility across varieties of foreign-accented English, and therefore to be able to express grades of alignment along the native language background dimension, we need to devise a means of representing languages in a sound similarity “space.” This language sound similarity space should take into account a wide range of features of linguistic sound structure, including features of the phoneme inventories, phonotactics, and prosody, and should provide a means of assessing the sound structure distance between English, the target language, and each relevant source language, as well as the distance between each of the source languages. In a

[11] Because participants may be reluctant to share standardized test scores, we plan to develop a separate measure of proficiency to be administered in the lab in the future.

separate line of work we are pursuing this issue (for a preliminary report, see Bradlow, Clopper, & Smiljanic, 2007), with an eye to incorporating it into the selection of future Diapix dyads.

The development of the Diapix task was a major methodological goal of the present study. The systematic communicative efficiency patterns discussed above indicate that the general Diapix task approach holds significant promise as an effective method for eliciting spontaneous conversational recordings. The conversations described here also indicate that the task succeeded in eliciting a wide range of utterance types from participants and a relatively balanced amount of speech from the two participants. These characteristics of the Diapix task also make it a useful methodological tool for addressing a broad range of research questions about speech communication in which it is preferable for the participants to have equal roles in the interaction.

Nevertheless, in the course of our analyses of the Diapix recordings in the current Wildcat Corpus we noticed several aspects of our implementation of the basic Diapix idea that require improvement. Most importantly, in creating the Diapix scenes we included various target items with the intention that these items could be used for subsequent acoustic analysis. For example, we included items that were intended to elicit the point vowels in a relatively consistent local phonetic context: “boss” for /a/, “booze” for /u/, and “bees” for /i/. However, it soon became evident that the participants did not produce anywhere near enough repetitions of these target items to facilitate valid acoustic measurements of the target vowels. We also ran into some unanticipated problems with the various details of the Diapix scenes, including no provision for color-blind participants (which created trouble for color based differences such as the color of the woman’s shoes) and some differences across the scene versions being interpreted as multiple rather than a single difference for some participants (e.g., the “beehive” difference involved the presence vs. absence of both a beehive and several bees around the beehive).

As stated in the introduction to this paper, the goal of the present project was to create an extensive database of native- and foreign-accented English that could be used as an empirical base for understanding the theoretical and applied implications of English spoken language communication in a global context. While we consider this general project to still be at a relatively early stage of development, we are encouraged by the success of our methodological innovations (most notably, the development of the Diapix task) and by the initial support that we have found for our central hypothesis, namely that successful speech communication in a global context depends on (a) alignment of the talkers to the target language, and (b) alignment of the talkers to each other in terms of native language background. Consistent with current exemplar theories of speech perception and production (e.g., Goldinger, 1996; Johnson, 1997; Pierrehumbert, 2002), the dyad-dependent variation in communicative efficiency observed in this study suggests a tight correspondence between the details of the current speech input and the cognitive representations that underlie access to higher levels of linguistic structure for speech recognition and target selection for speech production. Moreover, these data are in-line with recent work on perceptual learning for speech which has demonstrated remarkable flexibility and adaptation in response to situation-specific variation (e.g., Norris, McQueen, & Cutler, 2003; Eisner & McQueen, 2005, 2006; Kraljic & Samuel, 2005; Kraljic & Samuel, 2006; Kraljic & Samuel, 2007; Bradlow & Bent, 2008; Clarke and Garrett, 2004; Maye, Aslin, & Tanenhaus, 2008). Our data represent an attempt to demonstrate the operation of these input-sensitive adaptive mechanisms in the context of relatively natural and spontaneous dialogues. An important future direction is to follow up these broad, communicative efficiency measures (time to complete the task and type-to-token ratio) with more fine-grained measures of production and perception adaptation within and between our various dyad types.

Acknowledgments

The authors would like to thank Chun Liang Chan and Janet Pierrehumbert for their contributions to the development of the speech database "toolbox." We also thank Jill Harris, Kelsey Mok, and Page Piccinini for their assistance in transcription and alignment. Finally, we gratefully acknowledge the staff and participants of the Northwestern University International Summer Institute of 2006 for their cooperation with this project. This work was supported by grant R01-DC005794 from NIH-NIDCD.

References

ANDERSON AH, BADER M, BARD EG, BOYLE EH, DOHERTY GM, GARROD SC, et al. The HCRC Map Task Corpus. Language and Speech. 1991; 34(4):351–366.

BAKER RE, BRADLOW AR. Variability in word duration as a function of probability, speech style, and prosody. Language and Speech. 2009; 52(4):391–413. [PubMed: 20121039]

BENT T, BRADLOW AR. The interlanguage speech intelligibility benefit. Journal of the Acoustical Society of America. 2003; 114(3):1600–1610. [PubMed: 14514213]

BENT T, BRADLOW AR, SMITH BL. Production and perception of temporal contrasts in native and non-native speech. Phonetica. 2008; 65:131–147. [PubMed: 18679042]

BRADLOW AR, ALEXANDER JA. Semantic-contextual and acoustic-phonetic enhancements for English sentence-in-noise recognition by native and non-native listeners. Journal of the Acoustical Society of America. 2007; 121(4):2339–2349. [PubMed: 17471746]

BRADLOW AR, BENT T. The clear speech effect for non-native listeners. Journal of the Acoustical Society of America. 2002; 112(1):272–284. [PubMed: 12141353]

BRADLOW AR, BENT T. Perceptual adaptation to non-native speech. Cognition. 2008; 106:707–729. [PubMed: 17532315]

BRADLOW, A.; CLOPPER, C.; SMILJANIC, R. A perceptual similarity space for languages. Proceedings of the XVIth International Congress of Phonetic Sciences; Saarbrucken, Germany. 2007.

CIEFL. Monograph. Hyderabad: CIEFL; 1972. The Sound System of Indian English; p. 7

CLARKE CM, GARRETT MF. Rapid adaptation to foreign-accented English. Journal of the Acoustical Society of America. 2004; 116(6):3647–3658. [PubMed: 15658715]

COSTA A, PICKERING MJ, SORACE A. Alignment in second language dialogue. Language and Cognitive Processes. 2008; 23:528–556.

DAVIS MH, JOHNSRUDE IS, HERVAIS-ADELMAN A, TAYLOR K, MCGETTIGAN C. Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences. Journal of Experimental Psychology: General. 2005; 134(2):222–241. [PubMed: 15869347]

DUPOUX E, GREEN K. Perceptual adjustment to highly compressed speech: Effects of talker and rate changes. Journal of Experimental Psychology: Human Perception and Performance. 1997; 23:914–927. [PubMed: 9180050]

EISNER F, McQUEEN JM. The specificity of perceptual learning in speech processing. Perception and Psychophysics. 2005; 67(2):224–238. [PubMed: 15971687]

EISNER F, McQUEEN JM. Perceptual learning in speech: Stability over time. Journal of the Acoustical Society of America. 2006; 119(4):1950–1953. [PubMed: 16642808]

GOLDINGER SD. Words and voices: Episodic traces in spoken word identification and recognition memory. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1996; 22:1166–1183.

GRADDOL, D. The future of English?. London: The British Council; 1997.

GRADDOL, D. English next. London: The British Council; 2006.

GREENSPAN SL, NUSBAUM HC, PISONI DB. Perceptual learning of synthetic speech produced by rule. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1988; 14(3):421–433.

HAYES-HARB R, SMITH BL, BENT T, BRADLOW AR. The interlanguage speech intelligibility benefit for native speakers of Mandarin: Production and perception of English word-final voicing contrasts. Journal of Phonetics. 2008; 36(4):664–679. [PubMed: 19606271]

HITCHINGS, H. Financial Times. London: 2008 May 3. One language fits all.

IMAI S, WALLEY AC, FLEGE JE. Lexical frequency and neighborhood density effects on the recognition of native and Spanish-accented words by native English and Spanish listeners. Journal of the Acoustical Society of America. 2005; 117:896–907. [PubMed: 15759709]

INTERNATIONAL PHONETIC ASSOCIATION (IPA). Handbook of the International Phonetic Association. Cambridge: Cambridge University Press; 1999.

JENKINS, J. The phonology of English as an international language. Oxford: Oxford University Press; 2000.

JILKA, M.; ANUFRYK, V.; BAUMOTTE, H.; LEWANDOWSKA, N.; ROTA, G.; REITERER, S. Assessing individual talent in second language production and perception. In: Rauber, AS.; Watkins, MA.; Baptista, BO., editors. New Sounds 2007: Proceedings of the Fifth International Symposium on the Acquisition of Second Language Speech. Florianópolis, Brazil: Federal University of Santa Catarina; 2008. p. 243-258.

JOHNSON, K. Speech perception without speaker normalization: An exemplar model. In: Johnson, K.; Mullenix, J., editors. Talker variability in speech processing. San Diego: Academic Press; 1997. p. 145-166.

JOHNSON, K. Speaker normalization in speech perception. In: Pisoni, DB.; Remez, RE., editors. The handbook of speech perception. Oxford: Blackwell Publishing; 2005. p. 363-389.

JONGMAN, A.; WADE, T. Acoustic variability and perceptual learning. In: Bohn, OS.; Munro, MJ., editors. Language experience in second language speech learning: In honor of James Emil Flege. Amsterdam/Philadelphia: John Benjamins; 2007. p. 135-150.

KRALJIC T, SAMUEL AG. Perceptual learning for speech: Is there a return to normal? Cognitive Psychology. 2005; 51:141–178. [PubMed: 16095588]

KRALJIC T, SAMUEL AG. Generalization in perceptual learning for speech. Cognitive Psychonomic Bulletin & Review. 2006; 13(2):262–268.

KRALJIC T, SAMUEL AG. Perceptual adjustments to multiple speakers. Journal of Memory and Language. 2007; 56:1–15.

KRAUSS RM, PARDO JS. Comment on Pickering & Garrod. Brain and Behavior Science. 2004; 27(2):203–204.

MAJOR R, FITZMAURICE CSM, BUNTA F, BALASUBRAMANIAN C. The effects of nonnative accents on listening comprehension: Implications for ESL assessment. TESOL Quarterly. 2002; 36:173–190.

MARSHALL RC, FREED DB, PHILLIPS DS. Communicative efficiency in severe aphasia. Aphasiology. 1997; 11(4):373–384.

MAURANEN A. The corpus of English as lingua franca in academic settings. TESOL Quarterly. 2003; 37(3):513–527.

MAYE J, ASLIN R, TANENHAUS M. The weckud wetch of the Wast: Rapid adaptation to a novel accent. Cognitive Science. 2008; 32(3):543–562. [PubMed: 21635345]

McGARR NS. The intelligibility of deaf speech to experienced and inexperienced listeners. Journal of Speech and Hearing Research. 1983; 26:451–458. [PubMed: 6645470]

MUNRO MJ, DERWING TM, MORTON SL. The mutual intelligibility of L2 speech. Studies in Second Language Acquisition. 2006; 28:111–131.

NORRIS D, McQUEEN JM, CUTLER A. Perceptual learning in speech. Cognitive Psychology. 2003; 47:204–238. [PubMed: 12948518]

PALLIER C, SEBASTIAN-GALLÉS N, DUPOUX E, CHRISTOPHE A, MEHLER J. Perceptual adjustment to time-compressed speech: A cross-linguistic study. Memory & Cognition. 1998; 26(4):844–851.

PARDO JS. On phonetic convergence during conversational interaction. Journal of the Acoustical Society of America. 2006; 119(4):2382–2393. [PubMed: 16642851]

PICKERING MJ, GARROD S. Toward a mechanistic psychology of dialogue. Behavioral and Brain Sciences. 2004; 27(2):169–226. [PubMed: 15595235]

PICKERING MJ, GARROD S. Alignment as the basis for successful communication. Research on Language and Computation. 2006; 4:203–228.

PIERREHUMBERT, J. Laboratory Phonology VII. Berlin: Mouton de Gruyter; 2002. Word-specific phonetics; p. 101-139.

RYAN, C. Master's thesis. University College; London: 2007. A comparison of the acoustic-phonetic features of clear speech in natural communication and reading.

SCHWAB EC, NUSBAUM HC, PISONI DB. Some effects of training on the perception of synthetic speech. Human Factors. 1985; 27(4):395–408. [PubMed: 2936671]

SMILJANIC, R.; BRADLOW, AR. Clear speech intelligibility: Listener and talker effects. Proceedings of the XVIth International Congress of Phonetic Sciences; Saarbrucken, Germany. 2007.

SMILJANIC R, BRADLOW AR. Speaking and hearing clearly: Talker and listener factors in speaking style changes. Language and Linguistics Compass. 2009; 3(1):236–264. [PubMed: 20046964]

SMITH LE, RAFIQZAD K. English for cross-cultural communication: The question of intelligibility. TESOL Quarterly. 1979; 13:371–380.

STIBBARD RM, LEE JI. Evidence against the mismatched interlanguage intelligibility benefit hypothesis. Journal of the Acoustical Society of America. 2006; 120:433–442. [PubMed: 16875239]

UCHANSKI, RM. Clear speech. In: Pisoni, DB.; Remez, RE., editors. The handbook of speech perception. Oxford: Blackwell Publishing; 2005. p. 207-235.

VAN WIJNGAARDEN SJ. Intelligibility of native and non-native Dutch speech. Speech Communication. 2001; 35:103–113.

VAN WIJNGAARDEN SJ, STEENEKEN HJM, HOUTGAST T. Quantifying the intelligibility of speech in noise for non-native listeners. Journal of the Acoustical Society of America. 2002; 111:1906–1916. [PubMed: 12002873]

WADE T, JONGMAN A, SERENO J. Effects of acoustic variability in the perceptual learning of non-native-accented speech sounds. Phonetica. 2007; 64(2–3):122–144. [PubMed: 17914280]

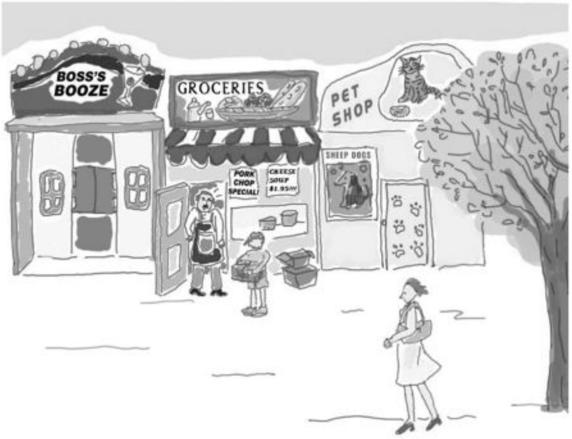
WEIL SA. Foreign-accented speech: Encoding and generalization. Journal of the Acoustical Society of America. 2001; 109:2473 (A).

WEINBERGER, SH. The Speech Accent Archive. George Mason University; n.d. http://accent.gmu.edu/

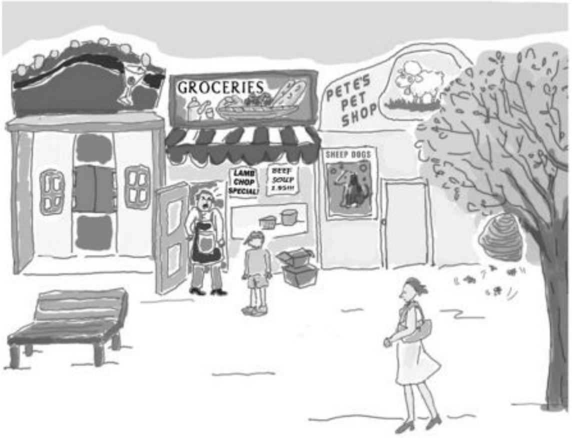
WILTSHIRE CR, HARNSBERGER J. The influence of Gujarati and Tamil L1s on Indian English: A preliminary study. World Englishes. 2006; 25(1):91–104.

Appendix: Diapix image pair used for spontaneous speech recordings in the Wildcat Corpus

Appendix A



Appendix B



Appendix C

Changed items		Missing items	
Version A	*Version B*	Version A	Version B
cat on pet shop sign	sheep on pet shop sign	no beehive	beehive
pork chop sign	lamb chop sign	paw prints on door	no paw prints on door
cheese soup	beef soup	Boss's Booze	no sign
woman—red shoes	woman—green shoes	just Pet Shop	Pete's Pet Shop
		no bench	bench
		boy carrying box	boy not carrying box

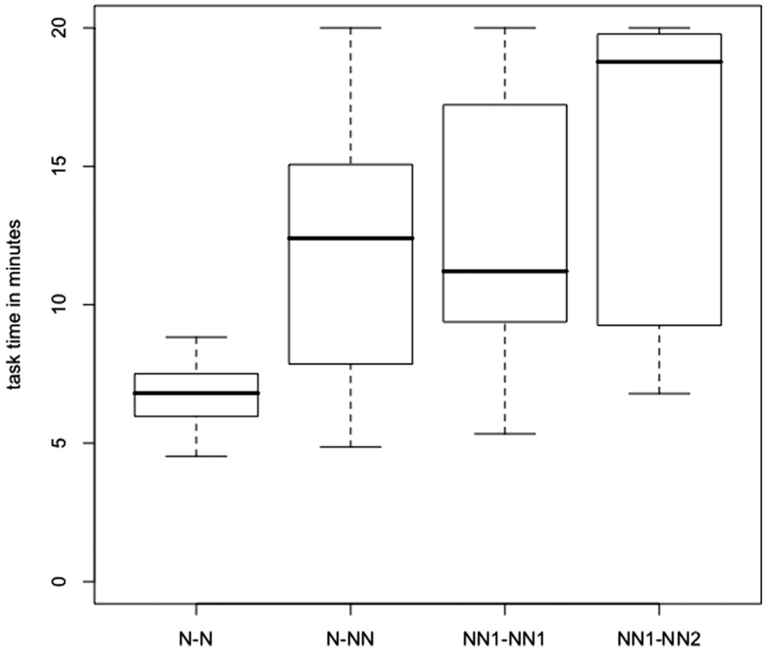


Figure 1.
Total duration of the Diapix task by pair type. A 20-minute time limit was imposed

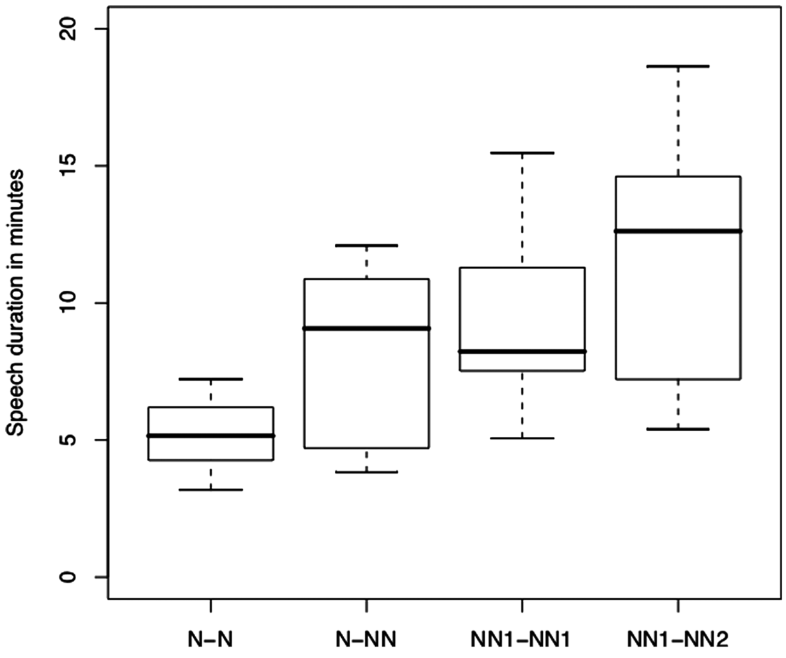


Figure 2.
Total speech duration by pair type

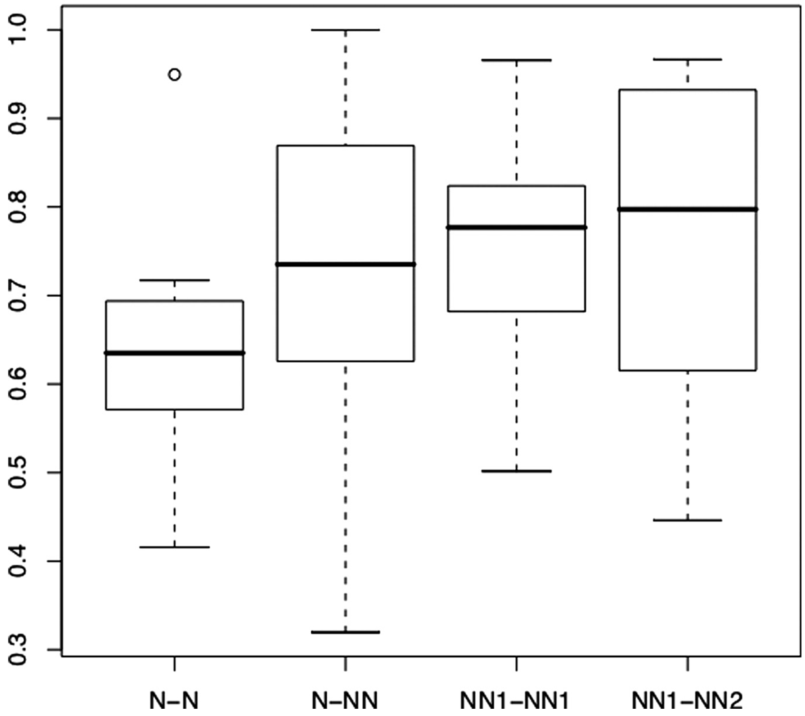


Figure 3.
Total speech duration of talker 1 (T1) / total speech duration of talker 2 (T2), where T1 is defined as the partner with shorter total speech duration

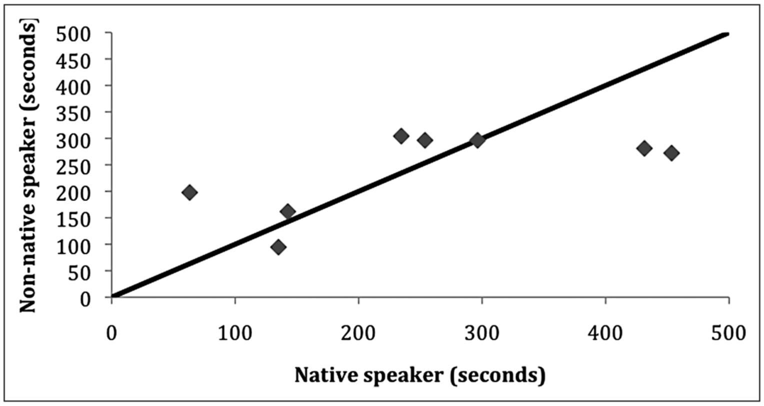


Figure 4.

Speech duration (in seconds) by native and non-native talkers who were paired in the Diapix task. Each point represents a single Diapix conversation

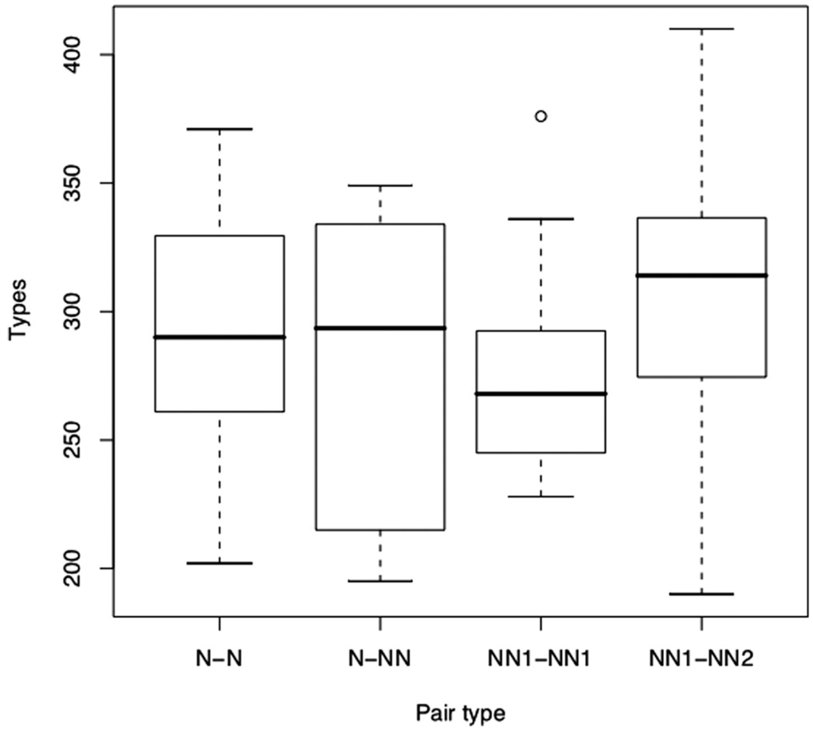


Figure 5.
The number of types (unique words) used in conversations of the four pair types

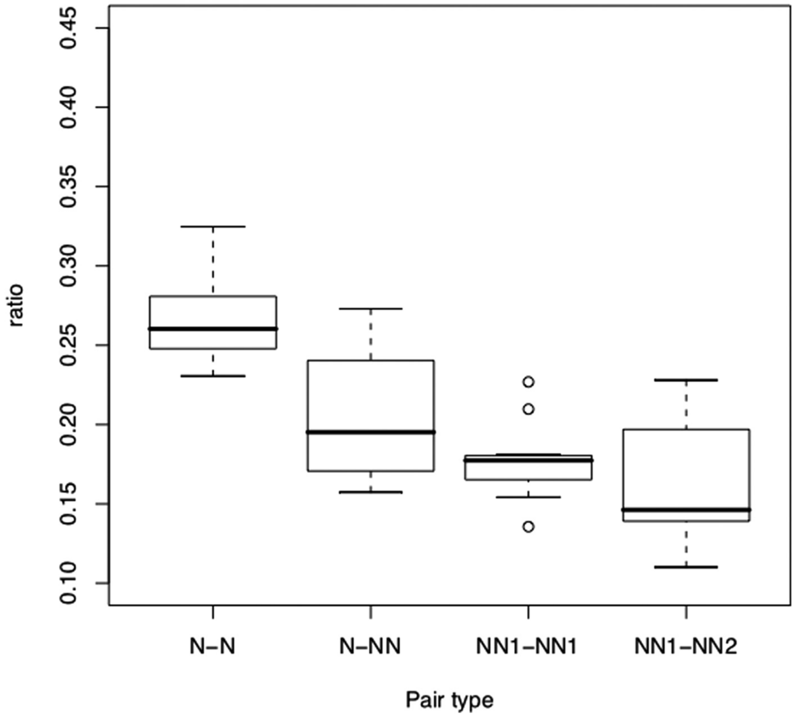


Figure 6.
Type-to-token ratios calculated over conversations

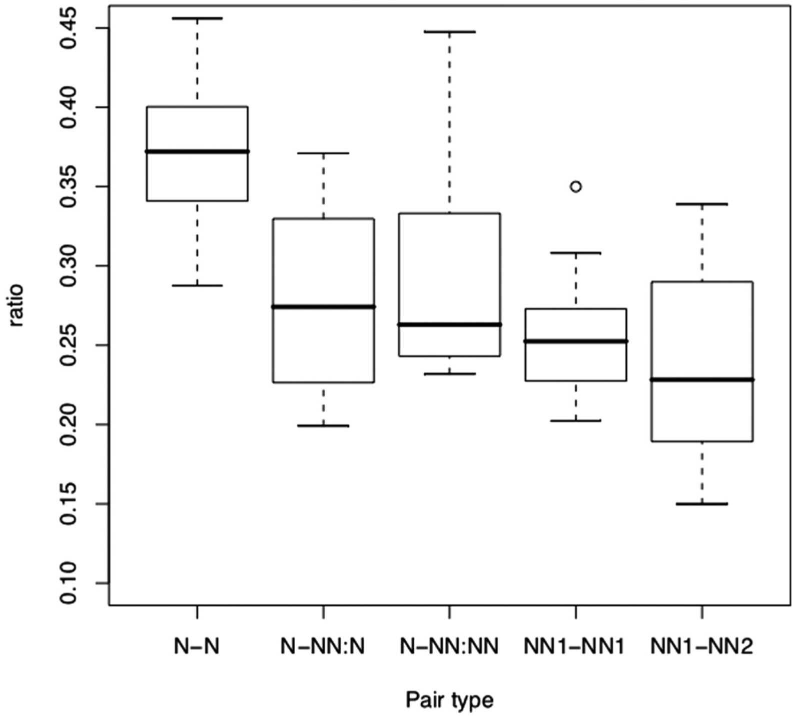


Figure 7.
Type-to-token ratios calculated by individual talkers. For ratios of individual talkers, N and NN talkers in the N-NN condition have been separated. Note that this is the only group that can be meaningfully sub-divided

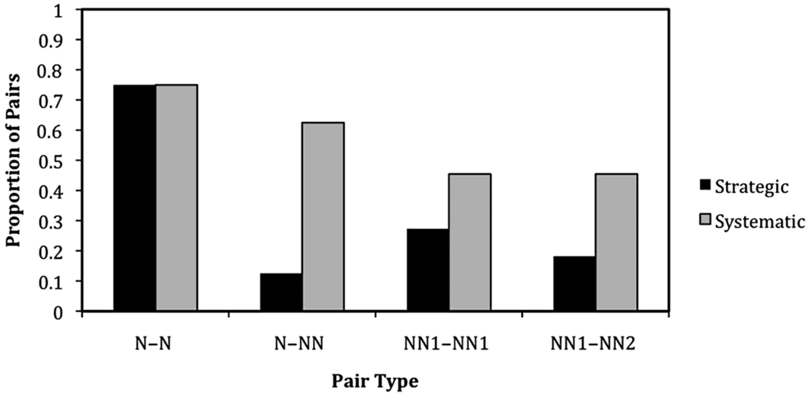


Figure 8.

Proportion of each pair type that a) began the task with discussion about strategy and b) progressed around the scene systematically

Table 1

The two dimensions of alignment that underlie the structure of the Wildcat Corpus

	Native language background	
Target language	**Aligned**	**Misaligned**
Aligned	Native+Native (N-N)	X
Misaligned		Native+Non-Native (N-NN)
	Non-Native+Non-Native (NN1-NN1)	Non-native+Non-Native (NN1-NN2)

Table 2

Native languages of Wildcat Corpus participants

Native language	Number of participants
Chinese *	20
English	24
Hindi/Marathi	1
Italian	1
Japanese	1
Korean	20
Macedonian	1
Persian **	1
Russian	1
Spanish	2
Telegu	1
Thai	1
Turkish	2

*One participant self-identified as a native speaker of Cantonese; 3 participants self-identified as Mandarin; the rest listed Chinese as their native language.

**Participant did not provide any further information about his native language.

Table 3

Summary of Diapix conversation participants. For further explanation of the pair types (in column 2) see Table 1 and the accompanying text above

Task language	Pair type	Talker		Sex	Pair N
		1	2		
English (38 pairs)	N-N (8 pairs)	English	English	F	4
				M	4
	NN1-NN1 (11 pairs)	Chinese	Chinese	F	3
				M	2
		Turkish	Turkish	M	1
		Korean	Korean	F	2
				M	2
		Indian[*] (Hindi/Marathi)	Indian (Telegu)	M	1
	NN1-NN2 (11 pairs)	Chinese	Spanish	M	1
			Russian	M	1
		Korean	Chinese	F	2
				M	2
			Japanese	M	1
			Persian	M	1
			Thai	M	1
			Macedonian	F	1
		Italian	Spanish	M	1
	N-NN (8 pairs)	English	Chinese	F	2
				M	2
			Korean	F	2
				M	2

[*] Because of the unique status of Indian English, the conversation between the two Indian participants (who have different mother tongues) was categorized as an NN1-NN1 pair. English is an official language in India, used extensively in government and in education from elementary school through university. Furthermore, two features crucially distinguish learners of Indian English from other non-native learners of English (see discussion in Wiltshire & Harnsberger, 2006). First, English is learned not simply to speak to foreigners, but serves as a lingua franca among Indians with many different first languages. Second, English learners in India study the Generalized Indian English dialect (CIEFL, 1972), rather than British or American dialects.

Table 4

Descriptive statistics for Diapix task duration and the duration of speech within the task (the sum of the two interlocutors' actual speech durations). Presented by pair type: native-native (N-N), native-non-native (N-NN), non-natives with matched native language (NN1-NN1), non-natives with different native languages (NN1-NN2)

Pair type	Mean		Median	
	Task duration	Speech duration	Task duration	Speech duration
N-N	6.74	5.20	6.81	5.16
N-NN	11.94	8.15	12.40	9.07
NN1-NN1	12.82	9.42	11.21	8.23
NN1-NN2	14.82	11.41	18.77	12.62