Lexical and grammatical features of spoken and written Japanese in contrast: exploring a lexical profiling approach to comparing spoken and written corpora

Itsuko FUJIMURA, Shoju CHIBA, Mieko OHSO
Nagoya University; Reitaku University; Nagoya University
Furo-cho, Chkusa-ku, Nagoya, Japan
fujimura@nagoya-u.jp, schiba@reitaku-u.ac.jp, ohsomk@ac.auone-net.jp

This paper statistically demonstrates the lexical and grammatical characteristics of conversational Japanese by comparing a 100 hour spontaneous spoken corpus: the NUCC (Nagoya University Conversation Corpus) with a written corpus: the Balanced Corpus of Contemporary Written Japanese (monitor version). 1) The conversation corpus contains more involved production than the compared written corpus. 2) The comparison between the spoken and written interactional corpora shows that the participants leave much more metalinguistic and illocutionary traces in their speech than their writing. This is explained by the difference of degree of elaboration of the emitted messages and the difference of degree of closeness between/among participants of exchanges. 3) Fragmented utterances are much more frequent in spoken conversation than written texts. In Japanese, because of its grammatical structure (=SOV type language; particles come after their head), fragmentation, omnipresent conversational phenomenon, easily causes a functional and grammatical change in the role of particles.

Keywords: conversation; internet exchanges; metalinguistic; norm; linguistic change; Japanese; fragmentation.

1. Introduction

In this paper, we describe the lexical and grammatical characteristics of Japanese face-to-face spoken conversation and show how they differ from written registers. The aim of this research is to elucidate the characteristics of spoken Japanese, so we can later compare them with the results piled in the literature of this domain (Blanche-Benveniste, 1990; Biber, 1995 among others). For this purpose, we compare a spoken corpus: the NUCC (Nagoya University Conversation Corpus) with a written corpus: the BCCWJ (Balanced Corpus of Contemporary Written Japanese, monitor version). The former is a corpus of 100 hours built by our research team. The latter is a 45 million morpheme-sized written corpus. Our method is mainly quantitative. We perform this research with a tool named Lexical Profiling System, devised by one of the co-authors of this paper.

2. Corpora and tool

2.1 NUCC

The NUCC was constructed between 2001 and 2003, and is available for research purposes from the site (https://dbms.ninjal.ac.jp/nuc/index.php?mode=viewnuc) free of charge. It is composed of transcriptions of 129 uncontrolled, natural conversations between or among friends, family members or colleagues. Each conversation has 2 to 4 participants and lasts 30 to 60 minutes. The participants are 198 native speakers of Japanese of various ages and from diverse academic backgrounds. Each conversation constitutes a file so that the corpus NUCC consists of 129 files.

Conversations were recorded and transcribed in standard Japanese orthography. The Japanese orthography currently used is quite phonemic, but suprasegmental features are not captured. Hence, accent, intonation, and prominence are not transcribed. Only the rising intonation that indicates questioning is marked with a question mark at the end of an utterance.

The corpus contains about 1.5 million morphemes ("short unit words" according to UniDic (cf. Ogiso *et al.*, 2012)), which shows that this is the largest corpus currently available of spontaneous spoken Japanese. As a caveat, there are more female participants (161) than male (37), and many of the participants are graduate students majoring in linguistic subjects. The lack of balance of the participants may be reflected in the data taken from this corpus.

2.2 BCCWJ (monitor version) [1]

The integral BCCWJ, published in 2012, includes about 170,000 samples of written texts, which are classified into carefully designed subcorpora (genres), namely books, newspapers, magazines, whitepaper texts, Internet texts, Diet minutes, among others. We see the BCCWJ as a good sample of written Japanese, because the corpus contains the samples from many genres, each of which is relatively large. It also utilizes unique sampling strategies so that the corpus represents the most recent status of contemporary written Japanese (Maekawa, 2007).

In this work, we used the monitor version of the BCCWJ earlier released in 2009, which is a part of the integral version. The monitor version consists of 4 subcorpora indicated in Table 1. We use the BCCWJ in two ways. One is the whole BCCWJ (monitor version) for the grammatical study in section 4, and the other, its subcorpora: Books (BK) and Internet Bulletin Boards (IBB) for the lexical studies in section 3. The BK is composed of 10423 samples taken from various genre of books published between 1971-2005. We used it because it is the largest part of the BCCWJ and for its standardized nature as written corpus. The IBB consists of "Questions and Answers" type written exchanges between anonymous writers and readers, published on Yahoo Japan's web site in 2005. The IBB is an interesting material to compare with the NUCC, because of their shared characteristics and for its novelty as a medium of communication. Both of them involve interaction

[1] Cf: http://www.ninjal.ac.jp/english/products/bccwj/. The BCCWJ refers to the BCCWJ (monitor version) from section 3 below.

Heliana Mello, Massimo Pettorino, Tommaso Raso (edited by), *Proceedings of the VIIth GSCP International Conference : Speech and Corpora*
ISBN 978-88-6655-351-9 (online)

between/among participants. The relation between/among participants is different though; the participants in the latter have close relationships while those in the former are strangers. They made real-time interactions in the latter, while there is a time lag between questions and answers in the former.

Table 1 indicates the characteristics of the studied corpora.

Subcorpus of BCCWJ and NUCC	Number of morphemes (millions)	Characteristics
Books (BK)	36.0	No interaction Elaborated production
White Paper	5.8	
Internet Bulletin Boards (IBB)	6.7	Long-distance interaction Prepared production
Minutes of the National Diet	5.5	
NUCC	1.5	Close interaction Real-time production

Table 1: Subcorpora of the BCCWJ (monitor version) and the NUCC

2.3 Lexical profiling system

The Lexical Profiling System is designed to compare corpora of different size, genre, or even an individual part of a corpus with the whole. The data to be compared are morphologically analyzed by a GUI program Chamame (ver. 1.71) (composed by a part-of-speech and morphological analyzer: Mecab (ver. 0.98) and a dictionary: UniDic (ver. 1.3.12)), and the frequency of lemmas, word forms, bigrams are counted and stored in a database. The tool then computes the frequencies of these units using different statistical measures such as LLR (Log-Likelihood Ratio) among others.

3. Lexical studies

3.1 60 Basic morphemes in the NUCC

First of all, we identified the 60 morphemes employed in all 129 conversations of the NUCC as in Table 2 in order to compare later the use of these morphemes in the NUCC and the IBB and the BK. We could say that these are basic morphemes of Japanese conversation. These consist of 6 adjectives, 4 adverbs, 1 conjunction, 4 interjections, 6 nouns, 18 particles, 1 prefix, 2 pronouns and 12 verbs[2]. Among the 18 particles, there are 4 utterance-final interactional particles, 13 sentence-internal casual or conjunctive particles and "*no*". "*No"*, one of the most frequently used morphemes in Japanese, is subcategorized into three according to the dictionary UniDic: genitive (*of* in English), quasi-nominal (*thing*, nominalizer) and interactional. The first two are sentence-internal particles and the last one, utterance-final particle.

POS	No	Morpheme
ADJ	6	*nai (not to exist), yoi (good),* *you (to look like), sugoi (superb),* *sonna (that kind of), sono (that)*
ADV	4	*mou (already), dou (how),* *sou (so, in such a way), kou (this way)*
AUX	6	*da, desu (*DEC*), reru (*PASS/POT/HON*),* *ta (*PAST*), nai (*NEG*),* *teru (*PROG, PERF*)*
CONJ	1	*de (and)*
INTJ	4	*un (yeah, I see), ah , a! (wow),ano (well)*
NOUN	6	*koto (matter), hito (person), toki (time, when), hou (side), ato (behind, afterward), mono (thing)*
PRT	18	Utterance-final, interactional: *ne (*TAGQ*, you know),* *yo (I tell you), ka (*Q*), na (I tell you)* Sentence–internal: *wo (*ACC*), ga (*SUB*),wa (*TOP*),* *ni* (DAT, LOC, TEMP, ADVL), *to (and with), keredo (although),* *kara (from), mo (also), kurai (about)* *te, de (and (*V/ADJ Suffix*))* *tte (*QUO*), made (until),* *no*: GEN,QN (sentence-internal), INTA (utterance-final)
PREFIX	1	*o* (POLITE)
PRO	2	*nani (what), sore (that)*
VERB	12	*iru (to exist, to be), dekiru (to be able to),* *miru (to see, to look at), naru (to become),* *wakaru (to understand), omou (to think),* *aru (to exist), kuru (to come), suru (to do),* *yaru (to do), iku (to go), iu (to say)*
total	60	

Table 2: 60 Morphemes used in all 129 conversations of the NUCC[3]

The fact that there are no personal pronouns in the list should not be interpreted as lack of active interaction. In Japanese, one can speak even for 30 minutes long without mentioning "me" or "you". Especially the

[2] These are the output of the Analyzer Chamame. We only modified the result of the automatic analysis by grouping "Rentai-shi", "Keijo-shi" and "Keiyo-shi" in Adjective, since the major function of these three categories is noun modification.

[3] Glosses are approximate due to lack of space. The list of abbreviations is following. ADJ: Adjective, ADV: Adverb, ADVL: Adverbial, ACC: Accusative, AUX: Auxiliary, CONJ: Conjunction, DAT: Dative, DEC: Declarative, HON: Honorific, INTJ: Interjection, INTA: Interactional, NEG: Negation, GEN: Genitive, PASS: Passive, PAST: Past Tense, PERF: Perfect, POT: Potential, PRO: Pronoun, PROG: Progressive, SUB: Subject, TAGQ: Tag-Question, Q: Question, TEMP: Temporal, QN: Quasi-Nominal, TOP: Topic, PRT: Particle, QUO: Quotation, V: Verb.

reference to the interlocutor with a personal pronoun meaning "you" is considered to be rude. The frequent uses of interactional particles like *ne*, *yo*, deictic verbs like *iku* (*to go*), *kuru* (*to come*) and honorific expressions fill the gap caused by the lack of personal pronouns.

3.2 NUCC compared with Books (BK)

The statistic measure: LLR demonstrates the degree of typicality for these 60 morphemes compared with the BK. Even if they are used in every conversation of the NUCC, their degree of typicality is not homogeneous. The most typical 10 morphemes relative to the BK with the highest degree of LLR and the least typical 5 with the lowest degree of LLR are shown in Table 3. The MPM indicates the number of morphemes per million.

no	Morpheme	Function	LLR	MPM
1	*un*	*Yeah, I see*	310,539	30,003
2	*ne*	TAGQ,	127,327	19,754
3	*tte*	QUO (contracted)	80,628	12,575
4	*ka*	Q	67,541	22,884
5	*teru*	PROG/PERF (contracted)	59,022	9,714
6	*sou*	*so*	51,485	11,024
7	*yo*	*I tell you*	44,561	9,790
8	*nani*	*what*	39,340	9,820
9	*keredo*	*although*	36,307	6,436
10	*a!*	INTJ	36,090	4,273
...	...	...	…	…
56	*suru*	*to do*	-2,899	14,343
57	*wa*	TOP	-4,030	25,419
58	*ni*	IO etc.	-4,301	29,498
59	*iru*	to exist, to be	-6,440	1,200
60	*wo*	ACC	-20,037	3,939

Table 3: Typical and atypical morphemes in the NUCC compared with the BK

We can easily see that interactional expressions and contracted forms are typical in face-to-face conversation. The backchannel *un* appears 30,000 times par million. This is 3% of the morphemes used in the NUCC. In contrast, the least typical 5 are indispensable grammatical morphemes in any Japanese utterance regardless of spoken or written. Negative value means that the morpheme is less used in the conversation than in books. In fact, the least typical morpheme with the lowest degree of the LLR, the accusative marker "*wo*" is often not pronounced in conversation.

3.3 NUCC compared with the IBB

We then compare the uses of these 60 morphemes in the NUCC with the IBB in order to show the difference in spoken and written interactional exchanges. These interactions are characterized by two points of view: social closeness and physical distance between two participants of communication.

3.3.1. Typical Morphemes

The most typical 10 morphemes of the NUCC compared with the IBB are following (LLR is in bracket).

1. *un* *yeah, I see* (324,691)
2. *da* DEC (159,975)
3. *ne* TAGQ, *you know* (146,670)
4. *no/n* GEN, QN or INTA[4] (108,044)
5. *ka* Q (101,483)
6. *sou* *so, in such a way* (95,564)
7. *tte* QUO (contracted) (85,429)
8. *ta* PAST (75,684)
9. *nani* *what* (67,687)
10. *iu* *to say* (61,961)

The high frequency of *da* (declarative marker) is noteworthy. Its occurrence seems to derive from the frequent use of short turn taking in face-to-face conversation, especially the large number of casual backchannel feedback finishing with "*da*", such as "*sou-na-n-da"* (*so*-DEC-QN-DEC, *"Indeed"*), whereas this is not the case in written correspondence on the Internet. The participants are not in real-time interactions in "Questions and Answers" type exchanges, so that the frequent use of short turn taking is not common. Also the participants of the IBB do not have a close relationship between them, because in fact they do not know each other and in general the written communication does not allow them to make intimate interactions in Japanese. These are the reasons for which the informal declarative form "*da*" is typical in the NUCC, whereas the formal one "*desu*" is numerous in the IBB.

3.3.2. Verb: *To Say* in the Conversation

Among the 12 verbs in the Table 1, "*iu*" (*to say*) is the most typical one of the NUCC with LLR: 61,961, followed by *iku* (*to go*, LLR: 20,919), *yaru* (*to do*, LLR: 17,603), *suru* (*to do*, LLR: 14,343), *kuru* (*to come*, LLR: 13,558), *aru* (*to exist*, LLR: 12,403), *omou* (*to think*, LLR: 10,903), *wakaru* (*to understand*, LLR:8,613), *naru* (*to become*, LLR: 5,970), *miru* (*to see, to look at*, LLR: 5,599), *dekiru* (*to be able to*, LLR: 1,489) and *iru* (*to exist, to be*, LLR: 1,200) in descending order. This metalinguistic verb *to say* is used much more often in oral conversation than in written correspondence. It may be explained at least partially by the fact that in real-time exchanges, we talk a lot about "how to say" something. The speaker leaves traces of metalinguistic activity in his speech. For example, when we hesitate in seeking an expression, we say: "*How should I say?*". In the example

[4] The occurrence of numerous "*no*" in conversation primarily comes from the frequent use of the interactional usage of this morpheme placed at the end of utterances. However there are also many "*no*" placed before the declarative "*da*" often realized "*n-da*". This frequently used bigram is often analyzed as a compound auxiliary in Japanese linguistics. This is not the case in this study, as to our morphological analyzer processes them as QN-DEC.

(1), having once used the word "*room*", the speaker corrects it with the word "*entrance*" while talking about the process of this correction: *heya-tte-iu-ka* (*Can-I say "room"?*). In this type of metalinguistic utterance, the verb: *to say* plays the main role.

(Ex.1) conversation 019

Gozenchu-wa zuutto heya-ni
morning-TOP throughout room-LOC
heya-tte-IU-ka genkan-ni haitte-ta-n-da
room-QUO-SAY-Q entrance-LOC
enter-PAST-QN-DEC
"I was in a room all morning, can-I SAY "room"?, in the entrance. "

In contrast, in the activity of writing, even private texts like those found in the IBB are prepared and elaborated. That would be why there is a big gap in the use of the verb: *to say* between the IBB and the NUCC.

4. Grammatical study: fragmentation

Finally, we will discuss how to end an utterance in Japanese conversation.

4.1 13 basic utterance-final morphemes in the NUCC compared with the BCCWJ

We analyze 13 morphemes employed at the utterance-final position in all 129 conversations of the NUCC. This position is defined by a period or a question mark in the transcription. We can consider these 13 items as the basic utterance-final morphemes in Japanese informal face-to-face exchanges. The Table 4 indicates that when compared with the BCCWJ, the most typical utterance-final morpheme of the NUCC is the interactional particle: "*ne*", while the least typical one is the auxiliary: "*ta* (Past Tense)".

These are classified into three groups. The first includes 4 final interactional particles (Final PRT): "*ne, yo, na, ka*". The second, 3 auxiliaries (AUX): "*da, nai, ta*" and the third, 6 sentence-internal conjunctive particles (PRT): "*te, keredo(kedo), tte, kara, de, ni*" as indicates the Table 4.

Of these three groups, the frequent use of interactional particles in conversation is entirely predictable. The normal position of these morphemes is at the end of utterances. The use of auxiliaries at the final position is also ordinary in every type of text. The most interesting phenomenon is the use of sentence-internal conjunctive particles at the utterance-final position. It is not normative in Japanese traditional grammar and absent in the written formal texts, while it is found in every conversation of the NUCC.

POS	morpheme	function	LLR
Final PRT	*ne*	TAGQ, Alignment	55,092
PRT	*te*	*and*	22,516
PRT	*keredo(kedo)*	*although*	14,129
PRT	*tte*	QUO	13,949
Final PRT	*yo*	*I tell you*	12,305
Final PRT	*na*	*I tell you, I know*	10,520
PRT	*kara*	*because*	7,526
PRT	*de*	*and*	6,583
Final PRT	*ka*	Q	6,329
PRT	*ni*	DAT, LOC, TEMP, ADVL	4,672
AUX	*da*	DEC	1,027
AUX	*nai*	NEG	270
AUX	*ta*	PAST	-7,774

Table 4: LLR of final morphemes of the NUCC compared with the BCCWJ

4.2 From sentence-internal particle to utterance-final particle or vice versa

We could say first that there are many syntactically incomplete sentences in Japanese conversation as in other languages[5] This could be due to the pragmatics of conversation: the participants of communication collaborate to finish a sentence as in example (2). The utterance of the speaker A stops at the end of the subordinate clause marked by an adversative conjunction *KEDO* (=*KEREDO* "*although"*). The speaker B completes A's utterance by adding the main clause.

(Ex.2) conversation 035

A: sensei-ni mikkahodo tomatte-morae-ba
professor-IO several days stay-make-if
ii-n-desu KEDO.
good-QN-DEC(formal) ALTHOUGH
"Although it would be better if we could ask the professor stay here for several days."
B: A! deki-nai-n-desu-ka.
ah can-NEG-QN-DEC(formal)-Q
"Ah, you can not do so."

However in most cases, this kind of collaboration between the participants of conversation is not obvious. The particle at the end of the utterance no longer has the conjunctive function linking the subordinate and main clauses but rather has a modal function. The example 3 shows that the utterance emitted by speaker B does not adversative with that of speaker A, despite the existence of *KEDO*. The function of *KEDO* in this case is to attenuate the assertive power of the predication and to show the intention of continuing the dialogue to the interlocutor (cf. Saegusa, 2007).

[5] Syntactic fragmentation does not necessarily correspond to informational fragmentation (cf. Matsumoto 2010).

(Ex.3) conversation 092

A: dou-iu-hanashi?
how-say story
"what story?"
B: tabun shi-ta-to-omou-n-da KEDO.
Perhaps do-PAST-QUO-think-QN-DEC ALTHOUGH
"Perhaps I have already spoken to you about. KEDO."
A: jaa, kika-nai-wa.
.so ask-NEG-PRT
"So I will not ask you."

In written normative texts, these morphemes have only one conjunctive function, while having two in conversational discourse.

This phenomenon could be viewed from a diachronic point of view. In Japanese, a SOV type language, particles are placed after their head, either conjunctives or interactionals. The resulting fragmentation can easily cause a functional and grammatical change in the role of particles. We could say first that these sentence-internal particles create new interactional functions in conversation. This is the direction from the norm to usages. However we could also point out the opposite direction: from usages to the norm in written texts. In standard written Japanese the interactional use of these particles may be put aside, while they always remain in conversation. Figure 1 indicates these two directions. This issue deserves a full review. It would be interesting to consider this question within the Macro-Syntaxe analytical framework (Blanche-Benveniste, 1990).

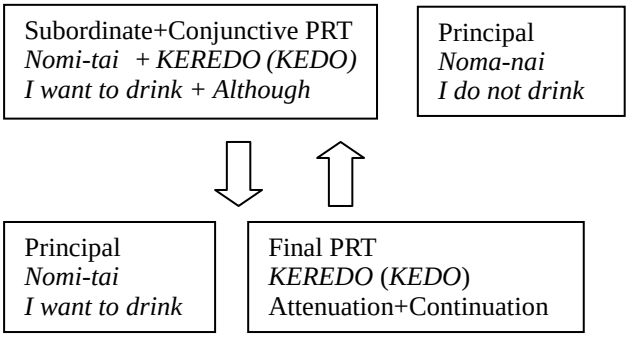

Figure 1: Linguistic change from sentence-internal PRT to utterance-final PRT or vice versa

5. Conclusion

Having compared the NUCC with the BCCWJ, several lexical and grammatical characteristics of Japanese conversation have been recognized.

1) 60 basic morphemes of spoken Japanese are identified. Personal pronouns are not included in the list. This is explained by the grammatical characteristics of the language.
2) Typical morphemes of conversation: interactional particles, interjections, markers of agreement and "*what*", reflect the involved nature of this activity, when compared with books.
3) The typical auxiliary of conversation, compared with written correspondence, is "*da* (declarative)". It may reflect the high frequency of short answers and backchannels in conversation.
4) The typical verb in conversation is "*iu* (*to say*)". This could come from frequent metalinguistic use of this verb in spontaneous speech, which, unlike written discourse, is not elaborated.
5) 13 basic utterance ending forms within conversation have been identified. Some of them are only used at the sentence-internal position in written texts. This is due to close and frequent exchanges between participants which cause incomplete utterances. In Japanese, because of its grammatical structure the fragmentation easily causes a functional and grammatical change in the role of particles.

Lastly, we summarize some of the features of conversational Japanese in contrast with written Japanese. It has more involved production, more metalinguistic and illocutionary traces. It also has more fragmented structures, which could cause a dynamic linguistic change. These are universal characteristics of spoken exchanges mentioned in Biber (1995), primarily due to the lack of time in real-time interactions (Biber, 2010) and secondarily to the closeness between two participants during exchanges. We also found some specific characteristics of Japanese conversation, like the absence of personal pronouns. This is explained only by the individual language structure.

6. Acknowledgements

This work was supported by MEXT/JSPS KAKENHI Grant Number (23520504).

7. References

Biber, D. (1995). Dimensions of Register Variation: A Cross-Linguistic Comparison. Cambridge: Cambridge University Press

Biber, D. (2010). Linguistic Styles Enabled by the Technology of Literacy. In M. Moneglia, A. Panunzi (Eds.), *Bootstrapping Information from Corpora in a Cross-Linguistic Perspective*. Firenze: Firenze University Press.

Blanche-Benveniste, Cl. (1990). *Le français parlé, Études grammaticale*. Paris: Editions du CNRS.

Maekawa, K. (2007). Design of a balanced corpus of contemporary written Japanese. In *Proceedings of Symposium on Large-Scale Knowledge Resources. (LKR2007)*, pp.55--58.

Matsumoto K. (2000). Japanese intonation units and syntactic structure, *Studies in Language*, 24(3), pp.515--564.

Ogiso, T., Komachi, M., Den, Y. and Matsumoto, Y. (2012). UniDic for Early Middle Japanese: a Dictionary for Morphological Analysis of Classical Japanese. In *LREC 2012 Proceedings*. Available at: <http://www.lrec-conf.org/proceedings/lrec2012/index.html>.

Saegusa, R. (2007). Usage of GA and KEREDO in Spoken Japanese. In *Center for Student Exchange journal*, 10, pp.11--27. Available at: <http://hdl.handle.net/10086/14360> (in Japanese).