In the Image of Language

Corpus Construction as Discourse Representation

John W. Du Bois
Linguisics Department, University of California, Santa Barbara
Director, Santa Barbara Corpus of Spoken American English

Kyoto, December 2011

ABSTRACT

Linguists don't often ask themselves what language is, or what it looks like, but perhaps they should. When the time comes to create an image of language, which will be contemplated as an object of scientific study, the question becomes more urgent. What is the object that we trying to represent? And how can we represent it in a way that captures its unique qualities, in a way that best supports our efforts to understand it? The question of representation is two-fold. The selection of materials for inclusion in a corpus is designed to "represent" the language of a certain group of people, or a certain ways of using language by that group. That is, the corpus is expected to be "representative" of the language or variety in question. At the same time, if the corpus materials include audio or video recordings of spoken language in use, there is the additional task of "representing" in writing what is happening on the recordings. We thus find that representation is a critical component of both the design and the transcription of a corpus of language in use.

But the real meaning of a corpus remains elusive. We can start by defining a corpus as a systematic, unified, representative body of observational data on a language. Or more vividly: a corpus is a slice of language. We cut a small slice in an attempt to represent the larger body of a language, whether English or French or Japanese. But this confronts us with the question: what is language itself, the object of our representation? Traditional corpus linguistics views language, and hence the corpus, as a collection of words in structured sequence -- nouns and verbs and other grammatical elements which it studies in order to arrive at generalizations about units, structures, rules, meanings, and frequencies. This is fine as far as it goes. But what if language also includes the meaning of silences, the pragmatic thrust of a conversation, the interactional goals of its participants, the social consequences that propel the talk from one construction and one activity to the next? I will suggest that the best way to take a slice of language is to take a slice of life; and this is what it means to represent a language. The approach to discourse representation that I propose is essentially ethnographic in its intent, seeking to capture the full scope of human life as represented in the way a group of people use language. But if the goal is to allow us to answer the larger questions about how and why speakers use language as they do, I would argue that such a corpus is of value not only for the sociocultural linguist but also for the scholar of linguistic structure -- the grammarian who really wants to understand grammar. This approach to creating a portrait of discourse, an image of language, seeks to put language in a new light, and to open up new questions, including questions about the relation between structure and use. I present examples of audio recordings of conversations and other speech events drawn from the Santa Barbara Corpus of Spoken American English, transcribed according to a method that I have been developing over a number of years. The goal is to show how an ethnographically sensitive approach to representing spoken discourse can create a corpus that has meaning for a diverse audience ranging from grammarians to language teachers to pragmaticists to social scientists.

Abstract

1. Examination of Spoken Dialogues to Enable Realistic Linguistic Study:
The Case of *Doo*-type Multi-unit Questions in *the Corpus of Spontaneous Japanese*

Masanobu Masuda（増田　将伸）Koshien Universit y （甲子園大学）

This presentation argues that examination of spoken dialogues is highly beneficial in enabling realistic linguistic study. The argument is illustrated in three aspects that the study can take into consideration actual usage, temporality and sequentiality, and each of them is exemplified by the analysis of *doo*-type multi-unit questions in *the Corpus of Spontaneous Japanese*.

Actual usage sets the basis of valid study that is free from arbitrary theorization. For instance, though *Kuchoo wa doo desu ka (How about speaking tone?)*, a question uttered in a sequence to ask how a participant have felt in making a speech, may seem understandable, the actual data shows that the addressed participant has much trouble in understanding what is asked and cannot decide what to answer until a clarifying utterance follows. Accumulation of such actual usage will contribute to valid study.

Temporality is a significant factor as well. When we see a multi-unit question *Aayuu niku tte doo nano, ii niku ja nai desho (How is that kind of meat, it isn't good, is it?)*, the second question may look like a mere clarification of what is asked by the first question. When we take temporality into consideration, however, we will notice that the multi-unit question is delivered with many pauses as in (1).

(1) aayuu niku toka tte doo nan- (0.5 sec.) no (0.3 sec.) ii (1.2 sec.) niku (micropause) ja nai desho

It illustrates that the delivery is delayed as much as possible, which would allow the recipient to take over the turn before the second question is delivered. This fact reveals the speaker's hesitation to deliver the second question which may imply a problematic assumption.

Sequentiality provides the precise characterization of utterances. For instance, a question *Are wa juugo-hun-kan no wa tsukare mashita (Did you get tired in the 15-minute interview?)* should be characterized as a persistent attempt, rather than a mere question, to elicit a response that the recipient got tired, to consider that the question is asked after a similar question is denied. By considering the position in a sequence like this, we can acquire the precise characterization of utterances.

2. Investigating spontaneous speech for a cross-linguistic study of interaction: Some empirical evidence in spoken narratives and task-oriented dialogues

Etsuko Yoshida and Midori Tanimura

This presentation aims to investigate the spontaneous speech data and shows that the traditional notion of the sentence in written mode does not automatically apply to utterance units in spontaneous speech. It is clarified, especially, that sentences in dialogues are not always represented by an individual speaker but are constructed as a product of collaborative effort involving more than one participant.

We introduce three types of spontaneous speech data and give the results of analysis on referential choice and clause construction in interactive discourse. The data we analyse are (1) spoken narratives of English and Japanese collected by an experiment based on the film of *the Pear Stories* (Chafe 1980), (2) task-oriented dialogues called English and Japanese (Labelless) Map Task Corpus (MTC), and (3) another task-oriented dialogues based on a Lego block task by a pair of native Japanese speakers, by a pair of native English speakers, and by a pair of Japanese learners of English. In data (1), comparing a corpus of English and Japanese narratives, the choice and the distribution of referring expressions are investigated. In data (2), focusing on the clause constructions observed in exchanges between two participants, discourse entities can be realized by explicit referring expressions rather than by implicit referring expressions. The research also highlights a particular sentence construction in particular context, 'conditional clause': How are conditional clauses used in spontaneous spoken language? Data (3) is constructed for a pedagogical research purpose. The representation of joint attention and the organisation of common ground are investigated.

Our findings in data (1) are that English and Japanese narratives show interesting correlations between the referential choices of discourse entities and local coherence of utterances, but most of the entities are represented by pronouns rather than noun phrases in English, and mainly by bare nouns in Japanese. In dialogue data (2), it is clear that the chains of NPs can contribute to the topic chains as local focus of discourse in both English and Japanese. Comparing with the original MTC, the labelless map task dialogue is more complicated due to the additional task design: naming the landmark. The lack of 'ready-made' written labels on the maps encourages the participants to construct their own descriptions to identify entities of landmarks. This task can require more effort into the participants' cooperation, especially at the initial stage of the dialogue. Furthermore, conditional clauses that stand alone function as instructions or mild orders. This type of instruction implicitly requires back-channels from the interlocutor. In dialogue (3), comparing the different workspace in either 'hidden' or 'visible' conditions, both participants contribute to dialogue processing by using more questions and answers in the hidden condition, and the different pairs tend to use different information structures in giving instruction.

Finally, despite the grammatical differences between the two languages, the ways of discourse development in both data sets show distinctive similarities in the process by which the topic entities are introduced, established, and shifted away to the subsequent topic entities.

A Sequential Analysis of the Utterance Initial *wa* in Japanese Conversation

Maki Shimotani
Kansai Gaidai University

The particle *wa* in Japanese has been well-known as a 'topic' and/or 'contrastive' marker (Kuno 1973; Shibatani 1990, etc.), and it has been extensively investigated in terms of its syntactic, semantic, and discourse-pragmatic characteristics and functions (Clancy and Downing 1987; Hinds 1987; Iwasaki 1987; Maynard 1980, 1987; McGloin 1986, 1987, etc.). In previous studies, however, it has been generally understood that *wa*, as other particles in Japanese, is an essentially postpositional grammatical element that attaches to a host noun or a noun phrase. Thus, researchers have conducted their study based on the presupposition that *wa* occurs in the canonical structure such as [X (NP) *wa* Y (Predicate)], and they have focused on the uses of *wa* directly accompanying an NP.

However, observing spontaneous conversations in recent years, we can actually find cases of *wa* that are detached from a possible host NP in the preceding utterances. Observe the following example from a spontaneous conversation.

1. R: *ichioo* *ima* *seekagaku* *tte* *yuu* *koto* *de.*
   tentatively now biochemistry COM say thing COP
   'Now, I('m) tentatively (majoring) in Biochemistry.'

2. A: *fu:::n °fun fun°*
   'I see.'

3. (0.5)

4. A: *e-* *de-* *shoorai* *doo* *suru* *tsumori* *nan* *desu?* ←
   eh then future how do intend COP-NML COP
   'Well then, what are you planning to do in the future?'

5. R: ***↑WA::****hah* *a::* *nanka (.)* *moshi* *iketara* *kokuren* *toka::,*
   wa (laugh) *um* *something-like* *if* *go-if* *the United Nation* *etc.*
   '***WA****::*(laugh), um if possible, I'd like to work for the United Nation or something.'

6. A: [*fu:::n.*
   'I see.'

7. R: [*soo* *yuu* *nanka* *kokusaikikan* *de:::,*
   such say something-like international organization at
   '(I'd like to work) for such kind of international organizations…'

As in this fragment, in response to speaker A's question (line 4), speaker R initiates her turn with *↑WA::* and provides an answer (line 5), where *wa* is uniquely removed from the canonical structure of [X (NP) *wa* Y (Predicate)] at a sentential level. The usage of this type of *wa* can be typically found in daily conversations with younger generations, but it also appears that the range of its use has been widen recently.

This paper explores this newly emerging usage of *wa* in Japanese conversation and explicates how detachability of Japanese postpositional particles incorporates the thematic and contrastive characteristics of *wa* to achieve a particular kind of interactional work. More specifically, based on the data from naturally occurring conversation, I will examine the sequential patterns and contexts in which the utterance initial *wa* occurs and demonstrate that it typically appears as a second pair part of a question-answer sequence. I will also argue that the utterance initial *wa* serves to bracket the interlocutor's response-soliciting utterance in the immediately preceding turn(s) as a whole, rather than to latch onto a distant possible host NP element, as other postpositional particles do (cf. Hayashi 2004). Further, I will maintain that this type *wa* used in response to a question simultaneously indicates the speaker's interactional stance toward the interlocutor to provide an affiliative response.

Selected References

Clancy, M. P. and Downing, P. (1987) *Wa* as a Cohesion Marker. In Hinds, J., Maynard, S. and Iwasaki, S. (eds.) *Perspectives on Topicalization. The Case of Japanese wa*. 3-56, Philadelphia: John Benjamins Publishing.

Hayashi, M. (2004) Discourse within a sentence: an exploration of postpositions in Japanese as an interactional resource. *Language in Society* 33, 343–376.

Hinds, J. (1987) Thematization, Assumed Fmiliarity, Staging, and Syntactic Binding in Japanese. In Hinds, J., Maynard, S. and Iwasaki, S. (eds.) *Perspectives on Topicalization. The Case of Japanese wa*. 83-106, Philadelphia: John Benjamins Publishing.

Iwasaki, S. (1987) Identifiability, Scope-Setting, and the Particle wa: A study of Japanese Spoken Expository Discourse. In Hinds, J., Maynard, S. and Iwasaki, S. (eds.) *Perspectives on Topicalization. The Case of Japanese wa*. 107-141, Philadelphia: John Benjamins Publishing.

Kuno, S. (1973) The Structure of the Japanese Language. Cambridge: The MIT Press.

Maynard, S. (1980) Discourse functions of the Japanese theme marker *wa*, Dissertation, Northwestern University.

_____. (1987) Thematization as a Staging Device. In Hinds, J., Maynard, S. and Iwasaki, S. (eds.) *Perspectives on Topicalization. The Case of Japanese wa*. 57-82, Philadelphia: John Benjamins Publishing.

McGloin, N. H. (1986) *Negation in Japanese*. Edmonton: Boreal Scholarly Publishers.

_____. (1987) The Role of *Wa* in Negation. In Hinds, J., Maynard, S. and Iwasaki, S. (eds.) *Perspectives on Topicalization. The Case of Japanese wa.* 165-183, Philadelphia: John Benjamins Publishing.

Shibatani, M. (1990) *The languages of Japan*. Cambridge: Cambridge University Press.