Real Time Utterance Production and Disfluencies

Takehiko Maruyama
Senshu University / NINJAL

1 Introduction

In speech, the production of utterance is largely affected by "real time constraints." A real time constraint is one that is imposed on the speaker in that the speaker "must conduct the linear and real time production of linguistic form that has well-formed syntactic structure" (cf. Levelt 1989). Spontaneous speech, in particular, differs from reading a previously prepared script aloud because one must organize one's thoughts on the spot and continuously and dynamically form utterances. Often, the fluency of utterances can be lost. In such situations, various disfluencies emerge, such as long pause, filled pause, pronunciation error, self-repair, cut-off, inversion, and insertion.

Such disfluencies inevitably occur in real utterance production. However, in actuality, even if a disfluency is included in a certain utterance, in most cases, this does not cause the listener to have difficulties in understanding. This fact suggests the following.

- In the process of utterance production, for the speaker, there exists a meta-linguistic strategy for the purpose of managing disfluencies; revising in real time, if necessary; and devising ways to show consideration for the listener. Such a strategy is shared among the participants.

As long as disfluencies inevitably emerge as a result of real time constraints, both the speaker and the listener consider their emergence a precondition, and it can be thought that they manage speech by preparing a strategy to adequately process these disfluencies. If certain regularities and patterns can be recognized in such a strategy, this composes an important part for the description of the "speech grammar." In other words, by categorizing disfluencies into several types and analyzing/describing their morphological/functional characteristics, this can contribute to the clarification of a mechanism that supports the management of speech communication.

The following sections show examples of disfluencies and discuss their types and the factors that cause them.

2 Defining Disfluency

The definitions of the term "disfluency" and the scope of the phenomena that it refers to differ among researchers. In this paper, disfluency is defined as "a linguistic phenomenon that occurs through the obstruction of a certain stage of the utterance production process because of any factor and also the linguistic phenomenon that occurs as an act to revise this."

A fluent utterance refers to the production of an utterance without faltering, having a well-formed linguistic form, during which the speaker does not pause or misspeak. In contrast, disfluency refers to the phenomenon that obstructs the flow of fluent utterance, such as when the speaker becomes silent instead of producing an utterance, mispronounces a word, utters a word that is different from what he/she intended to say, or confuses the word order. Furthermore, because the production of utterance for the purpose of revising these difficulties derails the original fluent utterance, it is included in the scope of disfluency. The following phenomena are typical examples of disfluency: pause, filled pause, elongation, mispronunciation, word cut-off, cut-off, false start, repair, repetition, addition, deletion, inversion.

3 Utterance Production Process and the Occurrence of Disfluency

Let us look at (1) as an example of utterance data that include disfluency. This example is taken from CSJ (Corpus of Spontaneous Japanese), which consists of 651 hours and 7.52 million words of spontaneous speech. Tags (F **) corresponds to a filled pause, (D **) to a word cut-off, ":" to an elongation, "/" to a pause less than 0.8 seconds, and bracketed number shows a length of pause more than 0.8 seconds.

(1) (F *e:tto desune*) *taijoo hoosin toyuu no wa* / (F *ano:*) (1.23) *mizuboosoo no uirusu ni yotte okoru byooki de* / (F *e:*) / *taigai tiisai koro ni mizuboosoo o yatta hito wa kanarazu* (F *ano*) *karada no naka ni* / *karada no naka*

no dokoka ni / mizuboosoo no uirusu toyuu no ga nokotte ite / de: (F *e*) *sutoresu: da toka ato: sugoi kyokudo no tukare toka ni yotte / totuzen* (F *sono:*) *mizuboosoo no uirusu ga mata abaredasite /* (F *ano:*) */ hatubyoo suru toyuu mono na n desu ga* (0.82) *mizuboosoo no toki ni dekiru* (F *ano:*) *suihoo / o tomonatta butubutu* (0.86) *jinmasin mitai na butubutu ga /* (F *e*) *totuzen dekiru n desu keredomo /* (F *e:*) */ watasi no baai wa desu ne: /* (F *e:*) / (F *ano*) */ kubi no usiro no sekitui / ni /* (F *e*) *uirusu ga nokotte ite / tyoodo kubi no / kono* (0.95) *bubun ni* (2.15) *koko kara /* (F *e:*) (D *na*) *kubi mune no ue kara kao no kao made no aida ni deru / toyuu koto de /* (F *e:*) */ karada no: desu ne daitai katagawa / no kagirareta bubun ni /* (F *e:*) (0.9) *sono hossin ga / deru / byooki desu* (S00F0210)

Well, shingles is an illness that is, uh, caused by the chickenpox virus, and in general, uh, those who experience chickenpox when they are young, um, always have in their body, somewhere in their body, what is called the chickenpox virus, and, well, from stress or because of extreme exhaustion, suddenly, um, the chickenpox virus may break out, and, well, illness can be caused, and the rashes that accompany the blisters, well, that are formed during chickenpox, rashes that look like hives, um, suddenly emerge, well, but in my case, um, well, the virus remained in my spine, behind my neck, um, in this area of my neck, they formed from here to my neck, above my chest, and my face, well, so rashes appear in limited areas of the body, um, generally on one side of the body, well, so the shingles are something like that.

Example (1) is an utterance that explains "shingles." The overall construction of the utterance shows that it is formed by a "multiple clause-chaining structure" that strings together a multiplex of adverbial clauses (cf. Maruyama et al. 2017). When the utterance begins, the speaker should have started it upon assuming a broad structure of "Shingles is an illness that is X." However, by hesitating during the explanation, adding background knowledge, or explaining the specific symptoms, an uninterrupted and very long utterance is formed as a result.

This utterance includes various disfluencies. For example, regarding "*karada no naka ni* (in their body)," the speaker realized immediately after the utterance that there was a need for revision, and right after, the speaker restates this as "*karada no naka no dokoka ni* (somewhere in their body)." Similarly, "*suihoo o tomonatta butubutu* (rashes that accompany the blisters)" is restated as "*jinmasin mitai na butubutu* (rashes that look like hives)." "*Kono bubun ni* (this area)" is replaced by a more detailed explanation of "*koko kara kubi mune no ue kara kao no kao made no aida ni* (from here to my neck, above my chest, and my face)." Moreover, with regard to "(F *e:*) *watasi no baai wa desu ne:* (um, in my case)" or "(F *e:*) *karada no: desu ne* (um, the body)," the interjectory particle of "*desu ne*" occurs together with an elongation and filled pause. These are elements that are uttered to secure time to prepare the content of the utterance that follows.

Each of these disfluencies can be thought of as occurring as a result of being affected by the constraints of the real time utterance production. In other words, even though the speaker must progressively produce linguistic forms in real time, when the speaker's smooth utterance production processing cannot keep up, the speaker buys time using such methods as the filled pause, elongation, pause, and interjectory participle, self-repairing the already uttered content to a more appropriate form or through repetition along with added information. In doing so, the speaker can respond to the occurrence of these difficulties.

Next, let us look at Example (2). In this example, the speaker begins to hesitate right after she started speaking. In doing so, a separate utterance is inserted in the original utterance, and the overall fluency is lost.

(2) *sono* (0.91) *tomatta heya tteyuu no ga:* (1.45) (D *na:*) */ nan joo gurai ka na* (D *ke*) */ nan joo gurai ni naru no ka na* (2.32) (F *n*) *moo /* (F *n:*) (2.25) *hati joo / hati joo wa semai desu ka ne* (D *nan*) *nan daroo tonikaku sugoku hirokutte:* (0.89) (F *eto:*) */ funatabi na noni: / futuu funatabi toka tte / nitoo toka de sika itta koto nai kara: atasi ni totte wa syoogeki datta n desu ga: / moo tonikaku heya ni beddo ga hutatu atte: sofaa mo atte: / de kurozetto toka mo zenbu tuite te barukonii ga tuite ite: / de: sikamo: nanka* (F *sono*) */* (F *ma*) *ohuro toka basu toire mo* (D *be*) *sikkari* (D *n*) */ heya no naka ni aru mitai na / kanji* (1.66) *desita* (S01F0183)

Well, the room that I stayed in, uh, how many mat sizes, about how many mat sizes was it, well, um, it was an eight-mat size, eight-mat might be small, what, what is it, it was very spacious, and um, it was a trip by ship but, normally traveling by ship, I've only done second class, so it was a shock to me, but, well, there were two beds in the room and a sofa, and everything was included like a closet and a balcony, and on top of it all, it was like the bathtub and toilet were also firmly in the room.

The speaker of Example (2) began the utterance to describe the room she stayed in but immediately got stuck with the words and hesitated with regard to the content of her utterance that had already been spoken and her future utterance by adding on-the-spot new content that the speaker had not thought of initially. In the end, the speaker ends the series of explanations with a sentence-final expression of "*desita*." Linguistic phenomena that represent the speaker as being stuck or hesitating are as follows: insertion into the utterance that causes hesitation regarding the content to be stated next or content that was already stated (e.g., "*nan joo gurai ka na* (how many mat sizes)," "*nan joo gurai ni naru no ka na* (about how many sizes was it)," "*hati joo wa semai desu ka ne* (eight-mat might be small)," "*nan daroo* (what is it)"); repetition to repronounce a word that was not fully stated ("(D

nan) *nan daroo* (what, what is it)"); and a false start that stops what was being stated ("*funatabi na noni:* (it was a trip by ship)") to begin a new utterance ("*futuu funatabi toka tte* (traveling by ship is normally)"). As a result, there is a disfluency in the overall utterance. When the utterance began, the content to be uttered was decided (i.e., explanation of the room the speaker stayed in), but the construction of explanation took time, and the speaker had various failures with structural, well-formed utterances while attempting to continue or develop the utterance, and as a result of attempting to somehow conclude her own utterance with a sentence-final expression "*desita*," in the end, an utterance like Example (2) was formed.

Such examples of disfluency cannot be explained without considering the process through which utterances are dynamically constructed and produced over time. It is not possible to explain linguistic data naturally unless it is viewed as something that is constructed in a dynamic process, rather than as static, as in the past.

4 Toward the Design of Speech Grammar

In a monolog, in which one speaker continues to speak, because he/she is obliged to continue to produce utterances for a certain period of time, it is not acceptable for the speaker to make a long pause. In addition, although one can directly monitor the hearer's response in a dialog, in a monolog, the reaction of the listener is extremely difficult to obtain. In other words, while the speaker of the monolog must assume the comprehension status of a listener, they have to summarize the content of their utterance in real time, spontaneously make it into a well-formed linguistic form and continue to produce it fluently without pause. Moreover, when the real time production of utterances does not progress fluently, various disfluencies occur.

Each of the disfluencies that were observed in examples (1) and (2) emerged because of difficulties in the process of utterance production. First, utterance production is a one-time individual behavior that is developed dynamically over time from when the utterance begins; when the utterance begins, the form of the utterance until it ends is not strictly decided. The speaker begins the utterance upon deciding on the general utterance plan, and he/she must construct in real time a linguistic form that has a grammatically well-formed structure and content that completes itself semantically; in addition, he/she must speak about it linearly. In particular, with regard to spontaneous speech, in contrast to written sentences that can be edited over time, the characteristic of spontaneity becomes strong, such as when changing or updating the original utterance plan on the spot or abandoning the original utterance plan to newly commence an utterance that is in line with the new plan depending on the situation. As a result, the flow of the utterance production that is at times fluent is obstructed, and like examples (1) and (2), an utterance that has structural disturbances or several disfluencies is formed.

The important point is that such characteristics of disfluencies must be shared between the speaker and the listener. Because the systematic strategy of disfluencies is shared between the speaker and the listener, even if disfluencies emerge in an utterance, the smooth understanding and transmission of the message are not obstructed. Regarding this point, it can be said that categorizing the several types of disfluencies and describing their morphological and functional characteristics consist a part of "speech grammar," which manages our speech-to-speech communication.

Spontaneous speech research involves picking out various linguistic phenomena that have not been noticed (or even have been neglected) and qualitatively and quantitatively analyzing them. Speech data are difficult to discuss based on the methodology of conventional descriptive grammar because it includes various "errors" and "noise." Nevertheless, by accumulating the observation and analysis of linguistic phenomena that appear in real speech, as well as the descriptions related to the assorted characteristics and functions of linguistic phenomena, it is possible to develop comprehensive descriptive speech grammar that can systematically cover extensive and real linguistic phenomena.

References

Levelt, Willem. J. M. (1989) *Speaking: From intention to articulation*. Cambridge: The MIT Press.

Maruyama, Takehiko. (2013) Descriptive Research on Disfluent Phenomena in Spontaneous Japanese: A Corpus-Based Approach. Doctoral dissertation, ICU.

Maruyama, Takehiko, Stephen W. Horn, Kerri L. Russell and Bjarke Frellesvig. (2017) On the Multiple Clause Linkage Structure of Japanese: A Corpus-based Study, *New Steps in Japanese Studies (Ca' Foscari Japanese Studies 5)*, 131-154. Venice: Edizioni Ca' Foscari. http://doi.org/10.14277/6969-152-2/CFJS-5-7