Published in final edited form as:

Exp Brain Res. 2011 June ; 211(2): 193–205. doi:10.1007/s00221-011-2664-5.

Speed-difficulty trade-off in speech: Chinese versus English

Mark L. Latash,

Department of Kinesiology, Rec. Hall-268 N, Pennsylvania State University, University Park, PA 16802, USA

Yao Sun,

Department of Kinesiology, Rec. Hall-268 N, Pennsylvania State University, University Park, PA 16802, USA

Elizaveta M. Latash, and

Department of Biology, 209 Mueller Lab, Pennsylvania State University, University Park, PA 16802, USA

Irina L. Mikaelian

Department of Germanic and Slavic Languages and Literatures, 427 Burrowes Building, Pennsylvania State University, University Park, PA 16802, USA

Mark L. Latash: mll11@psu.edu

Abstract

This study continues the investigation of the previously described speed-difficulty trade-off in picture description tasks. In particular, we tested a hypothesis that the Mandarin Chinese and American English are similar in showing logarithmic dependences between speech time and index of difficulty (ID), while they differ significantly in the amount of time needed to describe simple pictures, this difference increases for more complex pictures, and it is associated with a proportional difference in the number of syllables used. Subjects (eight Chinese speakers and eight English speakers) were tested in pairs. One subject (the Speaker) described simple pictures, while the other subject (the Performer) tried to reproduce the pictures based on the verbal description as quickly as possible with a set of objects. The Chinese speakers initiated speech production significantly faster than the English speakers. Speech time scaled linearly with $\ln(ID)$ in all subjects, but the regression coefficient was significantly higher in the English speakers as compared with the Chinese speakers. The number of errors was somewhat lower in the Chinese participants (not significantly). The Chinese pairs also showed a shorter delay between the initiation of speech and initiation of action by the Performer, shorter movement time by the Performer, and shorter overall performance time. The number of syllables scaled with ID, and the Chinese speakers used significantly smaller numbers of syllables. Speech rate was comparable between the two groups, about 3 syllables/s; it dropped for more complex pictures (higher ID). When asked to reproduce the same pictures without speaking, movement time scaled linearly with $\ln(ID)$; the Chinese performers were slower than the English performers. We conclude that natural languages show a speed-difficulty trade-off similar to Fitts' law; the trade-offs in movement and speech production are likely to originate at a cognitive level. The time advantage of the Chinese participants originates not from similarity of the simple pictures and Chinese written characters and not from more sloppy performance. It is linked to using fewer syllables to transmit the same information. We suggest that natural languages may differ by informational density defined as the amount of information transmitted by a given number of syllables.

Keywords

Speech; Picture description; Fitts' law; Reaction time; Speech rate; Chinese

Introduction

It is intuitively clear that one has to speak for a longer time to transmit more information. Recently, we have offered a method to quantify this relationship using picture description tasks (Latash and Mikaelian 2010, 2011). In one of the studies (Latash and Mikaelian 2011), we used very simple pictures, with only two objects in each picture, which allowed to introduce a quantitative index of difficulty (ID) and study how changes in this index affect speech time (ST). That study showed very strong relationships between ST and $\ln(ID)$, typically accounting for over 95% of the variance for each of the participants:

$$ST = a + k \cdot \ln(ID), \quad (1)$$

where a and k are constants. The functional form of this relationship is identical to that of the famous Fitts' law, which describes how movement time changes with the ratio of target distance to target size (Fitts 1954; Fitts and Peterson 1964).

One of the unexpected results of the mentioned study was that native Mandarin Chinese speakers spent substantially less time across all ID values as compared with native Indo-European (American English and Russian) speakers. Participants within both groups showed ST scaling with ID as expected from Eq. (1), but the Chinese-speaking participants had about 40% lower k values, while a values were similar across the two groups.

A few mutually non-exclusive hypotheses were suggested on the origin of this difference. In particular, the very simple schematic pictures could resemble Chinese characters, and so they could give an advantage to the Chinese speakers. This hypothesis was indirectly supported by the fact that the Chinese speakers showed significantly shorter reaction time, from the picture presentation to speech initiation, as compared with the Indo-European speakers. Another possibility was that the Chinese speakers spoke faster but potentially their speech could be harder to understand and it could contain more errors. Finally, it was possible that the Chinese speakers used fewer syllables to transmit the same amount of information. This could be facilitated by the use of tones in the Mandarin Chinese, which may be an effective channel of information transmission (Brown-Schmidt and Canseco-Gonzalez 2004; Valaki et al. 2004; Lee 2007).

This study was performed to test the mentioned hypotheses. We tested groups of subjects speaking Mandarin Chinese and American English. The subjects were tested in pairs: One of the subjects (the Speaker) was describing pictures, while the other one (the Performer) tried to re-create the pictures using actual objects as quickly and accurately as possible. This design allowed checking whether the faster speech in Chinese participants was associated with faster performance of motor tasks based on the transmitted information. A priori, faster speech might lead to lower quality of information transmission, that is, more time needed to understand and implement the information and/or higher error rate. In addition, the subjects were always watched by a native speaker who kept a log of errors made by the Speaker and by the Performer.

In other trials, the participants were shown the same sets of pictures and asked only to copy them using the actual objects. If the advantage of the Chinese speakers was in the similarity

of the pictures to Chinese characters, we could expect a similar advantage in those tasks that involved picture recognition and duplication but no speech production.

Finally, we quantified the number of syllables within each utterance and compared the speech rates between our subject groups. Ultimately, we tested the main hypothesis: The two natural languages, the Mandarin Chinese and the American English, differ significantly in the amount of time needed to describe simple pictures, this difference increases for more complex pictures, and it is associated with a proportional difference in the number of syllables.

Methods

Subjects

Sixteen healthy volunteers participated in the experiments, 8 men and 8 women with the age of 19–34 years. They all were right-handed and had normal or corrected to normal vision. Eight subjects were native Chinese (Mandarin) speakers and eight were native English (American) speakers. The Chinese speakers were also fluent in English. All the procedures were approved by the Office for Research Protection at the Pennsylvania State University.

Apparatus and procedures

Two subjects were tested at the same time. They were sitting in chairs at the same table facing each other (Fig. 1). One of the subjects (the Speaker) faced a laptop computer (Dell Latitude D830) with the 15" screen placed on the table about 0.6 m away from the subject. Plantronics digital DSP400 headphones with a microphone were placed naturally on the subject's head such that the microphone was about 5 cm from the mouth. The microphone was connected to the USB port of the computer. The sound signal was collected at 10 kHz using a Matlab-based program.

The second subject (the Performer) had two soft mats (2 cm thick) placed on the table in front of him/her; the mats were used to avoid accidental sounds during object manipulation. The bigger mat (about 0.7×0.5 m) was aligned with the midline of the subject's trunk. The second mat (0.3×0.3 m) was aligned with the right shoulder.

There was a round object in the center of the bigger mat at all times, the "plate", with a diameter of 0.2 m. There were four objects on the smaller mat, a computer mouse, a wooden stick, a plastic fork, and a plastic knife. The latter three objects were about 0.18 m long. The mouse was placed such that it was easy for the subject to place the right hand on it. Mouse left clicks were recorded by the same laptop computer. The other three objects were placed to the left of the mouse, parallel to each other, spaced by about 3 cm, always in the same order, the stick (about 6 cm from the mouse), the fork and the knife (from the mouse toward the large mat).

The experiment involved four series. In two of the series (Speaker–Performer series), one of the subjects (the Speaker) was describing the pictures that he/she saw on the laptop screen, while the other subject (the Performer) tried to reproduce the pictures based on the verbal description. Then, the subjects switched roles.

During the experiment, the Speaker paced the experiment. The Speaker was instructed to press the Enter key, wait for a picture to appear on the screen, and describe the picture as quickly and accurately as possible. The instruction was always given in English in the following way: "You have to tell your partner as quickly as possible what you see such that he or she is able to re-create this picture in all the important details". After describing the picture, the Speaker was to remain silent until the end of the trial (up to 15 s). Then, the

Speaker asked the Performer if he/she was ready for the next trial, pressed the key again, and the next picture emerged on the screen. The Speaker was told to take a breath prior to pressing the Enter key in order not to contaminate the sound signal with an accidental deep breath. The participants were reminded to take rest periods as needed.

The Performer sat with the right hand ready to click the mouse. The instruction to the Performer was “Your task is to make a picture from the objects in front of you matching the description given by the Speaker. As soon as you know what object to manipulate, click the mouse once, pick the object, place it in the correct position and orientation with respect to the plate, and click the mouse one more time. Try to perform all the manipulations as quickly as you can. After each trial, return the moved object to its initial location.” Data collection started with picture presentation and ended with the second mouse click. The Performer was allowed to use only the right hand during task performance.

Each picture showed two objects, one of which was always the same: a plate in the middle of the screen. The second objects could be a stick, a fork, or a knife (Fig. 2). Prior to the experiment, the subjects were shown the objects, one by one, and asked to decide how they would call them during the experiment. Note that the stick is a symmetric object with the two ends and two sides identical to each other. In contrast, the fork has two ends (the prongs and the handle), while the two sides are symmetrical. The knife has two ends (the sharp end and the handle) and also the two sides (the sharp and the dull ones). The second object could be located under the plate, above the plate, to the left of the plate, to the right of the plate, or centered about 1 of the 4 corners. The objects could be oriented vertically, horizontally, 45° to the vertical, or slightly tilted. Examples of different object orientations and locations are shown in Fig. 2.

The Speakers were free to select their preferred strategy of describing the objects, but they were always reminded to try to do this as quickly as possible. They were also reminded to use complete sentences. For example, equivalent descriptions of a picture could be: “The fork is at the upper-left corner, tilted 45° with the prongs facing down and away from the plate.” or “The fork is to the left and above the plate with the prongs down and to the left, turned by 45°.” Speakers did at times produce incomplete sentences; all those utterances were included into the analysis. Within this experiment, we did not study specific strategies of object description, rather regularities that were present in spite of the different strategies. As a result of the instruction, there was a lot of variability across the subjects and within the subjects across trials in the lexical material used.

Each participant took part in two more series that involved only one subject at a time (Performer-only). The order of Performer-only and Speaker-Performer tasks was balanced across subjects. The subject was sitting at the Performer’s site. The computer was oriented toward the subject, such that the subject faced the screen, about 0.8 m away. An experimenter paced the picture presentation. After a picture was presented, the subject clicked the mouse, moved the object to match the picture, and clicked the mouse again. Then, the subject moved the object back into its position on the smaller mat and got ready for the next trial. There were 2–3 s intervals before picture presentation, after the subject returned to the initial position and was ready for the next trial. The task was the same as for the Performer in the two-participant series. However, the subject did not receive verbal descriptions and was instructed to reproduce the pictures he/she saw.

Each series consisted of four blocks, 18 pictures in each block. The compositions of the 72 pictures were selected at random from all possible compositions to have a balanced representation of different difficulty levels. The combinations of factors such as *Object*,

Location, and *Orientation* were used to estimate difficulty index of each task (see the next section).

An experimenter always observed the performance of the subjects and made notes if one of the subjects made an error. If an error occurred, the experimenter asked the subject(s) to stop, identified the error for them, and then the block continued. Errors in object names, locations, and orientations were quantified separately for each of the participants.

Prior to each series, the subjects were given a detailed explanation of the task, a brief demonstration of the objects, their properties, and their possible relative configurations. They were told that the important features of the pictures were (1) object, (2) location, and (3) orientation. They were explicitly asked not to describe other features of the picture such as, for example, the distance between the object and the plate or the object location with respect to the computer screen borders. Then, the subjects performed ten to eighteen practice trials. If a subject asked for additional practice, it was always given to the subject; this happened in four out of the 18 subjects. There were 1-min intervals between the blocks and at least 3-min intervals between the series. The blocks were presented in a random order. Picture presentation within blocks was randomized.

Data analysis

Analysis was performed off-line. A Matlab-based program was used to define the times of initiation and termination of the speech signal in each trial of the two-participant series. For this purpose, the power of the signal was computed at each sample using the integral of the power spectral density distribution over all the frequencies with the time window of 30 ms. Then, the peak value (P_{PEAK}) of the signal power was computed. Further, the time when the signal first reached 10% of P_{PEAK} was detected and used as the time of speech initiation (T_{SI}), and the time when the signal dropped under 10% of P_{PEAK} and remained under this value until the end of the trial was found and used as the time of speech termination (T_{ST}). Reaction time (RT) was measured as the time from the trial initiation (picture presentation) to T_{SI} . Speech time (ST) was defined as the difference between T_{ST} and T_{SI} .

In some trials, one of the times, T_{SI} or T_{ST} , could not be determined reliably using the described routine due to coughing or heavy breathing by the subject, or extraneous noises. Such trials, as well as trials when the subject stopped in the middle of a trial and then repeated a portion of the previous text, were rejected from further analysis. On average, less than 5% of the trials were rejected.

The two times of mouse clicks were also recorded, T_{C1} and T_{C2} . The difference between T_{SI} and T_{C1} was computed as the “speech-movement delay” ($\Delta_{SM} = T_{C1} - T_{SI}$). Movement time was computed as $MT = T_{C2} - T_{C1}$. Total performance time (TT) was computed as the time T_{C2} with respect to trial initiation. The various timing indices were used to analyze whether shorter speech time by the Speaker was associated with shorter action time by the Performer and shorter time spent on performing the whole task by the two participants. We explored the RT index to test the earlier finding on an advantage of Chinese speakers in reaction time in such tasks. The Δ_{SM} index was analyzed as a component of performance by the Performer analogous to RT for the Speaker. We were interested in whether the Chinese participants were faster in both speech and non-speech actions and whether this advantage involved both reaction time and action time.

Similarly to the previous study (Latash and Mikaelian 2011), each task was assigned an index of difficulty (ID) according to the following equation: $ID = ID_{OB} \cdot ID_{LO} \cdot ID_{OR}$, where ID_{OB} is index of difficulty of the object, ID_{LO} is index of difficulty of the location, and ID_{OR} is index of difficulty of the orientation. ID_{OB} has been assigned three values, 1 (stick),

2 (fork), and 4 (knife). ID_{LO} has been assigned two values, 1 (above, below, to the left, and to the right of the plate) and 2 (at one of the corners). ID_{OR} has been assigned two values, 1 (vertical or horizontal) and 4 (tilted 45° and slightly tilted). These values have been assigned based on the number of degrees-of-freedom one has to describe for the mentioned characteristics. For example, the stick only had to be named; for the fork, the direction of the prongs had also to be described, while for the knife both the direction of the sharp end and the orientation of the blade had to be described. Simple locations ($ID_{LO} = 1$) required the subject to name only one coordinate (above, below, to the right, or to the left of the plate), while locations at the corners ($ID_{LO} = 2$) required the subject to name whether the corner was above or below and whether it was to the right or to the left of the plate. For tilted objects ($ID_{OR} = 4$), both the angle of the tilt and the direction of the upper end had to be described (for example, “the stick is to the right of the plate, tilted 45°, with the upper end pointing to the right”).

As a result, the trials with the stick had ID values of 1, 2, 4, and 8; the trials with the fork had ID values of 2, 4, 8, and 16; and the trials with the knife had ID values of 4, 8, 16, and 32. The three examples in Fig. 2 illustrate pictures with ID ranging from 1 to 32. The data were averaged within each subject across trials that had the same ID values. Further, linear and logarithmic regression analyses were used to analyze the dependences of the timing characteristics (RT, ST, MT, TT, and Δ_{SM}) on ID .

The number of syllables (N_{SYL}) in each utterance of the Chinese and English speakers was quantified by a native speaker of the corresponding language. Speech rate was computed for each subject and for each ID value, across the trials corresponding to that ID , as the ratio of the average number of syllables to the average speech time: $R_{SP} = N_{SYL}/ST$.

Multivariate analysis of variance (MANOVA) was performed to explore whether the two groups differed in the timing indices and whether this difference depended on ID . The factors were *Timing-Index* (5 levels corresponding to the five timing indices), *Language* (a between-factor, two levels, Chinese and English), and ID (6 levels). Further, analysis of variance (ANOVA) with repeated measures was used to explore the significant effects of MANOVA for each timing index separately. We also used one-way ANOVAs with the factor *Language* to explore a number of dependent variables such as the coefficients within the equations linking specific timing indices to ID and the number of syllables to ID . To study the number of errors between the two groups, ANOVA with an additional factor *Action* (two levels, Speaker and Performer) was used. Tukey’s pairwise contrasts were used to analyze significant effects at $P < 0.05$.

Results

General performance characteristics in the speaker–performer tasks

All the subjects made at least a few errors during the experiment. These errors were present for each subject in the Speaker–Performer tasks, while the Performer-only tasks showed less than one error per participant. The number of errors varied within a broad range. For the Chinese group, the average number of errors per subject over the two Speaker–Performer series (144 trials) was 8.1. These involved 3.7 errors/Speaker and 4.4 errors/Performer. For the English group, the average number of errors per participant was higher, 13.25 with 8.5 errors/Speaker and 4.75 errors/Performer. Because of the large variability across subjects, the differences in the number of errors between the groups and between Speakers and Performers were under the significance level as confirmed by a two-way ANOVA, *Language* $\times$ *Action* which showed no significant effects ($P > 0.1$). Only 6% of all the errors were made in naming/picking the correct object, 23% of the errors were in the location of the object, while over 71% of all the errors were made in the orientation of the object.

When average timing indices were computed across all the trials within each subject, the Chinese group was faster than the English group in all the timing indices. In particular, on average, reaction time (RT) in the Chinese group was 0.971 s, while it was 1.138 s in the English group. Speech time (ST) was 3.05 s in the Chinese group, as compared with 4.17 s for the English group. Similar differences were observed for the speech-movement delay (Δ_{SM}), movement time (MT), and total time (TT): 0.54 versus 0.98 s for Δ_{SM} , 4.47 versus 4.98 s for MT, and 5.98 versus 7.10 s for TT. These results are illustrated in Fig. 3 with standard deviation bars. For better visualization, the RT and Δ_{SM} data are presented in panel a, while the ST, MT, and TT data are shown in panel b.

The differences between the Chinese and English groups were tested using a MANOVA with factors *Timing-Index*, *Language*, and *ID*. The MANOVA showed significant effects of all factors and significant two-way interactions: *Timing-Index* ($F_{(4, 108)} = 2.304$; $P < 0.0001$), *Language* ($F_{(1, 230)} = 204$; $P < 0.0001$), *ID* ($F_{(5, 227)} = 305$; $P < 0.0001$), *Timing-Index* $\times$ *Language* ($F_{(4, 160)} = 18.4$; $P < 0.001$), *Timing-Index* $\times$ *ID* ($F_{(20, 320)} = 57.8$; $P < 0.0001$), and *Language* $\times$ *ID* ($F_{(5, 218)} = 8.36$; $P < 0.001$). There was also a significant three-way interaction ($F_{(20, 318)} = 2.01$; $P < 0.01$).

To further explore the significant effects of MANOVA, ANOVAs were run on each timing index separately with the factors *Language* and *ID*. For RT, main effects of both factors were found, *Language* ($F_{(1, 12.3)} = 7.86$; $P < 0.05$) and *ID* ($F_{(5, 56.4)} = 2.52$; $P < 0.05$), without a significant interaction ($P > 0.1$). Pairwise contrasts showed significantly longer RT for *ID* = 8 as compared with *ID* = 1 without differences within other pairs of *ID* levels. The effect of *Language* reflected shorter RT for the Chinese group. For Δ_{SM} , there was only a significant effect of *Language* ($F_{(1, 10.1)} = 7.47$; $P < 0.05$) reflecting smaller values for the Chinese group.

For ST, both main effects and the interaction were significant, *Language* ($F_{(1, 14)} = 33.7$; $P < 0.0001$), *ID* ($F_{(5, 70)} = 257$; $P < 0.0001$), and *Language* $\times$ *ID* ($F_{(5, 70)} = 9.30$; $P < 0.001$). Pairwise contrasts showed significantly smaller values for the Chinese group for all *ID* levels ($P < 0.01$) except *ID* = 1. Within each group, the data differed across all the pairs of the *ID* levels ($P < 0.05$).

For MT, both main effects and the interaction were significant, *Language* ($F_{(1, 14.3)} = 7.75$; $P < 0.05$), *ID* ($F_{(5, 59.7)} = 312$; $P < 0.0001$), and *Language* $\times$ *ID* ($F_{(5, 49.0)} = 2.65$; $P < 0.05$). Pairwise contrasts showed a significant difference between the groups for *ID* = 32 ($P < 0.05$), and a close to significant difference for *ID* = 16 ($P = 0.059$), while there were no differences for other *ID* levels. There were significant differences within each pair of *ID* levels for the Chinese group, except within the pairs {*ID* = 8; 4} and {*ID* = 16; 32}. For the English group, all the differences were significant except within the pair {*ID* = 8; 4}.

Similarly, for TT, both main effects and the interaction were significant, *Language* ($F_{(1, 20.1)} = 62.7$; $P < 0.001$), *ID* ($F_{(5, 53.6)} = 239$; $P < 0.0001$), and *Language* $\times$ *ID* ($F_{(5, 49.4)} = 6.70$; $P < 0.05$). Pairwise contrasts showed significantly smaller values for the Chinese group for all *ID* levels ($P < 0.01$) except *ID* = 1. There were significant differences within each pair of *ID* levels for the Chinese group, except within the pairs {*ID* = 8; 4} and {*ID* = 16; 32}. For the English group, all the differences were significant except within the pair {*ID* = 8; 16}.

Since several timing indices showed close to linear scaling with log-transformed *ID* for each subject and within each group, further we describe the dependences of these indices on *ID* using corresponding linear regressions. Coefficients within the regressions are used as dependent variables in some of the further analyses.

Characteristics of speech as a function of ID

Reaction time showed no significant scaling with the log-transformed index of difficulty (ID) in either group. However, the Chinese subjects showed a trend for an increase in RT with $\ln(ID)$. The averaged across subjects data with standard deviation bars are shown for both groups as functions of $\ln(ID)$ in Fig. 4.

Speech time showed strong linear scaling with $\ln(ID)$ in all the subjects. The data for individual subjects are shown in Fig. 5 with linear regression lines and equations. Each point corresponds to the average of all the trials with the same ID value. Note the very high correlation coefficients. Averaged across all subjects ST data showed a linear scaling with $\ln(ID)$ accounting for over 99% of variance for each of the two groups.

Further, regression equations, $ST = a_{ST} + k_{ST} \cdot \ln(ID)$, for individual subjects were analyzed. The intercept coefficient (a_{ST}) was about 28% smaller in the Chinese group as compared with the English group, with the average values of 1.44 ± 0.237 and 1.84 ± 0.249 . The regression coefficient (k_{ST}) was about 44% smaller in the Chinese group as compared with the English group, 0.933 ± 0.144 and 1.348 ± 0.243 . Student's t -tests confirmed the significantly smaller coefficients for the Chinese group as compared with the English group: $t = 3.28$; $P < 0.01$ for a_{ST} and $t = 4.16$; $P < 0.001$ for k_{ST} .

Characteristics of movement as a function of ID

The delay between speech initiation by the Speaker and movement initiation by the Performer did not show significant scaling with the index of difficulty. Average across subjects data can be seen in Fig. 6.

Movement time (MT) and total performance time (TT) both showed significant scaling with $\ln(ID)$ in each individual subject. The regression equations $MT = a_{MT} + k_{MT} \cdot \ln(ID)$ and $TT = a_{TT} + k_{TT} \cdot \ln(ID)$ for the data of individual subjects all showed correlation coefficients $R_{(4)} > 0.94$ ($P < 0.01$). The data pooled across the subjects within each group showed $R_{(4)} > 0.99$. Figures 7 and 8 present the data for individual subjects and averaged across the two groups data with the regression lines and equations.

The coefficients of the regression equations were analyzed similarly to how this was described for speech time in the previous subsection. For MT, there was only a minor, statistically non-significant difference in the intercept value between the two groups, $a_{MT} = 2.67 \pm 0.554$ for the Chinese group and $a_{MT} = 2.46 \pm 0.356$ for the English group. In contrast, the regression coefficient was significantly lower in the Chinese group; $k_{MT} = 1.039 \pm 0.09$ for the Chinese group and $k_{MT} = 1.46 \pm 0.301$ for the English group ($t = 3.79$; $P < 0.005$).

For TT, the difference between the intercepts was not significant, $a_{TT} = 4.076 \pm 0.449$ for the Chinese group and $a_{TT} = 4.55 \pm 0.601$ for the English group. The regression coefficient was significantly lower for the Chinese group; $k_{TT} = 1.099 \pm 0.09$ for the Chinese group and $k_{TT} = 1.468 \pm 0.318$ for the English group ($t = 3.21$; $P < 0.01$).

Characteristics of movement during the performer-only series

All the subjects performed the Performer-only task much faster as compared with the task that involved speech production. Typical total performance time (TT) values were between 1.5 and 3 s. In contrast to the Speaker–Performer tasks, the Performer-only tasks were performed faster by the English speakers as compared with the Chinese speakers.

Reaction time showed no scaling with the index of difficulty, while total performance time did. The average reaction time for the Chinese-speaking subjects was 0.577 ± 0.101 s, while

for the English-speaking subjects it was significantly longer 0.716 ± 0.202 s ($t = 3.01$; $P < 0.05$). Figure 9 shows the TT data for the individual subjects and the averaged across subjects data for each of the two groups. Seven out of 8 subjects within each group showed statistically significant linear scaling of TT with $\ln(ID)$. The data averaged across the subjects showed significant scaling of TT with $\ln(ID)$, $R_{(4)} = 0.918$ for the Chinese group and $R_{(4)} = 0.948$ for the English group (both $P < 0.01$).

Further analysis of the intercept and regression coefficient values in the equation $TT = a_{TT0} + k_{TT0} \cdot \ln(ID)$ showed similar values for the intercept (a_{TT0}), 1.824 ± 0.251 and 1.835 ± 1.78 for the Chinese and English group, respectively. The regression coefficient (k_{TT0}) for the Chinese group was significantly larger than for the English group, 0.187 ± 0.045 as compared with 0.128 ± 0.053 ($t = 2.41$; $P < 0.05$).

Analysis of the number of syllables and speech rate

The shorter speech time in the Chinese group was associated with significantly fewer syllables per utterance. Figure 10 (panels a and b) shows the dependences of the number of syllables per utterance (N_{SYL}) on the log-transformed index of difficulty. There were strong linear relationships between N_{SYL} and $\ln(ID)$ for most subjects. These relationships were significant for each of the subjects ($R_{(4)} > 0.92$) except two whose linear correlation coefficients were not significant. Data averaged across subjects within each group showed very strong linear relationships between N_{SYL} and ID for each group ($R_{(4)} > 0.97$, $P < 0.01$).

Speech rate (number of syllables/second, R_{SP}) was comparable between the two groups. R_{SP} values varied substantially among the subjects as shown in panels c and d of Fig. 10. The average values did not differ significantly between the two groups. There was, however, a significant effect of ID on R_{SP} confirmed by a two-way ANOVA *Language* $\times$ *ID*. There was a main effect of ID ($F_{(5, 70)} = 21.4$; $P < 0.001$) and no other effects. Pairwise contrasts showed a significantly higher R_{SP} for the pictures with the smallest ID ($ID = 1$; $\ln(ID) = 0$) as compared with all other ID values ($P < 0.001$). In addition, R_{SP} for $ID = 2$ was significantly higher than R_{SP} for $ID = 32$ ($P < 0.05$). Other differences were not significant.

An additional analysis was run using linear regressions between N_{SYL} and speech time over all ID levels. In all subjects, there was a significant linear increase in N_{SYL} with ST. On average, the amount of variance accounted for by the linear regression was 88% in the Chinese group and 90% in the English group. For data averaged across subjects within each group, linear regression accounted for over 97% of variance for each group. The regression coefficient was 2.98 N_{SYL}/s in the Chinese group and 3.01 N_{SYL}/s in the English group; the difference between the groups was not significant.

Discussion

Our results support only one of the hypotheses on the origin of faster picture description by Chinese speakers mentioned in the introduction (see also Latash and Mikaelian 2011). The time advantage of the Chinese speakers disappeared when they were asked not to describe the pictures but to duplicate them. This finding speaks against the hypothesis that the Chinese speakers were helped by the similarity of the pictures to Chinese written characters. The shorter speech times of the Chinese participants were not associated with lower quality of information transmission; indeed, their performance contained fewer errors as compared with that of the English speakers (although the difference was not statistically significant).

In contrast, our main hypothesis received strong support: The Chinese speakers used significantly fewer syllables to describe the pictures, while their speech rate (syllables/second) did not differ from that of the English speakers. The difference in the number of

syllables between the two groups increased for more complex pictures. In the rest of this section, we discuss implications of these findings for such issues as speed-difficulty trade-off in speech and motor performance and differences between languages in speed of information transmission.

Speed-difficulty trade-off across tasks and languages

Several timing indices analyzed in this study showed a logarithmic scaling with the index of difficulty (ID). This scaling resembles strongly the classical Fitts' law, which links movement time to an index of difficulty computed as the ratio of distance to target size (Fitts 1954; Fitts and Peterson 1964; reviewed in Meyer et al. 1982; Plamondon and Alimi 1997). Fitts discussed this scaling based on the information theory, while later studies tried to link it to such diverse processes as movement planning, movement mechanics, and movement corrections (Meyer et al. 1982; Gutman et al. 1993; Plamondon and Alimi 1997). Recently, Fitts' law was shown in postural preparation to movement, that is, the duration of postural adjustments prior to movement initiation scaled with ID (Duarte and Latash 2007). This happened even when movement time did not scale with ID during manipulations of target width because the movement was ballistic and movement time was defined by the laws of mechanics (Juras et al. 2009). These results corroborate the idea that the scaling of movement time with task difficulty originates at the level of movement planning (also see Bradi et al. 2009), that is, it involves cognitive processes.

In our experiments, the same functional form $T = a + k \cdot \ln(ID)$, where T is a timing index, described the relationships between a variety of timing indices and ID across three languages (two in this study plus one more in an earlier study, Latash and Mikaelian 2011). The universal functional form of this relationship across language and motor tasks corroborates the hypothesis on a common, likely cognitive, origin of Fitts' law.

Despite the universal functional form, $T = a + k \cdot \ln(ID)$, which explained a large amount of variance in the data, commonly over 95%, individual participants differed in the coefficients a and k . These differences were, however, modest across speakers of the same language, but they could be rather dramatic between the languages. Indeed, the regression coefficient k for the speech time was 44% larger for the English speakers as compared with the Chinese speakers. The differences in the intercept (a) were smaller in magnitude but still significant and in the same direction: The Chinese speakers had the smallest a values. In a previous study (Latash and Mikaelian 2011), the results for k were similar while there were no significant differences in a between the Chinese and Indo-European speakers. This may be due to the fact that the current study used a more homogenous group of English speakers (in contrast to the mixed, English–Russian group in the earlier study), and the number of subjects per group was larger.

The observed reaction time (RT) values were within the range reported for picture description tasks (Smith and Wheeldon 1999; Allum and Wheeldon 2007; Marian et al. 2008). As in our previous study, the Chinese speakers showed significantly shorter RT values (on average, saving over 170 ms or nearly 20% of RT), but no significant scaling of RT with $\ln(ID)$ was observed. The shorter RT in the Chinese speakers may be explained by a hypothesis that the cognitive process of task analysis and phrase construction starts before RT and continues until the end of the utterance. If there are differences in the speed of this process between subject groups, the differences start to accumulate from the initiation of the process such that relatively minor differences can be seen in RT values while stronger differences are accumulated over ST. As of now, however, this hypothesis has not been supported experimentally.

It is common to take a deep breath before uttering a sentence: Breathing pauses occurring before sentences can take about 400–800 ms and show a weak positive correlation between the pause duration and the sentence duration in some persons (Fuchs et al. 2008; note that subjects in that study read sentences, not created them). So, it is possible that a portion of the reaction time in our study was related to inhalation in parallel with cognitive processes, and the differences in the duration of the planned utterances were responsible for the RT differences. We would like to note, however, that our subjects paced the experiment themselves. Therefore, they could take a deep breath prior to pressing the key that triggered the next trial; moreover, they were instructed to inhale before pressing the key to avoid contamination of the sound signal by a deep breath.

Other timing characteristics, such as movement time, total performance time, and speech-movement delay showed strong linear scaling with $\ln(ID)$. Note that performing the task in the Performer-only trials took much less time as compared with the trials involving speech production. Hence, the bottleneck for the Performer in the Speaker–Performer series was not in the speed of his/her actions but primarily in the rate of speech by the Speaker. Indeed, anecdotally, the Performers typically picked up the indicated object and quickly moved it close to the plate at the initiation of the utterance. Then, they waited for the Speaker to describe where the object was to be placed and how it was to be oriented. It is not surprising, therefore, that our movement time (MT) and total time (TT) indices showed scaling with ID similar to that of ST.

The way we computed ID could be criticized as somewhat arbitrary and based on a mixture of information theory and linguistic considerations. As explained earlier (Methods), we computed it as the product of ID components related to the object, its location, and its orientation. Each of these components was computed as the total number of degrees-of-freedom the subject was expected to describe. From the information theory, the amount of information transmitted when a person names one of the four simple locations (above, below, to the left, and to the right) is the same as the amount of information transmitted when one of the four corners is named. Linguistically, however, naming a corner requires naming two degrees-of-freedom, its left–right coordinate, and its up–down coordinate; hence, we assigned higher ID values to the corner locations.

Our Performer-only series has provided indirect support for the method of ID computation. Indeed, when required to perform only the motor part of the task, the subjects showed a classical Fitts’-like linear scaling of movement time with $\ln(ID)$. So, with respect to movement time, our method of ID computation led to qualitatively the same results as could be expected from the classical method based on the ratio of target distance to target width. Note that the regression coefficients and correlation coefficients within the linear regressions in the Performer-only series were smaller than the corresponding coefficients of the regressions that linked ST to $\ln(ID)$ in the Speaker–Performer series (this was true for all 16 subjects, and hence this finding is significant at $P < 0.01$ according to the sign test). This may be seen as providing further indirect support for our method of estimating ID.

Constructing phrases as quickly as possible $\neq$ speaking as fast as possible

In the classical studies of quick speech production by the group of Sternberg (Sternberg et al. 1978, 1980, 1988), a somewhat different question was asked: How long does it take a person to pronounce an utterance consisting of ‘ n ’ elements (words, numerals, or non-words) with each element consisting of ‘ m ’ syllables? The text had been presented to the subject in advance. As a result, the subjects did not need to construct a phrase but rather to utter the elements as quickly as possible after an imperative signal. In those experiments, there was a close to quadratic increase in the speech time with the number of elements and

the number of syllables per element. As a result, average speech rate increased for longer samples.

In our study, the subjects spoke much slower than in the mentioned studies by Sternberg's group, with speech times 3–4 times longer for utterances with comparable total numbers of syllables. In contrast to Sternberg's studies, there was little effect of the changes in speech time (induced by ID changes) on the speech rate quantified as the number of syllables per utterance. The only significant effect was observed for the shortest utterances, but it was opposite to what Sternberg described: Shorter utterances in our study were associated with faster speech rate. So, we conclude that the mechanisms that define speech production under the "speak as fast as possible" instruction are qualitatively different from those that act under the instruction "construct a phrase as fast as possible".

The significantly faster speech rates for the shortest utterances could be related to the fact that the simplicity of such pictures brought it closer to the "speak as fast as possible" instruction, such that (nearly) the whole phrase was created during the reaction time delay and then uttered at a high rate. Indeed, the speech rate of 150 ms/syllable was closer to the speech rates observed by Sternberg's group (although still lower by a factor of 2).

Speaking faster ≠ doing faster

To test one of the main hypotheses, we compared the performance of our subjects in two-participant (Speaker–Performer) and one-participant (Performer-only) tasks. The logic was simple: If the presumed similarity between the presented pictures and Chinese characters gave the Chinese-speaking participants advantage, its effects could be expected in both types of tasks. The result was directly opposite to that expected based on this hypothesis: In the Performer-only task, the Chinese speakers were significantly slower as compared with the English speakers. So we conclude that, even if the pictures resembled Chinese characters, this factor by itself did not give the Chinese speakers advantage that could lead to lower speech time values.

Why were the Chinese speakers slower in the Performer-only series? We have only a speculative answer. This result may reflect a cultural difference between students in a major US University who came from China and those who were raised and studied in the USA. The Chinese students were trying to be, first and foremost, accurate (and fast, if this was compatible with being accurate), while their American counterparts seemed to be trying to be, first and foremost, fast (even if this led to a somewhat sloppy performance). Data on the number of errors confirm this interpretation: The Chinese students showed fewer errors, although the group difference did not reach significance because of the large inter-subject variability.

Why did Chinese speakers use fewer syllables?

The results of our study show that the main factor that led to lower speech time values (and was also reflected in other timing indices such as speech-movement delay and performance time) in the Chinese speakers was the fewer syllables they used as compared with the English speakers. As a result, their speech time was proportionally shorter, while speech rates were similar in the two groups.

Several studies reported an increase in speech duration and a decrease in speech rate occurring at grammatical and prosodic junctures (among others Byrd 2000; Byrd and Saltzman 2003; Krivokapić 2007). Thus, the differences in speech time between the Chinese and English speakers might be a result of their sentence, and more generally, text structure.

On the scale analytic/synthetic, English occupies the most analytic position of all the Indo-European languages, but it is far less analytic than Chinese. English is a highly structured language with almost every sentence requiring a subject and a verb, while spoken Chinese is much more forgiving. For example, Chinese syntax can tolerate the omission of not only a single subject in a sentence, but also several subjects in a group of clauses or even sentences as long as these clauses or sentences are within the same topic chain. It is not unusual to see a single subject that takes several verb phrases without using any conjunctions in a sentence. These similar particularities of Chinese sentence construction might have contributed to the reduced number of syllables and, therefore, shorter speech time to describe the same sets of pictures for the Chinese speakers as compared with the English ones.

Whether using tones by itself allows to express the same number of meanings with fewer syllables (Brown-Schmidt and Canseco-Gonzalez 2004; Valaki et al. 2004; Lee 2007) remains a highly speculative interpretation. We plan to address it using subjects who speak other tone languages such as Cantonese Chinese (6+ tones) and Vietnamese (6 tones). It would be also interesting to test our hypothesis on an isolating language lacking tones, such as Khmer, but finding enough participants may be challenging.

Informational density of natural languages

Panels a and b of Fig. 10 show the number of syllables/utterance (N_{SYL}) as a function of $\ln(ID)$. These graphs show that the same amount of phonetic material can transmit significantly more information in Chinese than in English in the simple picture description tasks used in this experiment. For example, on average, 15 syllables in English can describe a picture with $\ln(ID) = 1$, while in Chinese, they allow to perform a task with $\ln(ID) = 2.5$. So, the results point at an important characteristic of natural languages that make them different in the ability to transmit simple information. This characteristic can be called “informational density”.

Since different languages emerged in the process of evolution, they have to have a balanced combination of relative advantages and disadvantages. Our current study has demonstrated an advantage of Chinese in the time required to describe simple pictures that require object location/orientation descriptions in the two languages, which might or might not generalize to other types of description/discourse. In particular, we do not know whether this advantage persists or disappears with an increase in picture complexity and introduction of subtleties. This is an exciting problem to study in future. Addressing this problem would require the analysis of much more complex corpora involving a larger variety of levels and kinds of cognitive and linguistic complexities as well as a broader range of languages.

Let us conclude by formulating a couple of hypotheses based on the current results that we hope to test in future studies. (1) Languages at the synthetic end on the analytic/synthetic scale require more time to transmit comparable amounts of simple information as compared with languages at the analytical end on this scale. (2) An advantage of one natural language over another in the time required to transmit simple information disappears (and may even turn into a disadvantage) for more complex tasks.

References

- Allum PH, Wheeldon LR. Planning scope in spoken sentence production: the role of grammatical units. *J Exp Psychol: Learn, Mem, Cognit.* 2007; 33:791–810. [PubMed: 17576154]
- Bradi AC, Adam JJ, Fischer MH, Pratt J. Modulating Fitts’ law: the effect of disappearing allocentric information. *Exp Brain Res.* 2009; 194:571–576. [PubMed: 19241067]

- Brown-Schmidt S, Canseco-Gonzalez E. Who do you love, your mother or your horse? An event-related brain potential analysis of tone processing in Mandarin Chinese. *J Psychol Res.* 2004; 33:103–135.
- Byrd D. Articulatory vowel lengthening and coordination at phrasal junctures. *Phonetica.* 2000; 57:3–16. [PubMed: 10867568]
- Byrd D, Saltzman E. The elastic phrase: modeling the dynamics of boundary-adjacent lengthening. *J Phone.* 2003; 31:149–180.
- Duarte M, Latash ML. Effects of postural task requirements on the speed-accuracy trade-off. *Exp Brain Res.* 2007; 180:457–467. [PubMed: 17273871]
- Fitts PM. The information capacity of the human motor system in controlling the amplitude of movement. *J Exp Psychol.* 1954; 47:381–391. [PubMed: 13174710]
- Fitts PM, Peterson JR. Information capacity of discrete motor responses. *J Exp Psychol.* 1964; 67:103–112. [PubMed: 14114905]
- Fuchs, S.; Hoole, P.; Vornwald, D.; Gwinner, A.; Velkov, H.; Krivokapic, J. In: Sock, R.; Fuchs, S.; Laprie, Y., editors. The control of speech breathing in relation to the upcoming sentence; Proceedings of the 8th international seminar on speech production; December 8–12, 2008; Strasbourg, France. 2008. p. 77-80.
- Gutman SR, Latash ML, Gottlieb GL, Almeida GL. Kinematic description of variability of fast movements: analytical and experimental approaches. *Biol Cybern.* 1993; 69:485–492. [PubMed: 8274547]
- Juras G, Slomka K, Latash ML. Violations of Fitts' law in a ballistic task. *J Mot Behav.* 2009; 41:525–528. [PubMed: 19564148]
- Krivokapić J. Prosodic planning: effects of phrasal length and complexity on pause duration. *J Phone.* 2007; 35:162–179.
- Latash ML, Mikaelian IL. How long does it take to describe what one sees? The first step using picture description tasks. *Hum Move Sci.* 2010; 29:369–385.
- Latash ML, Mikaelian IL. A logarithmic speed-difficulty tradeoff in speech production. *Mot Control.* 2011; 15:52–67.
- Lee CY. Does horse activate mother? Processing lexical tone in form priming. *Lang Speech.* 2007; 50:101–123. [PubMed: 17518105]
- Marian V, Blumenfeld HK, Boukrina OV. Sensitivity to phonological similarity within and across languages. *J Psychol Res.* 2008; 37:141–170.
- Meyer DE, Smith JE, Wright CE. Models for the speed and accuracy of aimed movements. *Psychol Rev.* 1982; 89:449–482. [PubMed: 7178330]
- Plamondon R, Alimi AM. Speed/accuracy trade-offs in target-directed movements. *Behav Brain Sci.* 1997; 20:1–31. [PubMed: 10096994]
- Smith M, Wheeldon L. High level processing scope in spoken sentence production. *Cognition.* 1999; 73:205–246. [PubMed: 10585515]
- Sternberg, S.; Monsell, S.; Knoll, RL.; Wright, CE. The latency and duration of rapid movement sequences: comparisons of speech and typing. In: Stelmach, GE., editor. Information processing in motor control and learning. Academic Press; NY: 1978. p. 117-152.
- Sternberg, S.; Wright, CE.; Knoll, RL.; Monsell, S. Motor programs in rapid speech: additional evidence. In: Cole, RA., editor. Perception and production of fluent speech. Lawrence Erlbaum Associates; Hillsdale: 1980. p. 507-534.
- Sternberg S, Knoll RL, Monsell S, Wright CE. Motor programs and hierarchical organization in the control of rapid speech. *Phonetica.* 1988; 45:175–197.
- Valaki CE, Maestu F, Simos PG, Zhang W, Fernandez A, Amo CM, Ortiz TM, Papanicolaou AC. Cortical organization for receptive language functions in Chinese, English, and Spanish: a cross-linguistic MEG study. *Neuropsychologia.* 2004; 42:967–979. [PubMed: 14998711]

Fig. 1.

A photograph of the setup with two participants. The Speaker watches the screen that shows pictures with pairs of objects, the plate and one more object. The Performer is sitting across the table being ready to move the named object to a proper position and orientation

Fig. 2.

Examples of the pictures with different objects. The values of ID related to the object (ID_{OB}), location (ID_{LO}), and orientation (ID_{OR}) are presented with the total ID computed as $ID = ID_{OB} \cdot ID_{LO} \cdot ID_{OR}$

Fig. 3.

Grand average values of the timing indices with standard deviation *bars* computed across the subjects are shown for the Chinese speakers (*hatched bars*) and English speakers (*solid bars*). **a** Reaction time (*RT*) and speech-movement delay (Δ_{SM}). **b** Speech time (*ST*), movement time (*MT*), and total performance time (*TT*)

Fig. 4. Reaction time (averages across subjects with standard deviation *bars*) as a function of the log-transformed index of difficulty, $\ln(ID)$, is shown for the Chinese (*open circles*) and English (*filled circles*) speakers. Regression lines and equations are presented

Fig. 5.

Speech time as a function of the log-transformed index of difficulty, $\ln(ID)$, for the Chinese (**a**) and English (**b**) speakers. Individual subject data are shown for both groups with linear regression lines and equations. Averaged across subjects data with standard deviation *bars* are shown in panels **a** and **b** with larger *square symbols*. Corresponding regression lines are shown with *thicker lines*, and the equations are shown in a *larger font*. Note the high coefficients of correlation and low regression coefficients in the Chinese group as compared with the English group

Fig. 6. Speech-movement delay, Δ_{SM} , (averages across subjects with standard deviation *bars*) as a function of the log-transformed index of difficulty, $\ln(ID)$, is shown for the Chinese (*open circles*) and English (*filled circles*) speakers

Fig. 7.

Movement time as a function of the log-transformed index of difficulty, $\ln(ID)$, for the Chinese (**a**) and English (**b**) speakers. Individual subject data are shown for the two groups with linear regression lines and equations. Averaged across subjects data with standard deviation bars are shown in panels **a** and **b** with larger *square symbols*. Corresponding regression lines are shown with *thicker lines*, and the equations are shown in a *larger font*. Note the high coefficients of correlation and lower regression coefficients in the Chinese group as compared with the English group

Fig. 8.

Total performance time as a function of the log-transformed index of difficulty, $\ln(ID)$, for the Chinese (a) and English (b) speakers. Individual subject data are shown for the two groups with linear regression lines and equations. Averaged across subjects data with standard deviation bars are shown in panels a and b with larger square symbols.

Corresponding regression lines are shown with thicker lines and the equations are shown in a larger font. Note the high coefficients of correlation and low regression coefficients in the Chinese group as compared with the English group

Fig. 9.

Total performance time in the “Performer-only” series as a function of the log-transformed index of difficulty, $\ln(ID)$, for the Chinese (**a**) and English (**b**) speakers. Individual subject data are shown with linear regression lines and equations. Averaged across subjects data with standard deviation *bars* are shown with *larger square symbols*. Corresponding regression lines are shown with *thicker lines*, and the equations are shown in a *larger font*. Note the high coefficients of correlation and higher regression coefficients in the Chinese speakers as compared with the English speakers

Fig. 10.

Number of syllables (N_{SYL}) as a function of the log-transformed index of difficulty, $\ln(ID)$, for the Chinese (a) and English (b) speakers. Individual subject data are shown with linear regression lines and equations. Averaged across subjects data with standard deviation bars are shown with *larger square symbols*.

Corresponding regression lines are shown with *thicker lines*, and the equations are shown in a *larger font*. Note the high coefficients of correlation and lower regression coefficients in the Chinese group as compared with the English group. Speech rate for the Chinese (c) and English (d) speakers. Individual subject data points are shown. Averaged across subjects data with standard deviation bars are shown with *larger square symbols*. The data points are interpolated