Distinct Functions of quotative markers: Evidence from Meidai Kaiwa Corpus

Chad Nilep

1. Introduction

The Japanese *joshi* or particle *to* serves as a quotative marker, either indicating the content of quoted speech or thought, or serving related functions such as indirectly attributing ideas to parties other than the speaker (functions of evidentiality) or distancing the speaker from those ideas (hedging, self-mocking, and the like). The particle *tte* is frequently identified as an informal variant of *to*, serving identical or nearly identical functions in casual speech. Scholars have suggested that the two forms may have different distribution or function (e.g. Jorden 1990, [1962]; Martin 1975; Okamoto 1995; Hamano 1998), but to date there has been little empirical work to distinguish the two forms using broad-based corpus methods. This paper presents an analysis of the Meidai Kaiwa Corpus, a collection of 129 informal conversations collected in ten prefectures throughout Japan. The data show that both *to* and *tte* occur frequently in the conversations. Contrary to the common assertion that *tte* is the variant of *to* used in casual speech, this suggests that both particles occur in informal contexts. Furthermore, the data show a clear but non-categorical tendency for the particle *tte* to appear with verbs of speech, especially *iu* "say", or at the ends of utterances with no following verb, and for the particle *to* to appear with verbs of cognition, especially *omou* "think". This tends to support suggestions by Jorden (1990) and Martin (1975) among others that the former particle indicates speech while the latter indicates the content of thoughts.

2. Data

Data come from the Meidai Kaiwa Corpus (Ohso 2003), a collection of transcripts of spontaneous conversations among previously acquainted native speakers of Japanese. The 129 conversations that make up the corpus total approximately 100 hours and 28 minutes of speech. Speakers range in age from early teens to over 90 years old, though precise ages are not recorded. Each conversation involves two, three,

or four participants, with the largest proportion of conversational groups dyads. Six conversations involve pairs of men, 78 pairs of women, 13 mixed-gender dyads, 17 groups of three women, ten mixed-gender groups of three participants each, one group of four women, and four mixed-gender groups of four participants each.

Seventy-eight of the conversations that make up the Meidai Kaiwa Corpus were recorded in Aichi Prefecture. Other conversations were recorded in Tokyo (n=22), Hokkaido (n=8), Gifu (n=7), Kanagawa (n=5), Shizuoka (n=3), Yamanashi (n=2), Niigata (n=2), Kyoto (n=1), and Osaka (n=1) prefectures.

3.　Methods

The 129 text files that make up the Meidai Kaiwa Corpus were combined into a single file for processing. When the file was scanned for errors, 431 lines out of 325,086 total lines were found to contain corrupted characters and were removed. The resulting corpus was analyzed and tagged for part-of-speech using the ChaSen morphological analyzer (Matsumoto et al. 1999) in KH Coder (Higuchi 2001).

Part-of-speech analysis classified words as nouns, verbs, adjectives, adjectival nouns, adverbs, or 'other'. The last category includes postpositions and all grammatical particles. Classifying all grammatical particles together has the unfortunate effect of conflating several items that have the same form but distinct functions.

In addition to the quotative particle, a particle marking conditional, sometimes translated as "if", and an associative or conjoining particle, often translated "with" or "and", all have the same surface form, *to*. Example 1 illustrates the conditional particle; example 2 illustrates the associative particle.

1. F064：スカート、スカート 楽しい よね。でも 気合　　入って ない と、
       skirt   skirt   fun   PT   but spirit enter NEG ***TO***
    なんか　　スカート はけない。
       somehow skirt   wear-NEG
   "Skirts, skirts are fun, right. But if for some reason I don't like it, I can't wear a skirt."

2. F034：あれ、な、妹さん　　　　と　お母さん と
       that PT young-sister ***TO*** mother ***TO***
    3人　　　　じゃない の。(↓)
       3-people isn't   PT
   "Wait, like, with younger sister and with mother, that's three people, isn't it?"

In example 1, the conditional particle *to* indicates the condition (*kiai haittenai to* "if you don't like it") under which skirts are not worn. In example 2, the associative particle indicates the individuals (*imoto-san to okaasan to* "with younger sister and mother") associated in the group of three people. These three distinct functions are generally recognized as three different particles with the same surface form.

Since all three particles with the form *to* and the morphosyntactic class 'other' were counted together, the most common collocate of *to* was *ka*, as in examples 3-4.

3. F023：ドールコレクション　とか　ある　って。（↓）
       doll collection ***TOKA*** exist TTE
  "There's a doll collection and so on."

4. F048：サンドイッチとか　ケーキとか　も　あるし。（↓）
       sandwich ***TOKA*** cake ***TOKA*** also exist
  "There are also sandwiches and cakes and so on."

While many analyses view the string *toka* as a combination of the associative particle *to* and the interrogative particle *ka*, at least historically (Suzuki 1998; Kinuhata 2012),[1] in contemporary usage *toka* is often treated as a lexical item in its own right (Kuno 1973). In light of this, part-of-speech tagging was re-run forcing identification of the string *toka* as a single word.

No other attempt was made to separate the various forms of *to* with automatic parsing or part-of-speech tagging.[2]

The final tagged corpus database consists of 3,656,677 tokens of 22,178 word types. It features 21,346 tokens of the particle *tte* and 27,827 tokens of *to* (conflating all uses). Using KH Coder, the most frequent collocates of each string were identified.

The second-most common item following the particle *tte* (the R1 collocate, first item to the right) and third-most common following *to* is the full-stop, accounting

---

[1] Additionally, some scholars view *toka* as a grammaticalized quotative marker derived from the quotative *to*. An analysis of ten randomly selected tokens of *toka* from the corpus found only tokens of the associative *to* + *ka*, though a more systematic search might reveal quotatives as well.

[2] Another common collocate of *to*, the verb *ikeru* "do, can do" (see Table 4), is similarly problematic. In every case that the parser recognized as *to* followed by the verb *ikeru*, the form of the verb is negative (e.g. *yaranai to ikenai*). The string 'negative + *to ikenai*' is a conventionalized marker of obligation (Fujii 2004).

for 14.19% of R1 collocates of *tte* and 5.79% of *to*. Since it occurs at the end of an orthographic sentence, the full-stop is useful for identifying sentence-final occurrences of the particles. The transcription is not consistent in using full-stops at the end of sentences, however. Question marks, parenthesis or angled brackets enclosing transcribers' notes, as well as other marks of punctuation also sometimes occur at the end of an utterance or turn at talk. Therefore all marks of punctuation were grouped together as a single category in counting R1 collocates.[3] The resulting category "all punctuation" occurred as the most common R1 collocate of *tte* and the second most common R1 collocate of *to*.

4. Results

The most common item following the particle *tte* was punctuation, accounting for 28.35% of R1 collocates. The second most common item was the verb *iu* "say" at 17.01% of R1 collocates. The ten most common items are summarized in Table 1.

Table 1 Right-collocates of *tte*

word type	tokens	percent	word type	tokens	percent
punctuation	6,052	28.35	の *nominal*	385	1.80
言う "say"	3,631	17.01	書く "write"	306	1.43
こと "thing"	879	4.12	何 "what"	278	1.30
感じ "feeling"	574	2.69	ね *interaction*	266	1.25
思う "think"	558	2.61	聞く "listen"	199	0.93

Among verbs following the particle *tte*, the verb *iu* "say" is by far the most common, accounting for 64.25% of verb collocates and 17.01% of all collocates. The second most common verb is the cognition verb *omou* "think", accounting for 9.87% of verb collocates. The third and fourth most common verb collocates, *kaku* "write" (5.41%) and *kiku* "hear" (3.52%) are, like *iu*, verbs suggesting verbal communication. The ten most common verb collocates of *tte* are summarized in Table 2.

---

[3] Punctuation marks following *tte* or *to* are the following, in order of frequency: 。 、 （ ？ ＊ ） ＜ ー 「

Table 2 Verb right-collocates of *tte*

word type	tokens	percent	word type	tokens	percent
言う "say"	3,631	64.25	感じる "feel"	60	1.06
思う "think"	558	9.87	出る "go out"	50	0.88
書く "write"	306	5.41	呼ぶ "call"	39	0.69
聞く "listen"	199	3.52	食べる "eat"	38	0.67
知る "know"	96	1.70	行う "do"	30	0.53

The most common item following the particle *to* is the verb *omou* "think", which accounts for 27.56% of R1 collocates. The second most common item is the combined category of punctuation with 24.00% of collocates. The words *issho* "together, same" (1.07%) and *onaji* "same, identical" (0.55%) were among the ten most common R1 collocates of *to*. Each of these words is likely to follow the associative particle *to* "and, with" rather than the quotative, as illustrated in examples 5-6.

5. F021：でも 嫁　　に　行って さー、親　　と 一緒　　に　うまく
   　　　　but bride LOC going PT　parent **TO together** LOC well
   　　　できる
   　　　do
   "But if she gets married, you know, can she get along with his parents"

6. オペラ歌手　　と 同じ　ように 歌わないでー、
   opera-singer **TO same** so-as sing-NEG
   でも 発声　は　きちんと　やって ね。
   but utter TOP precise do　　PT
   "Don't sing the same as an opera singer, but vocalize precisely, you know."

Table 3 Right-collocates of *to*

word type	tokens	percent	word type	tokens	percent
思う "think"	7,670	27.56	言う "say"	286	1. 03
punctuation	6,678	24.00	いける "do"	158	0.57
は *topic*	627	2.25	同じ "same"	153	0.55
ね *interaction*	539	1.94	も "also"	123	0.44
一緒 "together"	299	1.07	だめ "cannot"	117	0.42

The verb *ikeru* "do, be able to do" (0.57%) is also among the ten most common verbs. In every case where *ikeru* follows *to*, the form of the verb is the negative *ikenai*, as illustrated in examples 7 and 8. The string (*verb*) *to ikenai* indicates prohibition;

when the initial verb is negative, as was the case in 89.94% of *to ikenai* tokens in the corpus, the string indicates deontic modal obligation.

> 7. いちばん 早い　人　　 は　6時 過ぎ から　もう　　 いて、子ども が
>   first   early person TOP six pass from already be child   PT
>   来る　 と　いけないから　　 さー。
>   come ***TO*** can't from PT
>   "The first people have already come after six, since children can't come."

> 8. F159：ちょっと 休んじゃいけないの？（↓）F004：うん？
>   　　　 little rest   can't PT　　　　　 yes
>   F159：通し　　　 じゃないといけない　の？
>   　　　 straight NEG　 ***TO*** can't PT
>   "Can't you rest a bit." "What?" "Do you have to go straight through?"

The ten most common R1 collocates in the data are summarized in Table 3.

Among verbs following the particle *to*, the verb *omou* "think" is extremely common, comprising 81.32% of verb collocates (27.56% of all R1 collocates). The communication verb *iu* "say" is the second most common verb following the particle, but it accounts for only 3.03% of verb collocates (1.03% of all R1 collocates). The ten most common verb collocates of *to* are summarized in Table 4.

Table 4　Verb right-collocates of *to*

word type	tokens	percent	word type	tokens	percent
思う "think"	7,670	81.32	見る "look"	60	0.64
言う "say"	286	3.03	出る "go out"	45	0.48
違う "differ"	98	1.03	書く "write"	44	0.47
話す "talk"	77	0.82	比べる "compare"	43	0.46
行く "go"	66	0.70	会う "meet"	40	0.42

5.　Discussion

It is frequently asserted that *tte* is equivalent with *to*, differing primarily in formality (e.g. Hayashi 1997; Shinmura 1998; Kaiser et al. 2001). That is, *tte* occurs only in colloquial or casual speech while *to* may be used with various levels of formality. Analysis of a corpus of conversations among acquaintances shows the frequent use of both *tte* and *to*. While it is difficult to specify the degree to which particular uses are colloquial or casual with broad-based corpus methods, the Meidai Kaiwa Corpus

consists of conversation among previously acquainted speakers with no researcher-specified topics. Such data collection methods tend to produce less formal types of speech. While these data can say nothing conclusive about the relative level of formality of *to* and *tte*, they reveal that each particle is used in friendly conversation.

5.1 Particle-verb collocation patterns

The data further suggest that *tte* and *to* are not utilized in identical contexts. Although each particle can be followed by the verb *iu* "say" and each can be followed by the verb *omou* "think", there is a strong tendency to prefer the collocation *tte iu* (3,631 occurrences) over *to iu* (286 occurrences), and likewise to prefer *to omou* (7,670 occurrences) over *tte omou* (558 occurrences). This is consistent with speaker intuitions reported to Martin (1975) that *tte* is used to mark reported speech as opposed to the content of thoughts. As Martin noted, however, his informants' observations were overly categorical: despite the strong preference for *to*, *tte* does occur with *omou* to mark the content of thoughts. Examples from the data include 9-10.

9. F060：こう、空間　も　　ねー（うん）うまく使ってるって思う。（↓）
    here space also PT (yeah) well using ***TTE think***
    （ほー）うん。
    ho yeah
    "Here, this space too, right, [yeah] they're using it well, I think. [oh] Yeah."

10. そう　いう の　買う とき　には　　　なんか、これ　を　飲む って
    that say PT buy time LOC-TOP things this OB drink ***TTE***
    思われたくない　　な　って 思う。
    think-want-neg PT ***TTE think***
    "When I buy that stuff, like, I think, 'I don't want them to think I drink this'."

Table 5 Quotatives following *omou* "think"

quotative +verb	tokens	percent
と思う (*to omou*)	7,670	93.22
って思う (*tte omou*)	558	6.78

The corpus includes 7,670 occurrences of the string *to omou* compared to just 558 occurrences of *tte omou*. Thus, in 93.22% of cases of quotative + *omou*, the quotative particle is *to* (see Table 5). The preference for *to* in such constructions is very great, but not categorical. The examples cited above show no sign of disfluency or other indications that *tte omou* constitutes a speech error or an ungrammatical utterance.

　The frequency of *to omou* appears to confirm the notion that *to* is the preferred marker for the content of thoughts, as suggested by both Jorden (1990 [1962]) and Martin (1975). Two other verbs complicate this idea, however. The verbs *kangaeru* "consider" and *kimaru* "decide", can also occur with a clause or phrase indicating the content of thoughts. The strings *to kangaeru* and *tte kangaeru* each occur 14 times in the corpus. Such usage is illustrated in examples 11 and 12.

> 11. M023：明日　　　休む　と　考えた　　だけで、
>   tomorrow rest ***TO consider*** only
>   こんな時間　まで　起きてられる。
>   this time until rise
>   "Only by thinking, 'Tomorrow is a day off,' (I) can stay up this late."

> 12. （↓）どう しよう か って 考えてる　暇　　あったら、（＜笑い＞）
>   what do　Q ***TTE consider*** leisure exist　　(laugh)
>   "'What should (I) do,' (I'm) thinking, if there is free time"

The string *to kimaru* occurs eight times, while *tte kimaru* occurs ten times. Usage is illustrated in examples 13 and 14.

> 13. 何か　28 で　相手　　と　決まって も、　結婚　できるのは
>   what 28 LOC partner ***TO decide*** also marry do　　TOP
>   ちょっと 遅れて
>   little be-late
>   "Like, even if you decide on a partner at 28, it's a little late to get married"

> 14. F024：そう。（↓）余る　　もの　って 決まっててね。（↓）F140：そう。
>   so　　　remain thing ***TTE decide*** PT　　　　so
>   "Right. It has been decided which things to leave out." "Right."

The verbs *kangaeru* and *kimaru* do not show the strong tendency for occurrence with *to* that was observed with the verb *omou*. This suggests a possibility that the quotative *to* has an affinity specifically for the verb *omou* rather than for marking

the content of thought generally. Given the small number of occurrences of each verb, however, strong conclusions may not be warranted.

As the verb *omou* shows a strong tendency to occur with *to*, the verb *iu* "say" tends to occur with the quotative *tte*. The corpus includes 3,631 occurrences of the string *tte iu* and 286 occurrences of *to iu*. This means that in 92.70% of quotative + *iu* cases, the quotative is *tte*. Moreover, this tendency is also seen with other verbs indicating communication. Quotative particles *to* and *tte* occur with the verbs *kaku* "write", *kiku* "hear", *yobu* "call", and *hanasu* "speak". As shown in Table 6, the tendency holds for most verbs of communication. In cases of quotative + communication-verb, the particle is overwhelmingly *tte* (90.47% of combined cases).

Table 6 Particles following communication verbs

verb	*tte*	*to*	*% tte*
言う (say)	3,631	286	92.70
書く (write)	306	44	87.43
聞く (hear, ask)	199	25	88.84
呼ぶ (call)	39	9	81.24
話す (talk)	13	77	14.44

There is one apparent exception to this pattern. The verb *hanasu* "speak, converse" frequently occurs with *to*; the string *to hanasu* is nearly six times as frequent as *tte hanasu*. This appears to be an artefact of conflating associative *to* with quotative *to*. In most cases within the data, *to hanasu* appears to indicate a person or group spoken to, as illustrated in examples 15 and 16.

15. （＜笑い＞）やっぱこう　M030さん　と　話す　と、なりますね。（↓）
    (laugh) also this M030-san *TO* speak *TO* become PT
    M030：何、
          what
    "After all, (you) get that way if you speak to Mr. (M030)." "What?"

16. 年上の 人　　 とか いろんな 世界の 人　　 と 話さなきゃだめ だ という。
    elder person TOKA various world person *TO* speak-must COP *TO* say
    "They say you must speak with older people and various people of the world."

Of the 77 occurrences of the string to *hanasu*, 13 are preceded by the honorific *san* "Mr., Ms." and another four by *chan*, a diminutive variant of *san* used for children

or intimates. Forty-five of the 77 occurrences follow a noun, with the most common tokens *hito* "person" (13), *sensei* "teacher" (10), *kodomo* "child" (6), and *tomodachi* "friend" (5). In other words, the string *to hanasu* usually indicates the person with whom one is speaking – a use of the associative particle *to* – and not the content of the speech.

5.2 Particles in sentence-final position

While Japanese quotative particles often occur before a verb indicating speech or thought, they can also occur at the end of a sentence or utterance. Sentence-final occurrences may indicate direct or indirect quotations in much the same way that quotative + verb combinations do. In addition, such usage can serve pragmatic functions such as distancing a speaker from the ideas expressed, marking evidentiality, or marking as topic some idea not overtly expressed in the preceding discourse (Okamoto 1995; Nilep and Fujimoto 2003; Suzuki 2007, inter alia). Some scholars (e.g. Martin 1975; Tanaka 2001) have suggested that sentence-final *tte* may be a reduced form of a construction containing a verb, such as *tte itte* (quotative *tte* plus a non-finite inflected form of the verb *iu* "say"), though this suggestion remains somewhat controversial. While the methods employed here can shed little light on that controversy, by examining occurrences of *to* or *tte* before punctuation it is possible to approximate the relative frequency of each particle in sentence-final position.

As noted above, the combined category of all punctuation is the most common right-collocate of *tte* and the second most common following *to*. Furthermore, the raw number of occurrences is similar, with 6,052 tokens of *tte* followed by punctuation compared to 6,678 tokens of *to*. At first glance, this seems to suggest that the particles are more or less equally common in sentence-final position. However, as was noted above, the morphological analysis used here conflates forms that may be semantically and morphologically distinct but written or pronounced similarly. Therefore, close inspection of *to* or *tte* followed by punctuation is needed.

To facilitate close inspection, a random selection was made from the instances of each particle followed by punctuation. First, all keywords in context (KWIC) in which *to* is followed by a mark of punctuation were combined into one list using Microsoft Excel. From this list of 6,678 KWIC, fifty were randomly selected. Each selection was then analyzed and coded as either containing the quotative particle, the associative particle, some other occurrence of orthographic *to*, or in cases where the function was unclear coded as "uncertain". The same procedure was followed

for fifty randomly-selected KWIC in which *tte* is followed by punctuation.

Of the fifty KWIC containing *to*, only two were judged to be instances of the quotative particle, shown as examples 17 and 18.

17. 大学　　受ける　とき　に、（うん）何か　　経済　　とか 法科 とか 違う な
    college join time LOC (yeah)what econ TOKA law TOKA wrong PT
    違う　　な　と（うん）思ってて、（うん）で、ま、
    wrong PT ***TO*** (yeah)thinking (yeah)and um
    美大　　　　の　建築　　　か、
    art-school PT architect PT
    "When (I) go to college [yeah] (I'm) thinking, [yeah] 'not economics, not law, [yeah] so maybe architecture at an art school'"

18. 一応　午前中　　ぐらい は　　クーラー つけ　　ない ようにしようと（＜笑い＞）
    once morning about TOP cooler attach NEG so do ***TO*** (laugh)
    思って たん　だけど さー。F026：うん、無理　　　　　じゃ
    think did but PT yeah impossible COP
    "Like, sometimes I think, [laughter] 'I won't use the air conditioner in the morning,' but…" "Yeah, it's impossible."

In each of these cases, although the quotative *to* is followed by a mark of punctuation, neither is actually in sentence-final position. In each case, the punctuation following the quotative particle is a left-parenthesis, used to indicate overlapping speech from another speaker. Following the parenthetical overlap, each quotative is followed by the verb *omou*. Thus, at least in the 50 KWIC analyzed, *to* followed by punctuation does not actually indicate a quotative in sentence-final position.

One of the occurrences of *to* followed by punctuation which was coded as uncertain does appear to be serving a function similar to the functions of sentence-final quotatives enumerated above. It is shown as example 19.

19. もうちょっと、えーと（どちら）こっち より　と　いう の　か。（↓）
    little-more um (which)this more TO say PT Q
    F040：方向　　　だ　と、東　西　南　　北。（↓）
    direction COP ***TO*** east west south north
    F081：方向　　　では、えー、
    direction PT um
    東　西　南　　北、　　名古屋　側、うち　から
    east west south north Nagoya side home from
    "...a little more [which way?] this way maybe." "A direction: east, west, north, or south" "Direction, um, toward Nagoya from here…"

In this utterance, *houkou da to*, *touzainanboku* "A direction [*to*], east, west, north, or south," the particle *to* may mark the word *houkou* "direction, orientation," as the topic of the discussion, one of the enumerated functions of sentence-final quotative particles. It seems equally valid, though, to see the function of *to* in this utterance not as quotative but as a conditional, like those described below.

The conditional *to*, frequently translated as 'if', marks a possible or likely event or outcome. Thirteen of the 22 KWIC coded as "other" are uses of the conditional *to*, as illustrated in 20.

> 20. M028：大体、毎　　日　　2本　飲みます ね、あっち 行く と。（↓）（うん）
>   　　　　main each day 2　　drink　PT there go ***TO***　　　　yeah
>   （へー）（↓）M022：いや、飲んじゃいます　　よ。（↓）だって、お昼　　から
>   　oh　　　　　　　　no　drink-complete PT　　but　noon from
>   "Pretty much everyday I drink two bottles, if I go there." "Yeah" "Oh" "Really, I drink a lot, you know. Like, from afternoon…"

In example 20, *mainichi ni-hon nomimasu*, *acchi iku to* "I drink two bottles a day if I go there," the particle *to* indicates the condition that results in the speaker drinking. Returning to example 19, in the string *houkou da to* "(if) it is a direction", the particle *to* can be analyzed as marking a condition in which directions must be given, or naming "orientation" as the topic, or serving both functions.

Within the 50 KWIC analyzed with *to* followed by a mark of punctuation, none are clear examples of the quotative in sentence-final position. It is possible that some number of tokens serve functions similar to those described, for example, by Hayashi (1997) for quotative *to* in sentence-final position. But given the frequency of conditional or associative function in the KWIC analyzed, that number is presumed to be relatively small.

While sentence-final *to* revealed a range of uses and arguably includes three distinct particles, sentence-final *tte* is much more unified. Thirty of the fifty KWIC analyzed were clearly quotative markers, as illustrated in 21.

> 21. F110：それは 何、　あれ　ダウンロード して くる　の？ おまけ って。（↓）
>   　　　　that what that download do　come PT　free　***TTE***
>   F136：え、ダウンロード じゃなくて。（↓）F110：もともと　入ってるの？（↓）
>   　　　um download　not　　　　　　　　original enter　PT
>   "What is that, did you download it for me? Free, you say." "Um, not downloaded." "Was it already there before?"

Only nine instances had clear non-quotative functions. These were primarily verb inflections that the morphological analyzer identified as separate lexical items, as shown in 22.

22. （そう なん　だ）うん、あの　ほら、インターネット ホームページ 作りたい　とか
    that what COP yeah that look internet homepage　make-want TOKA
    言って。（↓）（うん うん）でー、そのー、あのー、何だっけ、えー、ディスプレイ
    say-***TTE***　　 yeah yeah and　that　 that　 what　 um　 display
    "Is that so." "So, look, like, if (you) say, 'I want to make a web page.'" "right, right" "and, um, that display…"

The eleven remaining KWIC coded as uncertain appear to reflect a mixture of quotatives, inflections, and artefacts owing to transcription conventions, errors, or inconsistencies. Example 23 illustrates one such unanalyzable token.

23. あなた へ の　愛　　を　ロンドン に　持って 帰ろう　って　いう 歌詞　が　ある。
    you　 to PT love OB London LOC carry return TTE say lyric PT exist
    （↓）＊＊＊ってー。（↓）（へー）（↓）
    　　　??? 　***TTE***　　　　 oh
    F083：なんか、＊＊＊詞　　って いう の
    　　　 what　 ??? 　words TTE say PT
    "There are lyrics about wanting to bring your love home to London. ***" "Oh" "What, *** words…"

Note that, while there are several tokens of quotative *tte* in example 23, it is the bold faced element, *tte* preceded by three asterisks and followed by a bar, that was randomly selected for analysis. Given the asterisks, marking an omission, and the other idiosyncrasies of this fragment, the function of *tte* is unclear.

   Although there is no guarantee that the analyzed cases are typical, if the proportion of quotative and other functions seen in these one hundred key words in context are representative of the 6,052 tokens of *tte* and 6,678 tokens of *to* followed by punctuation in the corpus, then once again a similar pattern emerges. There is a clear and strong preference for quotative *tte* rather than *to* at the end of sentences, though *to* may also appear occasionally. As with verbs of communication and verbs of thought, there are clear patterns, but the patterns are not categorical. Table 7 shows the number of each particle type followed by punctuation observed in the 50 KWIC analyzed. The number in parenthesis is the estimated number of each particle type in the entire corpus, if these proportions are consistent. The latter numbers should, of course,

be viewed skeptically. Furthermore, as shown above, the presence of punctuation does not necessarily indicate the end of an utterance. The estimated frequency of quotative *tte* followed by punctuation (approximately 3,631) is virtually the same as the observed frequency of *tte* followed by the verb *iu* (3,631). In contrast, the estimated frequency of *to* followed by punctuation (267 unambiguous quotatives plus 2,671 uncertain cases) is far less than the observed cases of *to omou* (7,670).

Table 7　Particles followed by punctuation

	quotative	associative	uncertain	other
って　(*tte*)	30 (3,631)	0 (0)	11 (1,331)	9 (1,089)
と　　(*to*)	2 (267)	6 (801)	20 (2,671)	22 (2,938)

Conclusion

An analysis of the more than 100 hours of multi-party conversation in the Meidai Kaiwa Corpus reveals that the quotative particles *to* and *tte* show clear and distinct co-occurrence patterns. When referring to the content of thought, the particle *to* is overwhelmingly preferred, most often appearing with the verb *omou* "think". When referring to verbal communication, including direct and indirect quotation, the particle *tte* is preferred, most often followed by the verb *iu* "say". When a quotative particle occurs at the end of a sentence and no verb follows, again the particle *tte* is strongly preferred. These preferences are very large; *to omou* is more than ten times as common as *tte omou*, while *tte iu* is more than ten times as common as *to iu*. Likewise, analysis of a random selection of sentence-final particles suggests that *tte* is fifteen times as frequent as the quotative *to* before a mark of punctuation, though the analysis also reveals that the presence of punctuation does not necessarily indicate the end of an utterance.

Despite the strength of the patterns revealed, however, these preferences are not categorical rules. There is no suggestion that less common strings such as *tte omou* are ungrammatical. These patterns confirm empirically some native-speaker and non-native-speaker analyst intuitions about the particles' usage (Jorden 1990 [1962]; Martin 1975). They are also consistent with the results shown for a much smaller and more closely analyzed corpus by Nilep and Fujimoto (2003).

In addition to the empirical findings, this study reveals some difficulties inherent

in the use of morphological parsers, which are necessary for large-scale corpus study. Because all grammatical particles were classified as part of the same category, orthographically and phonetically similar forms such as quotative, associative, and conditional particles (*to*) could not be distinguished. This suggests two problems to be overcome, one technical and one theoretical.

The technical issue relates to a possible problem of circularity. The current author is not trained in the technical design of parsers or morphological analyzers. However, in theory such systems generally rely on two elements: a lexicon of words and morphemes, and a set of morphosyntactic and morphophonological rules which apply to them. If multiple words have the same form, they must be distinguished primarily by syntactic role. In the case of Japanese particles, word order or other surface markers of syntactic role also partially overlap. It would seem that care must be taken in specifying syntactic rules that are sufficiently specific to distinguish one particle from another, and yet are not so over-specified that they fail to identify relevant instances. To state the risk of circularity concretely, over-specifying the morphosyntactic contexts in which a particle can occur risks only finding instances that support current models of grammar. This in turn risks missing empirical evidence from corpus methods that might allow refinement of those models.

The related problem for linguistic theory is no less thorny. This analysis has assumed that the quotative, associative, and conditional particles are homonyms, distinct forms within the lexicon that share surface form. However, it is sometimes difficult to draw clear lines between these forms. Consider, for example, the following utterance, adapted from Takanashi (2011, 244-245) with original morpheme gloss.

24. 名前で　呼んだり とか。たまに　　でも　私 も　ベイビイ とか　呼んじゃう。
　　 name by call　*or* sometimes but　I　too 'baby'　*QT*　call
　　 "I call him by name, or… But sometimes I also call (my boyfriend) 'baby'."

Takanashi glosses the first instance of *toka* as "or" and the second instance as a quotative marker (glossed QT). While the analysis of corpus data above discarded all instances of *toka* as a lexical item distinct from the quotative, Takanashi's analysis suggests that some tokens of *toka* may consist of the quotative *to* combined with the interrogative *ka*. Takanashi is far from alone in this analysis. Other scholars have likewise viewed some instances of *toka* (Suzuki 2002; Koike 2010) or *dato* (Okamoto 1995; Suzuki 2007) as quotative markers. One can posit that two homophones

pronounced *to* followed distinct paths to lexicalization as *toka*, resulting in another pair of homophones. However, it is equally possible to posit a single *to* particle that is sufficiently vague (Geeraerts 1993) to function as quotative, associative, or conditional, depending on context. In short, it is not clear whether these occurrences of *to* represent either a set of distinct, homophonous particles or a single, highly polysemous particle.

The problem of polysemy (multiple meanings of a single lexical item) versus homophony (distinct lexical items with the same phonological form) is a vexed one in linguistics (Nunberg 1979; Geeraerts 1993; Brown 2008, inter alia). In the absence of theoretical or philosophical agreement about the nature of lexical meaning, technical solutions to the issues presented here are likely far off.

References

Brown, Susan Windisch. 2008. "Polysemy in the mental lexicon." *Colorado Research in Linguistics* no. 21:1-12.

Fujii, Seiko. 2004. "Lexically (un)filled constructional schemes and construction types: The case of Japanese modal conditional constructions." In *Construction Grammar in a Cross-Language Perspective*, edited by Mirjam Fried and Jan-Ola Östman, 121-155. Amsterdam: John Benjamins.

Geeraerts, Dirk. 1993. "Vagueness's puzzles, polysemy's vagaries." *Cognitive Linguistics* no. 4 (3):223-272.

Hamano, Shoko. 1998. *The Sound-Symbolic System of Japanese*. Edited by Masayoshi Shibatani. Vol. 10, *Studies in Japanese Linguistics*. Stanford: CSLI.

Hayashi, Makoto. 1997. "An exploration of sentence-final uses of the quotation particle in Japanese spoken discourse." In *Japanese/Korean Linguistics 6*, edited by Ho-min Sohn and John Haig. Stanford: CSLI.

Higuchi, Koike. 2001. KH Coder. http://khc.sourceforge.net/en/.

Jorden, Eleanor Harz. 1990. *Beginning Japanese*. 14th ed. Tokyo: Charles E. Tuttle. Original edition, 1962.

Kaiser, Stefan, Yasuko Ichikawa, Noriko Kobayashi, and Hilafumi Yamamoto. 2001. *Japanese: A Comprehensive Grammar*. London and New York: Routledge.

Kinuhata, Tomohide. 2012. "Historical development from subjective to objective meaning: evidence from the Japanese question particle *ka*." *Journal of Pragmatics* no. 44 (6-7):798-814. doi: 10.1016/j.pragma.2012.03.004.

Koike, Chisato. 2010. An analysis of shifts in participation roles in Japanese storytelling in terms of prosody, gaze, and body movements. Paper read at The twenty-seventh annual

meeting of the Berkeley Linguistics Society, at Berkeley.

Kuno, Susumu. 1973. *The Structure of the Japanese Language*. Edited by Samuel Jay Keyser, *Current Studies in Linguistics*. Cambridge, MA: MIT Press.

Martin, Samuel E. 1975. *A Reference Grammar of Japanese*. New Haven: Yale University Press.

Matsumoto, Yuji, Akira Kitauchi, Tatsuo Yamashita, Yoshitaka Hirano, Hiroshi Matsuda, Kazuma Takaoka, and Masayuki Asahara. 1999. Japanese Morphological Analysis System ChaSen 2.0. Nara Institute of Science and Technology, Nara.

Nilep, Chad, and Junko Fujimoto. 2003. Animating functions of quotative particles *to* and *tte* in spoken Japanese. In *JSLS 2003 The 5th Annual Conference of the Japan Society for Language Science*, edited by Tamiko Ogura. Kyoto University: JSLS.

Nunberg, Geoffrey. 1979. "The non-uniqueness of semantic solutions: polysemy." *Linguistics and Philosophy* no. 3 (2):143-184.

Ohso, Mieko. 2003. 名大会話コーパス [Meidai Kaiwa Corpus]. Tokyo: National Institute for Japanese Language and Linguistics.

Okamoto, Shigeko. 1995. "Pragmaticization of meaning in some sentence-final particles in Japanese." In *Essays in Semantics and Pragmatics: In Honor of Charles J. Filmore*, edited by Masayoshi Shibatani and Sandra Thompson, 218-246. New York and Amsterdam: John Benjamins.

Shinmura, Izuru. 1998. *広辞苑 [Koujien dictionary]*. Tokyo: Iwanami.

Suzuki, Satoko. 1998. "Pejorative connotation: a case of Japanese." In *Discourse Markers: Descriptions and Theory*, edited by Andreas Jucker, 261-276. New York and Amsterdam: John Benjamins.

———. 2002. "Self-mockery in Japanese." *Linguistics: An Interdisciplinary Journal of the Langauge Sciences* no. 40 (1):163-189.

———. 2007. "Metapragmatic function of quotative markers in Japanese." In *Metapragmatics in Use*, edited by Axel Hübler and Wolfram Bublitz, 73-85. Amsterdam: John Benjamins.

Takanashi, Hiroko. 2011. "Complementary stylistic resonance in Japanese play framing." *Pragmatics* no. 21 (2):231-264.

Tanaka, Hiroko. 2001. The implementation of possible cognitive shifts in Japanese conversation. In *Studies in Interactional Linguistics*, edited by Margaret Selting and Elizabeth Couper-Kuhlen. Amsterdam/Philadelphia: John Benjamins.

Appendix: Transcription conventions

TO	any instance of *to*	NEG	negative inflection
TTE	any instance of *tte*	OB	object marker *o*
TOKA	any instance of *toka*	Q	interrogative marker
COP	copula	TOP	topic marker *wa*
LOC	locative postposition	PT	other particles not specifically analyzed