Bulletin of Tokyo Gakugei University Sect. II, 41, pp. 181~189, 1990

Speech Perception : Context and Environment

Tetsuyu NODA

Department of English Teaching

(*Received for Publication : September 30, 1989*)

Children have a remarkable ability to acquire the first language in a very short time and acquire the rhythm or the "feel" or "sound" of the language and tell their native language from the other languages. On the other hand, even adult native speakers who are supposed to have acquired the language often make mistakes when they speak. They have the competence, the knowledge of the language, but when they speak, they may occasionally make mistakes or falter or hesitate. When they write or speak, they make mistakes. When they "edit" or "monitor" what they have done, they notice their mistakes and correct them. Sometimes they do not notice. People often understand the intended meanings even if the mistakes are not corrected by the writer or the speaker. This does not mean that they do not know the language but shows that they do know the language.

People do not just listen to every sound or word when they speak in daily conversations. Whether on the street, or in a bar, or at a party, there is usually noise, sometimes a lot of noise. But native speakers can usually understand without much difficulty. They can communicate in spite of noise. Why do people understand and communicate even in bad conditions? There must be some cues in the context in which the communication takes place. Language cannot take place in a vacuum. There is some kind of cue for a person in a noisy environment to be able to communicate. Even if they cannot hear every phoneme, sound, word, phrase, or sentence, they may be predicting what will be said or hear what is not actually "uttered." Sometimes they have illusions that they are hearing some sounds which are not really present, or sometimes they mishear in the way the speaker intends. They may have an ability to fill in the "gap" in what is not clearly heard. Not only the linguistic or phonetic, but contextual cues may help our understanding.

In the following sections we are going to see the importance of context in understanding speech, how important it is, and how it affects understanding.

Warren (1970) studied the phnomenon they called "phonemic restoration." It was found that replacement of a phoneme in a recorded sentence by a cough resulted in illusory perception of the missing sound. They recorded the sentence "The state government met with their respective legislatures convening in the capital city." Then they cut out of the tape recording of the sentence one phoneme, or speech sound, "s" in "legislatures." They

also cut out the preceding and following phonemes, and then they spliced the recorded sound of a cough of the same duration. The result was that the missing sound was heard by the subjects as clearly as were any of the phonemes that were physically present. Even after being told that a sound was missing, the subjects could not distinguish the illusory from the real one. The "gis" in "legislatures" was heard clearly even when it was replaced by an extraneous sound. No subject identified correctly the position of the cough or the position of the tone when the cough was replaced with the tone. But when a speech sound was deleted and not replaced with a sound, the gap was recognized in its proper location, and illusory perception of the missing sound did not occur. The redundancy of language and the context supply the missing "s" in "legislatures."

Warren and Obusek (1971) confirmed their preliminary study (1970). The listeners restore the missing sound on the basis of both prior and subequent context. This illusory effect causes the physically absent phoneme to seem as the speech sounds which are present. The nonspeech sound is perceived as coexisting with, but not interfering with, the intelligibility of phoneme which is actually present. They got similar results even when three different types of sound (tones, buzzes, and coughs) were employed. People can comprehend speech even when portions are obliterated, and they can maintain intelligibility using the redundancy provided by the context. They maintain, "PhR (phonemic restoration) may represent an essential mechanism leading to such comprehension—so that PhR is not only an illusion, but a well-practiced skill aiding in extraction of meaning from discourse heard under the noisy conditions of everyday life" (261). This is part of the answer to the question why we can understand what a speaker is saying in a noisy situation. Even though we cannot hear every sound or word, we can still comprehend aided by the redundancy and context.

In the sentence "There was time to (?) ave ____." a cough is followed directly by the word fragment "ave." On the basis of the prior context, the missing fragment could be "shave," "save," "wave," "rave," etc. If subsequent portion of the sentence refers to departing friends, the PhR would be /w/. The experiments by Sherman have shown that "listeners can store auditory information until subsequent context identifies the phoneme obliterated by an extraneous sound and then "hear" the missing phoneme as a PhR" (361). Thus it is presumed that listeners identify the sounds or restore the missing sounds as present and comprehend after the context is made clear.

Warren and Warren (1970) found that people rely on syntactic and semantic constraints far beyond the missing segment itself. The subjects were presented with the sentence with a cough (*), "It was found that the *eel was on the ____." The words that could complete the sentence are "axle," "shoe," "orange," and "table." They heard "*eel" as "wheel," "heel," "peel," and "meal." People experience the appropriate phonemic restoration. It appears that they store the imcomplete information until the necessary context is supplied so that the required phoneme can be synthesized. In connection with the use of the context for correcting errors, they cite Miller's reasoning that a mistake once made while listening to spoken discourse would cause errors in interpreting the following portions of the message to pile up. They argue that storage of incoming language information is associated

with error correction. They also refer to Bryant and Harter (1890) that highly skilled telegraphers listening to Morse code did not translate the auditory signals that constitute a word until some six to 12 words after the signals were heard. Warren and Warren argue, "Verbal context. . . can determine completely the synthesis of illusory speech sounds; phonemic restorations are heard when the context is clear but part of the stimulus is absent. Another illusion arises when the stimulus is clear but the context is absent" (35).

Warren and Warren (1970) also show another illusion when the stimulus is clear but the context is absent. The subjects listened to a clear recording of a word or phrase repeated over and over again, and illusory changes occurred in what the voice seemed to be saying. The subjects listened to "tress," repeated without pause, and illusory forms "dress," "stress," "Joyce," "floris," "florist," and "purse" appeared. They call this finding the "verbal transformation effect." At different ages, the subjects experienced different transformations. Children at age five experienced either very few or no verbal transformations. At six half they heard illusory changes, and by age eight, all the children heard verbal transformations. From their observation with verbal transformations, they state, "As people grow older they employ different perceptual mechanisms appropriate to their familiarity with language and their functional capacities, both of which change with age" (36). They believe that the use of verbal context has much to do with the age differences in the frequency and nature of verbal transformations. "The absence of illusory changes at age five suggests that young children have not yet reached the stage in language development where storage with skilled reorganization comes into play" (36). They conclude that phonemic restorations and verbal transformations provide new techniques for studying the perceptual organization of heard speech, the grouping of speech sounds, the correction of the listener's errors and the resolution of acoustic ambiguities.

Obusek and Warren (1973) examined the relationship between illusory changes of repeated words (verbal transformations or VTs) and illusory presence of phonemes replaced by noise (phonemic restorations or PhRs). Separate groups of subjects were each presented with one of four variations of the repeated stimulus word "magistrate": stimulus intact (MAGISTRATE); speech sound "s" removed and replaced with a silent gap (MAGI TRATE); speech sound "s" removed and replaced with a louder extraneous sound (MAGI* TRATE); syllable "gis" removed and replaced with a louder extraneous sound (MA*** TRATE). All subjects believed that they were hearing the intact word, and none of them hearing MAGI*TRATE or MA***TRATE believed that a portion of the stimulus was replaced. All subjects hearing MAGI TRATE noticed the pause where "s" should have been. It appears that non-speech sounds replaced by louder sounds can also be perceptually restored on the basis of context. They suggest that both PhRs and VTs are related directly to perceptual processes employed normally for the correction of errors and resolution of ambiguities in speech.

Cherry and Wiley (1967) reported an experiment involving the intelligibility of speech with many gaps. These gaps were either left silent or filled with white noise. They tested how well selections read from various forms of literature such as newspapers and technical matter were understood. Intelligibility increased markedly when white noise rather than

silence filled the intervals between the fragments of speech sounds. Their experiment stresses the importance of the temporal patterning of speech to perception. They suggest investigation of conversation in similar conditions.

Cole (1973) had his subjects listen to a passage and indicate, as quickly as possible, whenever they heard a mispronunciatoin. Mispronunciation included one consonant change in a three-syllable word by changing one, two, or four distinctive features (e.g. "pusily," "visily," or "sizily" for "busily"). Mispronunciations in the first syllable involves the first phoneme in the syllable (e.g. *s*uggested-*z*uggested), and in the second and third syllable (messe*n*ger-messe*m*ger; intordu*c*e-introdu*sh*). Although mispronunciations involving a single feature change were seldom detected, two and four feature changes were readily detected. Reaction times to mispronounced words were slower when they occurred in the first syllable of the word than in the second or third syllable. The subjects need more information than is provided by the first syllable in a word, and must identify an entire word before they can identify a mispronounciation. It is supposed that they need to hear the entire word or the context in which the word is used to identfy it and understand the meaning. This suggests that the subjects do not attend to all of the acoustic information that is present. It is assumed that we expect a certain amount of noise, and a word altered by a single feature may fall within the normal limits of acceptability for the word. This may explain that a mispronounced word may go unnoticed by the listener if it is not a severe mispronunciation.

Picket and Pollack (1963) and Pollack and Pickett (1963) examined the intelligibily of excerpts obtained from the fluent stream of speech. In Picket and Pollack, the talker recorded a short text at three rates of continuous utterance : very fast, normal, and very slow. The intelligibility of one or more words removed from fluent speech was studied, and they found that as the duration of speech sample increased its intelligibility increased, but the rate of utterance did not produce large effects on intelligibility. They concluded that any slurring of articulation that occurs in fast utterance is compensated by covering more context, while a slow utterance may cover less context but be articulated more clearly.

Pollack and Picket (1963) examined the intelligibility of conversational speech words from the fluent stream of speech. They found that the average intelligibility of samples increased with the duration of the excised utterance, and is relatively independent of the average rate of speaking. They interpreted that talkers tend to maintain a constant average precision of communication, either by speaking less words or by speaking more words less clearly, or that the slower parts of conversation are slower because of greater information content. They observed that intelligibility of a word in an excerpt of fluent speech depended markedly upon the duration of the excerpt. They found that intelligibility improved as the duration of the excerpt was lengthened by adding successive words. They speculated that the additional contextual restrictions in the larger samples help the improvement. Thus, a listener who heard *Mary had a little lamb* could correctly complete the initial word of the phrase, even though he might have been unable to identify it when it was presented alone as an excerpt. They refer to this type of context as "structural context. " Then they thought that longer excised samples were more intelligible also

because they provided a clearer auditory picture of the speaker's communication act. They refer to the class of such activities as "auditory context. "

Pollack and Pickett (1964) found that the intelligibility of the intial word of a sample improved as more and more successive words were heard, and also that the intelligibility of the intial word of a speech sample increased as the length of the heard sample was increased. The result is that substantial gains in intelligibility were obtained for longer excised samples of speech. Continued improvements were observed with the longer samples, and they assumed that the advantage of longer samples result from the contribution of auditory context.

Marslen-Wilson (1973) examined the process of shadowing which suggests that speech perception involves more than auditory, phonetic, and phonological processes. It was found that people might be anticipating and identifying the speech on the basis of syntax and meaning too. Some errors the shadowers made were appropriate to the syntax and meaning of the original sentences. For example, in the sentence "It was beginning to be light enough so I could see," some inserted "that" after "so." In the sentence "He had heard at the bridge. . .", five subjects replaced "heard at" with "heard that." These examples show that the subjects' output can be constrained by the preceding context up to and including the word immediately before the error. This shows that people are aware of the syntax and meaning of the sentence.

In Marslen-Wilson (1975) the shadowers repeated semantically anomalous sentence— "The new peace terms have been announced. They call for the unconditional *universe* of all the enemy forces," or syntactically anomalous sentences— "He thinks she won't get the letter. He's afraid he forgot to put a stamp on the *already* before he went to post it. " These anomalies were corrected in shadowing. The words that were anomalous for syntactic and semantic reasons were corrected faster than the words that were anomalous for phonogolical reasons, like "tomorrance" for "tomorrow. " This study presents evidence that sentence perception is most plausibly modeled as a fully interactive process. Each word is immediately entered into the processing system at all levels (phonetic, lexical, syntactic, and semantic) of description, and is simultaneously analyzed at all these levels. In Marslen-Wilson and Tyler (1976), the data also support an interactive parallel model, and they conclude that a representation of the input at all available levels is initiated immediately in sentence processing.

Marslen-Wilson (1978) studied the interactions between the bottom-up analyses of the input and different forms of internally generated top-down constraint, using a shadowing task and a mispronunciation detection task. They found that the listener's dependence on bottom-up analyses in the shadowing task varies as a function of the syllable position of the mispronunciation within the word and of the contextual constraints on the word as a whole. Earlier the study (1975) examined the effects of context variables on the frequency with which the listeners restored mispronounced words to their original form (e.g. repeating "compsiny" as "company"). These restorations occurred frequently where the disrupted word was both syntactically and semantically congruent with its sentential context. In this study the subjects were not told that there were mispronounced words, but were tested

for their comprehension of the shadowed passage. They say this situation is more analogous to normal listening. In the shadowing condition, they examined the number of restorations of mispronounced words to their original form or "fluent restorations" (e.g. repeating "travedy" as "tragedy") and "exact repetitions." In the detection condition, they examined the detection miss-rate. In the detection task only syllable position effects were obtained. They propose an active direct access model. They suggest that top-down processing constraints interact directly with bottom-up information to produce the primary lexical interpretation of the acoustic-phonetic input.

Miller and Isard (1963) showed that the grammatical structure and meaningfulness of a sentence affect the perception of individual words in the message. Listeners were asked to shadow grammatical sentences, anomalous sentences, and ungrammatical strings of words all formed from the same words. For example, grammatical sentences : "Accidents kill motorists on the highways. Trains carry passengers across the country. Bears steal honey from the hive. " Ungrammatical strings of words : "Around accidents country honey the shoot. On trains hive elephants the simplify. Across bears eyes work the kill." Responses were scored both for the number of principal words (five per sentence) and for the number of complete sentence. The sentence scores were : 89% of the grammatical sentences, 80% of the anomalous sentences, and only 56% of the ungrammatical strings were repeated exactly. The scores for the principal words were : 98%, 96%, and 88% of the words in each category respectively were heard correctly. This shows that people were more accurate on grammatical sentences, a little less accurate on anomalous sentences, and least accurate on ungrammatical strings. They conclude that both syntactic and semantic rules are normally involved in the perception of sentences.

Lieberman (1963) had speakers read aloud meaningful grammatical sentences at a fast rate. Some of these words contained common maxims and stereotyped phrases (e.g. "A stitch in time saves nine."), and other sentences which were less familiar contained certain test words that occurred in the stereotyped sentences (e.g. "The number that you will hear is nine."). Test words were excised, and listening tests were performed. Apart from the listening tests, indexes of redundancy (redundant or non-redundant) were computed from the percent of correct guesses obtained from the written tests by the readers. When tape recordings were presented without any masking noise, excised words were mostly correctly identified. When noise was added, the redundant words were identified less often than the non-redundant words. When the complete sentences from which the words were excised were presented with noise, they were 100% correctly identified. He thinks that the speakers must have taken more care when they articulated the non-redundant words. Lieberman says, "People are aware of the semantic and grammatical environment of a word when they either speak or listen to a sentence. When a speaker reads a text rapidly he may utter a word with less care when he knows that a listener can identfy the words from the context. The speaker may modify his production of a word in the light of the subsequent context of the sentence" (184-185). The results of the listening tests show that the intelligibility of the excised words is inversely proportional to the redundancy index. He says, "Isolated words and even phrases are frequently unintelligible though the conver-

sation as a whole is perfectly intelligible."

Miller, Heise, and Lichen (1951) showed that a word is harder to understand if it is heard in isolation than it is heard in a sentence. They illustrate this by reading sentences containing five key words, and scored the listener's responses as the percentage of these key words that were heard correctly. For comparison, these key words were extracted from the sentences, scrambled, and read in isolation. They showed that words heard in sentences were identified more accurately.

Salasoo and Pisoni (1985) argue that both acoustic-phonetic information and syntactic contextual knowledge interacted to generate the set of hypothesized word candidates used in identification. They say that word identification in sentences is an interactive process that makes use of several knowledge sources. It is also documented that both acoustic-phonetic information from the speech signal and other sensory sources of knowledge contribute to spoken word identification (Cole and Rudnicky, 1983). They think that the listener's knowledge of morphology, syntax, and semantics may be labeled *context*. It appears that the listener uses all the available information in both stages of spoken word identification.

Lieberman (1967) suggests that the listener may make use of his knowledge of the articulatory constraints and employ "analysis by synthesis." He uses his knowledge of the syntactic and semantic constraints of the language and the total social context of the message. He goes on to say, "The listener apparently comprehends the message by a process of "hyphothesis formation" that involves analysis-by-synthesis where the context guides the recognition routine. The listener may consider a comparatively large "chunk" of speech, and he is often able to "guess" what the speech signal should be from the context that is furnished by the chunk. The speaker, in turn, simplifies his articulatory control problem, knowing when the listener probably will be able to guess what the message should have been. The speaker may neglect to manipulate his articulatory apparatus precisely when he believes that the listener will be able to guess what he should have said from the context of the message" (163). This has been shown by Miller and Isard (1963). Lieberman also mentions that a speaker may talk more distinctly to a foreign student believing that the listener is unfamiliar with the grammar of the launguage or that a husband may talk less distinctly to his wife believing that she is familiar with his dialect.

People often mihear or misunderstand in everyday conversations. Clark and Clark (1977) refer to Garnes and Bond (1975) and quote data from casual speech:

Original	*Misperception*
wrapping service	wrecking service
meet Mr. Anderson	meet Mr. Edison
I'm covered with chalk dust	I'm covered with chocolate
get some sealing tape	get some ceiling paint

Most of the misperceptions are similar in sound to the original, but some are misperceived by meaning. Clark and Clark say that these misperceptions are a sort of illusion. But

it often happens that we "restore" the "slips of the tongue" in the meaning the speaker originally intended. Sometimes the speaker corrects his mistake, and sometimes he doesn't when he doesn't notice it. Sometimes the listener can supply the correct word or phrase intended by the spaker.

Smith (1973) shows detailed data of the child's acquisition of phonology. In the early stages, the child can not always make detailed distinction in voiced or voiceless sounds, for example, /z/or/d/, /g/, /b/ in word final positions. These sounds are somtimes omitted (e.g. "s" in "always," or "because" or "news" or "noise", or this sound is pronounced as a voiceless sound as /s/. Other voiced sounds are pronounced as voiceless in the early stages (e.g. "d" in "side," "slide," or "hard" is pronounced as /t/ ; "g" in "big," "egg," or "leg" are pronounced as /k/; "b" in "rub," "cobweb," or "cube" is pronounced as/p/. These sounds were gradually acquired.

Brown (1977) gives some examples of elision made by native spakers: "banned for life," "last year," "most recent," "World Wild Life Fund, "discharged prisoners."

Gimson (1980) says that there is a silence of a certain duration in "rubbed gently" after /b/or in "walked back" after/k/, which signals the difference from the present tense "rub gently" or "walk back," or the tense is supplied by the general context. This means these sounds may not be actually pronounced but perceived to be pronounced by the listener, and sometimes only the context helps distinguish the difference.

Gimson also shows that the vowel before a voiced consonant is longer than the vowel before a voiceless consonant (e.g. ladder vs. latter ; eyes vs. ice), and the native speakers can tell the difference by the length of the vowel rather than the voiced or voiceless sound. It is remarkable that the native speakers are doing such a minor but important distinction. These distinctions may seem possible only in a quiet situation, but people are not always speaking in a noiseless, quiet place. There is usually a lot of noise or distraction, but people can understand each other and communicate without much difficulty. People can communicate with the particular conversation partner even in a crowed, noisy room. This is called the "cocktail party phenomenon" or "dinner party phenomenon." It seems interesting to test how native speakers can tell the minor difference in a noisy situation with context and without context, and how they differ from non-native speakers.

References

Brown, G. (1977). *Listening to spoken English.* London: Lomgman.

Cherry, C.,& Wiley, R. (1967). Speech communication in very noisy environments. *Nature*, 214, 1164.

Clark, H. H., & Clark, E. V. (1977). *Psychology and language: An introduction to psycholinguistics*. New York: Harcourt Brace Jovanovich.

Cole, R. A. (1973). Listening for mispronunciations: A measure of what we hear during speech. *Perception & Psychophysics*, 1, 153-156.

Garnes, S., & Bond, Z. S. (1975). Slips of the ear: Errors in perception of casual speech. In *Papers from the Eleventh Regional Meeting, Chicago Linguistic Society*, pp.

214-225.

Gimson, A. C. (1980). *An introduction to the pronunciation of English* (3rd ed.). London: Edward Arnold.

Lieberman, P. (1963). Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech*, 6, 172-187.

Lieberman, P. (1967). *Intonation, perception, and language*. Cambridge, Mass.: MIT Press.

Marslen-Wilson, W. D. (1973). Linguistic structure and speech shadowing at very short latencies. *Nature*, 244, 522-523.

Marslen-Wilson, W. D. (1975). Sentence perception as an interactive parallel process. *Science*, 189, 226-228.

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitice Psychology*, 10, 29-63.

Miller, G. A., & Isard, S. (1963). Some perceptual consequences of linguistic rules. *Journal of Verbal Learning and Verbal Behavior*, 2, 217-228.

Miller, G. A., Heise, G. A., & Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology*, 41, 329-335.

Obusek, C. J., & Warren, R. M. (1973). Relation of the verbal transformation and the phonemic restoration effects. *Cognitive Psychology*, 5, 97-107.

Pickett, J. M., & Pollack, I. (1963). Intelligibility of excerpts from fluent speech: Effects of rate of utterance and duration of excerpt. *Language and Speech*, 6, 151-164.

Pollack, I., & Pickett, J. M. (1963). The intelligibility of excerpts from conversation. *Language and Speech*, 6, 165-171.

Pollack, I., & Pickett, J. M. (1964). Intelligibility of excerpts from fluent speech: Auditory vs. structural context. *Journal of Verbal Learning and Verbal Behavior*, 3, 79-84.

Salasoo, A., & Pisoni, D. B. (1985). Interaction of knowledge sources in spoken word identification. *Journal of Memory and Language*, 24 (2), 210-231.

Smith, N. V. (1973). *The acquisition of phonology: A case study*. New York: Cambridge University Press.

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science*, 167, 392-393.

Warren, R. M., & Warren, R. P. (1970). Auditory illusions and confusions. *Scientific American*, 223, 30-36.

Warren, R. M., & Obusek, C. J. (1971). Speech perception and phonemic restorations. *Perception & Psychophysics*, 1, 358-362.