New Steps in Japanese Studies
Kobe University Joint Research
edited by Kazashi Nobuo and Marcella Mariotti

On the Multiple Clause Linkage Structure of Japanese
A Corpus-based Study

Maruyama Takehiko
(Senshu University / NINJAL, Japan)

Stephen W. Horn
(NINJAL, Japan)

Kerri L. Russell
(Oxford Brookes University, UK)

Bjarke Frellesvig
(University of Oxford, UK)

Abstract In this paper, we will describe the distribution of the multiple clause linkage structure within actual spoken and written Japanese. We will examine three Japanese corpora: BCCWJ, CSJ and OCOJ. By identifying distributions of multiple clause linkage structures in corpora of contemporary Japanese (BCCWJ and CSJ), we shed light on what kinds of settings give rise to what type of clause linkage structures through what processes. The dynamic rewriting rule proposed by Kondo (2005) is introduced as a model for the incremental production of multiple clause linkage structures. Some common patterns of such structures occurring in Old Japanese are identified by OCOJ and compared to patterns in BCCWJ and CSJ.

Keywords Multiple Clause Linkage Structure. CSJ. BCCWJ. OCOJ. Incremental Production.

DOI 10.14277/6969-152-2/CFJS-5-7 | Submission 2015-11-04 | Acceptance 2016-02-12
ISBN [ebook] 978-88-6969-152-2 | ISBN [print] 978-88-6969-153-9 |

1 Introduction

1.1 Multiple Clause Linkage Structure in Japanese

In Japanese sentences, subordinate and coordinate clauses are marked by non-final predicates in isolation or predicates of various inflections followed by conjunctive particles. Clauses immediately following these may be main, subordinate, or non-final coordinate clauses. A structure in which a clause is grammatically linked to another is called a clause linkage structure (Van Valin 1984; Haiman, Thompson 1993; Hasegawa 1996).

The clause linkage structure is one of the basic sentence structures that are generally observed in various languages. What is characteristic of the clause linkage structure in Japanese (and perhaps SOV languages in general) is that sometimes extremely long clause chains are formed by the concatenation of clauses using a surprising variety of clause linkage markers.[1] An example of a long clause linkage structure that was observed in the Corpus of Spontaneous Japanese is given in (1).[2] The symbol "/" indicates a boundary between linked clauses.

1. *watashi ga sundeita tokoro wa danchi: no nikai deshite* / (F *e:to*) *sono mae wa ōkina* (F *e:*) *Meiji: dōro ga hashitte itandesu keredomo* / *danchi to dōro no aida niwa kō danchi no niwa mitaina kanjide* / (F *e:to*) *dōro no temae ni ki ga takusan haete ita node* / (F *e:to*) (F *ma*) *tori ga* (D *tsutsu*) *tobidashita to shitemo* / *sugu niwa dōro ni denaide* / *sono: ki* (D *n*) (D *ki*) *ki no atari ni hikkakatterukana: toyū* (F *e:*) *kanji deshita node* / *mazu nikai kara kō ki o ki no dono hen ni iruka toyū no o atari tsukete* / *atari o tukeru toyū ka* (F *ma:*) / *sagashite mite* / *sugu ni wa mitsukaranakatta node / shōganai* (D *gu*) *node* / (D *susu*) *sugu ni soto ni tobidashimashite* (CSJ:S02M0076)

1 To ask whether clauses concatenated in this way are in a relation of subordination ([+ dependent] and [+embedded]) or coordination ([-dependent] and [-embedded]) is to oversimplify, especially in the case of Japanese speech. The point is difficult to determine solely on the basis of predicate morphology since, in Japanese in general, the clause linkages |te| and |cont.| (see table 2 and discussion) can be used for either coordination or subordination. Supra-segmental factors have been shown to play a role in resolving the syntactic ambiguity of clause linkages in speech (Tyler 2012). In Japanese speech, some otherwise ambiguous linkages interpreted as coordinating can be deprived of their logical semantics. Yuasa and Sadock (2002, 92) group *te*-clause linkages that morphologically appear to be embedded but are logically independent under the category of 'pseudo-subordination'. Tsunoda (2013, 22) applies Olson's (1981) term 'co-subordination' to the type of *te*-clause linkages that appear in many of our examples here. The distribution of multiple clause linkage structures at least in part involves the question of how the concatenation of independent clauses differs from the parataxis of sentences.

2 Words enclosed in (F) indicate filled pauses and words enclosed in (D) indicate word fragments. A colon ':' indicates an elongated vowel. A final annotation such as (CSJ:S02M0076) indicates talk-ID in the corpus.

> Where I used to live was the second floor of an apartment, / and in front of that ran the great Meiji Avenue, / but taking the form of, like, this kind of apartment garden between the apartment and the street, / a lot of trees were growing on this side of the street, / so it was like, well, even if the bird rushed out from the room, / it would not go to the street directly, / but it must get hung up somewhere around the trees, / so at first from the second floor, approximating which tree where among the trees the bird might be, / or making an approximation of it, well, / I tried to look for it, / but I couldn't find it right away, / so there was nothing else for it, / so I went rushing out straightaway.

(1) is a part of a narrative describing an episode in which the speaker's bird had flown away and he tried to find it. The speaker enumerates a set of background conditions informing his choice of action and then describes a series of events, but there is not a single break in the speech marked by a finite sentence-final form. Instead, by repeatedly linking various clauses, the speaker produces an utterance that is potentially indefinitely long. Any structure containing more than one linkage of this type is referred to as a 'multiple clause linkage structure' in this paper.

1.2 Research Questions

Just as indefinitely long sentences can be generated by the recursive embedding of complement clauses, there is no limit to the length of sentences produced by the application of clause linkage. While from the point of view of prescriptive grammar extremely long clause linkage structures are to be avoided, very long utterances with multiple clause linkage structures do appear in spoken Japanese, as is seen in (1). Nevertheless, there has not been any research that quantitatively investigates and describes the extent to which this phenomenon can actually be found in different registers of Japanese.

Thus, this paper will quantitatively and qualitatively describe the multiple clause linkage structure in contemporary spoken and written Japanese. The following two research questions are posed:

> What types of clauses are connected to form the multiple clause linkage structure and in what order?
>
> What factors bear on the variable use of the multiple clause linkage structure?

In this paper, we describe the distribution of the multiple clause linkage

structure within actual spoken and written Japanese using corpora. In section 2 we introduce two large-scale corpora of contemporary spoken and written Japanese, the Corpus of Spontaneous Japanese and the Balanced Corpus of Contemporary Written Japanese. Section 3 quantitatively examines the distribution of clause boundaries appearing in the two corpora. In section 4 we discuss the factors influencing the distribution of the multiple clause linkage structure in various registers and consider the proliferation of these structures in Early Middle Japanese prose.[3] Then, we examine the Oxford Corpus of Old Japanese and discuss whether the same tendencies can be found between contemporary and Old Japanese.[4] Finally, we present some data from the late Meiji era that suggest that the convention of contemporary formal writing calling for avoidance of the multiple clause linkage structure has been established only very recently.

2 Data

2.1 Corpora: CSJ and BCCWJ

As seen in (1), the multiple clause linkage structure can appear in a situation where a single speaker continuously holds the floor, talking in a spontaneous manner. In spoken Japanese in general, monologues contain more multiple clause linkage structures than dialogues. This paper primarily uses the audio data of monologues compiled in the Corpus of Spontaneous Japanese (CSJ), which has more monologues than other Japanese corpora, as its object of analysis.

The CSJ is a corpus that contains 651 hours and 7.52 million words of spontaneous speech. The audio data can be classified into two categories: Academic Presentation Speech (APS) and Simulated Public Speaking (SPS). The APS is composed of live recordings of academic presentations in various academic societies. The SPS contains general speeches and comments by laypeople on everyday topics, speaking before small audiences. A relatively formal speaking style is observed in the APS, while a casual speaking style is observed in the SPS. Most monologues in the APS and the SPS are 10-15 minutes long (NINJAL 2006).

The monologues used for this study are from the richly annotated subcorpus called CSJ-Core, which takes 18.8 hours of data from APS and 19.9 hours from SPS. The audio is transcribed and analysed morphologically, and major clause boundaries are annotated with clause boundary labels. In this paper, units ending with an explicit grammatical form to show the

3 Early Middle Japanese is the language of 9th to 12th century Japan.

4 Old Japanese is the language of 7th and 8th century Japan.

point of completion of an utterance are identified as sentences, and labeled with |EOS| ('End Of Sentence'), as we will see in section 2.2.

The written language data that we compare with the monologues are extracted from the Balanced Corpus of Contemporary Written Japanese (BCCWJ). The BCCWJ is a balanced corpus of written Japanese that contains 100 million words in extracts randomly sampled from a wide variety of texts. Designed to include many registers such as books, magazines, newspapers, white papers, laws, verses, textbooks, and Internet documents, the BCCWJ makes it possible to look into what styles of written language occur in what type of registers (Maekawa et al. 2014).

The written samples we analyse are sentences taken from the richly annotated BCCWJ-Core out of the three registers of Books, Magazines and Newspapers. The text is analysed morphologically by an electronic dictionary, the *UniDic*, and sentence boundaries have been assigned in the process. In this paper, only the lines that end with the punctuation marks ('。'; '?'; '!') were extracted and recognised as sentences.

Table 1 shows the size of the data sets that are the object of analysis. Here, speech and writing are referred to as channels, while APS and SPS (from the CSJ), and Books, Magazines, and Newspapers (from the BCCWJ) are referred to as registers. The numbers of words are measured on the basis of an analytical unit designated in the corpus as the "Short Unit Word", which approximates the level at which entries in traditional Japanese dictionaries are identified (Maekawa et al. 2014).

Table 1. Data Statistics

Corpus	Registers	Files	Sentences	Words
CSJ (spoken)	APS	70	5,389	191,591
	SPS	107	4,494	164,096
BCCWJ (written)	Books	83	8,780	204,050
	Magazines	86	9,342	202,268
	Newspapers	340	11,898	308,504

2.2 Clause Boundary Labels

In order to analyse the form and distribution of clause linkage structures, it is necessary to know what sorts of clauses appear in the text and the order in which they are connected. As a first step, we associate the morphology of rightmost clause boundaries with Clause Boundary Labels (CBLs). At the point at which we began our investigation, CBLs identifying 49 types of clause boundaries had already been annotated to the CSJ. As for the BCCWJ, a clause boundary analysis was newly conducted, using a Clause Boundary Analysis Program (CBAP) that detects rightmost clause boundaries and annotates 147 varieties of CBL (Maruyama et al. 2004).

Examples of texts annotated with CBLs are shown below. (2) is an example from SPS, and (3) is taken from Newspapers. The CBLs are assigned various names, such as |EOS|, |ga|, |node|, and so on, depending on the type of clause boundary identified. The CBLs involved in forming multiple clause linkage structures are fully listed in table 2 below.

2. *rukura no mura nandesuga* |ga| / *hikōjō wa hontō ni yama no naka ni arimashite* |te| / *shikamo hikōjō ga jarimichi toyū tokoro deshita node* |node| / *jissai ni chakuriku suru toki wa hontō ni shinzō ga tomarisōni nattandesu keredomo* |keredomo| / (F *e*) *tsuite mitara* |tara| / *igai ni sōitta shukuhaku shisetsu nado de* (F *e:*) *nigiwatta chīsana mura deshita* |EOS|/ (CSJ:S01F0151)
   When it comes to the village Lukla, / the airport really being located in the deep mountains, / and moreover, because it was a situation where the airport was gravel-paved, / when we actually touched ground my heart really almost stopped, / but once we arrived there, / to our surprise, we found that it was a small village flourishing with that sort of hotel and such. /
3. *isuraeru kara no hōdō ni yoruto* |to| / *dōkoku saidai no toshi teruabibu de jūgonichi yoru paresuchina jichiku gaza kara no isuraeru gun tettai to yudaya jin nyūshokuchi tekkyo o motomeru* |adnominal| / *shūkai ga hirakare* |continuation| / *jūgoman nin ijō ga sanka shita.* |EOS| / *saidai yatō no rōdōtō nado no yobikake ni yoru* |adnominal| / *mono de* |de| / n*isen'nen aki ni paresuchina funsō ga gekika shite irai* |irai| */ saidai no kibo to natta.* |EOS| / (BCCWJ:PN4f_00018)
   According to the report from Israel, / on the night of the 15th there was a gathering in Tel Aviv, the largest city in Israel, / to demand the withdrawal of Israel army and removal of Jewish settlement from the Palestinian Territory of Gaza, / and more than 150 thousand people participated. / It was an event / promoted by the largest opposition party, the Labour Party, / and was the biggest gathering since / the Palestinian conflict intensified in the autumn of 2000. /

In order to compare the results of the clause boundary analyses from the two corpora, we will limit the types of clause boundaries to high frequency clause-linking items common to both corpora. Thus, this paper took the CBLs shown in table 2 as marking the main clause boundaries that appear in multiple clause linkage structures and used them as the basis of comparison.

Table 2. Clause Boundary Labels Used in Multiple Clause Linkage

Types	Clause Boundary Labels
EOS	EOS
Coordinate	ga, keredomo, keredo, kedomo, kedo, shi
Reason	kara, node
Conditional	tara, taraba, nara, naraba, to, reba
Misc.	cont. (continuation), de, te, quote, toyu

The continuation CBL indicates the rightmost boundary of a clause ending with the conjunctive form of a predicate. The *de* CBL indicates the rightmost boundary of a clause ending with a non-finite form of the copula *da*. Furthermore, although the *toyu* CBL does not indicate clause-linking morphology but rather indicates a special kind of complementizer at the rightmost boundary of a relative clause, *toyu* CBLs have been included in the comparison because they appear to participate in a recurring pattern with clause-linking CBLs (discussed in detail in § 3.4).

Using the CBLs extracted from CSJ and BCCWJ, we conducted an analysis of their quantitative distributions and combinations in multiple clause linkage structures. The following section examines (1) the frequency of clause boundaries across different registers, (2) the number of clause linkages per sentence, (3) frequently occurring patterns of clause linkage, and (4) 'highly complex clause linkage structures'.

3 Analyses

3.1 The Frequencies of Clause Boundaries across Registers

The first analysis concerns the types of clause boundaries and how frequently they appear within each register. Tokens for each type of CBL that appeared in each register were totalled, and the totals for each were normalised to the rate of instances per 200,000 words per register. Table 3 shows the results.

Table 3. Frequencies of Clause Boundaries (per 200,000 words)

	CBL	APS	SPS	Books	Magazines	Newspapers
EOS	EOS	5,624	5,476	8,606	9,237	7,713
Coordinate	*Ga*	1,027	672	716	552	496
	Keredomo	382	800	14	2	1
	Kedomo	108	328	0	0	0
	Keredo	8	37	15	26	4
	Kedo	43	584	26	62	10
	Shi	54	230	108	90	21
Reason	*Kara*	78	261	307	185	69
	Node	310	735	150	164	40
Conditional	*tara(ba)*	60	303	184	172	29
	To	546	691	438	365	265
	nara(ba)	3	9	42	53	11
	Reba	153	225	450	288	178
Misc.	Continuation	556	277	1,908	1,837	2,023
	De	347	769	448	408	408
	Te	2,884	3,903	2,122	1,625	1,080
	Quote	1,006	1,577	1,130	732	881
	Toyu	1,454	1,163	445	267	150
Total		14,645	18,038	17,110	16,065	13,379

First, from the total number of clause boundaries in the CSJ, we see that there were more instances of clause boundaries in SPS (18,038) than in APS (14,645). On the one hand, since highly spontaneous speaking style (as characterised by factors other than clause linkage) is observed more in SPS than APS, the tendency for non-final clause boundaries (and multiple clause linkage structures) to appear more frequently in SPS can be interpreted as being correlated to highly spontaneous speech. In particular, among the clause boundary types that are frequently used in multiple clause linkage structures, SPS has 2 to 13 times more frequent instances than APS of the following types: *keredomo*, *shi*, *kara*, *node*, *tara*, and *de*-clauses. On the other hand, *ga*-clauses and continuation clauses appear 1.5 to 2 times more frequently in APS. This shows that these types of clauses are preferable in formal speaking styles.

From the total number of clause boundaries in the BCCWJ, we see that there were 13,379 in Newspapers, 16,065 in Magazines, and 17,110 in Books. Taking the fact that sentence boundaries (EOS) are fewer in Newspapers (7,713) – and thus that sentences in Newspapers are longer – than those in the other written registers, together with the fact that the total number of clause boundaries is also fewer in Newspapers than in other written registers, constructions other than clause linkage are clearly favored to extend the content of sentences in Newspapers. We surmise that what allows newspapers to have longer sentences (among written regis-

ters) without using proportionately many clause linkage structures is their tendency to employ relatively saturated argument structures filled with complex noun phrases. The highest number of continuation clauses among the three registers is found in Newspapers, which can be understood as a characteristic of the formal writing style.

When CSJ and BCCWJ are compared, the number of EOS is far smaller in CSJ, indicating that the average sentence length is greater in speech than in writing. Clause linkage also appears more often in spoken language than it does in written language. A comparison of the conjunctive forms across CSJ and BCCWJ shows that *keredomo*-clauses and *te*-clause are more frequent in the spoken language while continuation clauses are more frequent in the written language. This shows that there is a difference between the written and spoken language in their preference for the conjunctive forms creating clause linkage structures.

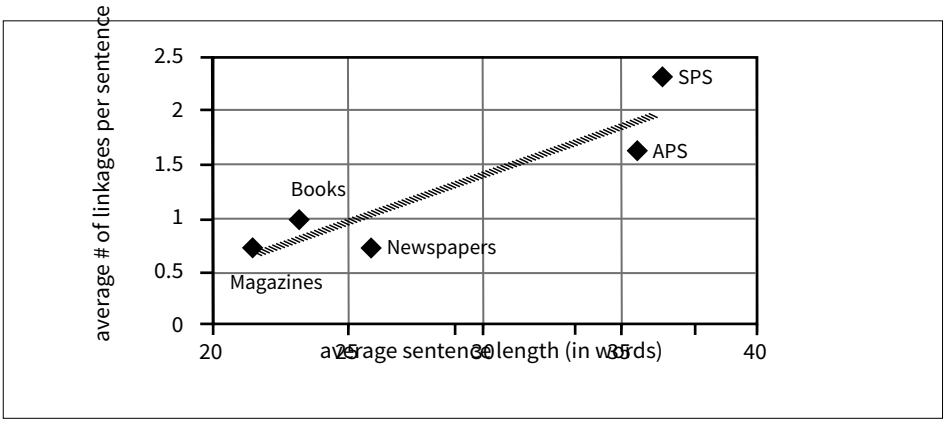

Figure 1. Average Sentence Length and Number of Clause Linkages

Generated from the statistics in table 3, figure 1 compares the average number of non-finite clause linkages with the average sentence length for each register. On average, as noted above, the registers in the speech channel show much longer sentences than those in the writing channel. With the exception of Newspapers, there is a clear trend for the number of clause linkages to increase in direct proportion to sentence length, but sentence length is obviously not the only factor. While the average number of linkages is well above the trend for the most casual-style register (SPS), the most formal register in the written channel (Newspapers) and the most formal register in the spoken channel (APS) both fall below the mean. We will discuss the effects that differences in channel and style have on the distribution of multiple clause linkage structures in more detail in section 4.

3.2 The Number of Sentences in Registers by Complexity

The second analysis groups sentences based on the number of CBLs in each (i.e. their complexity) and compares the frequency of instances from each group across registers. The number of CBLs for each register was normalised to the number of instances per 10,000 sentences. Figure 2 shows the results. The horizontal axis shows the complexity of the sentences of a given group; the vertical axis shows the frequency of instances for each group.

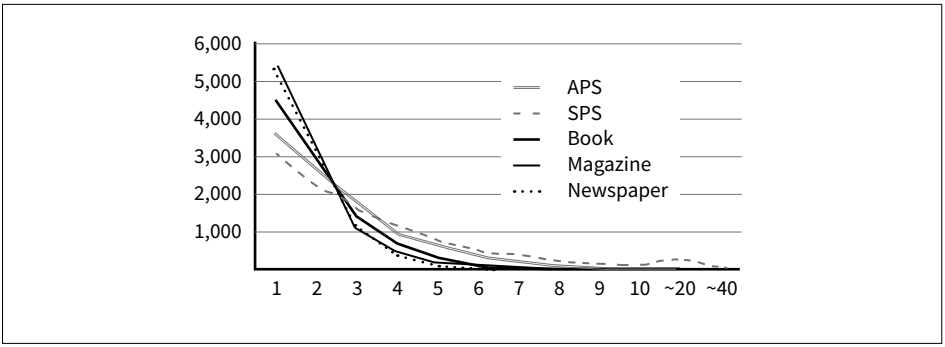

Figure 2. Number of Sentences in Registers by Complexity (out of 10,000 sentences)

figure 2 shows that the curves of the plots for APS and SPS (the CSJ registers) are clearly more gently sloped than those for the remaining registers from the BCCWJ. The data for Magazines and Newspapers have a nearly identical locus, where for each of these approximately 5,300 out of 10,000 sentences are simple sentences that do not contain a clause boundary except for the EOS. Sentences in the spoken language that only had EOS clause boundaries numbered approximately 3,200 for APS and 2,600 for SPS. In figure 2, the frequency of instances for written and spoken language is reversed at the point where the number of clause boundaries found in a sentence rises to 3. At subsequent values, the graph shows a long tail shape, especially for SPS. Based on this distribution, we can see that very long multiple clause linkage structures occur particularly often in casual, spontaneous speaking style. We will refer to structures of this type as ‘highly complex clause linkage structures’ (discussed in detail in § 3.4).

3.3 Frequently Occurring Patterns of Clause Linkage

Now we will examine the linkage pattern of the clause boundaries. In both written and spoken language, what types of clauses are connected and in what order?

Clause linkage patterns can be represented as ordered sets of non-final clause-linking CBLs. Accordingly, examples (2) and (3) have, respectively, the following types of clause linkage pattern:

(2') *ga* _ *te* _ *node* _ *keredomo* _ *tara* _ EOS

(3') *to* _ cont. _ EOS *de* _ EOS

Below we will look at clause linkage patterns that occur with the greatest frequency. The patterns of clause linkage found in each register were totalled and the percentages of these appearances over the total number of sentences were calculated. Tables 4 and 5 show the 10 top-ranking items in each register.

Table 4. Frequent Clause Linkage Patterns (CSJ)

APS		SPS	
32.0%	EOS	26.2%	EOS
9.1%	*te* _ EOS	5.9%	*te* _ EOS
3.8%	*toyu* _ EOS	3.6%	quote _ EOS
3.3%	*ga* _ EOS	1.8%	*toyu* _ EOS
2.7%	quote _ EOS	1.7%	*te* _ quote _ EOS
2.2%	*te* _ *te* _ EOS	1.6%	*de* _ EOS
1.9%	verb _ EOS	1.4%	*keredomo* _ EOS
1.5%	*te* _ *toyu* _ EOS	1.4%	*to* _ EOS
1.5%	*de* _ EOS	1.4%	*te* _ *te* _ EOS
1.5%	*te* _ quote _ EOS	1.2%	*ga* _ EOS

Table 5. Frequent Clause Linkage Patterns (BCCWJ)

Books		Magazines		Newspapers	
45.1%	EOS	53.7%	EOS	52.2%	EOS
7.8%	*te* _ EOS	7.8%	cont. _ EOS	11.5%	cont. _ EOS
6.4%	cont. _ EOS	6.3%	*te* _ EOS	5.6%	*te* _ EOS
3.5%	quote _ EOS	2.9%	quote _ EOS	3.6%	quote _ EOS
2.4%	*ga* _ EOS	2.4%	*ga* _ EOS	2.7%	*ga* _ EOS
1.8%	*to* _ EOS	2.0%	*de* _ EOS	2.6%	*de* _ EOS
1.8%	*de* _ EOS	1.7%	*to* _ EOS	1.4%	*to* _ EOS
1.6%	*reba* _ EOS	1.3%	*te* _ cont. _ EOS	1.2%	cont. _ cont. _ EOS
1.2%	*toyu* _ EOS	1.3%	*reba* _ EOS	1.1%	cont. _ *te* _ EOS
1.2%	*kara* _ EOS	1.0%	*toyu* _ EOS	1.0%	*te* _ cont. _ EOS

As we have seen in figure 1, the most frequent clause linkage pattern for all registers was the simplex sentence, with no clause boundary marker other than EOS. This includes not only very short simple sentences or fragmentary sentences (like those in (4)), but also relatively long sentences (like those in (5)).

4. a) *kore de owarimasu* |EOS| / (CSJ:S01F0050)
   We will end here. /
   b) *JRBungoTakedaekikaratohonanafun.*|EOS|/(BCCWJ:PM41_00182)
   It's 7 minutes on foot from the JR Bungo Takeda station. /

5. a) *mekishiko jin to shite wa son'na mazushī katainaka no machi yori mo motto ōbei nami ni hattatsu shita rizōtochi ya mekishiko shitī no naito raifu o an'nai shitakatta yō deshita* |EOS| / (CSJ:S00F0173)
   It seemed that, as a Mexican, he would have preferred to show me a resort area developed on a par with Western countries, or the nightlife in Mexico City, rather than such an impoverished town in the hinterlands. /
   b) *kugatsu no mekishiko kankun no sekai bōeki kikan* (WTO) *kakuryō kaigi no ketsuretsu ikō nikokukan kōshō ya chiiki jiyū bōeki kyōtei nado daitai kyōtei eno ugoki ga tsuyomatteiru* |EOS| / (BCCWJ:PN3b_00007)
   Since the breakdown of the WTO Ministerial meeting at Cancun, Mexico in September, the movement toward alternate agreements such as bilateral negotiations or regional FTAs is becoming stronger. /

Looking at items ranked 2nd place and lower in tables 4 and 5, we see that the '*te* _ EOS' pattern is ranked at 2nd or 3rd place for all the registers in both written and spoken language. This combination is likely the most basic simple clause linkage pattern. Furthermore, the three written registers share the following patterns ranked higher than 7th place: 'cont._ EOS', 'quote _ EOS', '*ga* _ EOS', '*de* _ EOS', and '*to* _ EOS'. We consider these patterns to be the basic clause linkage structures for the written language in general.

3.4 The Distribution of Highly Complex Clause Linkage Structures

All the examples of common clause linkage patterns that we have seen so far consist of fairly small sets of clause linkages forming rather simple structures. The more complex a clause linkage structure becomes, the more diversified its patterns become, so that these results do not appear high up in the order of frequency.

For each of the registers we analysed, the number of sentences that have more than 6 clause boundaries (our criterion to identify sentences with highly complex clause linkage structures) was normalised to the number of sentences per 10,000 sentences and totalled (which corresponds to the right half of figure 1). The complementizer *toyu* was also counted for this study. table 6 shows the results.

A comparison of the written and spoken language reveals that highly complex clause linkage structures occur more often in spoken language.

In particular, sentences with more than 11 clause boundaries are far more frequent in SPS than in other registers. The sentence with the most complex clause linkage structure has 34 linked clauses (this appears in SPS). Table 6 shows that highly complex clause linkage structures appear most frequently in casual speaking style.

Table 6. Number of Sentences in Registers by Complexity (2)

	x6	x7	x8	x9	x10	x20	x40
APS	321	186	95	41	17	37	0
SPS	465	367	185	116	78	220	11
Books	95	49	10	10	5	0	0
Magazines	54	16	5	3	2	1	0
Newspapers	35	4	2	2	0	1	0

Below, a highly complex clause linkage pattern is picked out and examined. Examples of a clause linkage pattern found in APS are shown in (6). The pattern was '*ga* _ *te* _ *te* _ *toyu* _ EOS' (complexity measure = 6).

6. a) *de: roku kyū jūni to* (D *yu*) *arimasu ga* |ga| / (F *e:*) *kore wa kono sūchi wa hikui hō no fukugō'on no mottomo takai shūhasū seibun o arawashite mashite* |te| / (F *ma:*) *sūchi ga fueru ni shitagatte* |te| / *baion seibun no kosū ga fueru toyū* |toyu| / *koto ni natte imasu* |EOS| / (CSJ:A01M0056)
   Now, here are (examples) six, nine, and twelve, / but as for these, their values show the highest frequency components of the lower compound tones, / and as the values increase, / the situation is that / the number of individual overtone components also increases. /
   b) *kore wa futatsu no bunsetsu no aida no kakariyasusa o motomeru tame no shuhō nanodesu ga* |ga| / (F *e:*) *omona tokuchō toshite* |te| / *bunpō to hyūrisutikku o mochiite* |te| / *kanō na kakari saki o sēgen suru toyū* |toyu| / *gihō o mochiite imasu* |EOS| / (CSJ:A03M0010)
   This is a method to calculate the degree of dependency between two *bunsetsu* phrases, / and as its main feature, it uses a technique in which, / using grammar and heuristics, / it restricts the possible targets for dependency. /

Including the two sentences in (6), the same clause linkage pattern was found in eight sentences in the APS and in five sentences in the SPS. It is conceivable that there are particular clause linkage patterns that are used as syntactic frames to produce ongoing speech, especially in the register of formal speaking style in academic presentations. Whether the distribution of this specific pattern rises to the level of statistical significance is a question we set aside for further research.

4 Discussion

4.1 Why does the Multiple Clause Linkage Structure occur?

Prompted by the quantitative analysis that was conducted in the preceding section, some thought should be given to the question of why the multiple clause linkage structure occurs more often in spoken language than it does in written language, and most often in spontaneous speech. In this section we will discuss this issue.

First, comparisons both within the channel of writing and within the channel of speech show that the multiple clause linkage structure appears much more frequently in casual style than it does in formal style. The effects of formality suggest that variability in the planning, monitoring and repair that go into the production of a communicative act may be a factor in the use of multiple clause structures. The register of APS is more likely to contain carefully prepared material than is the register of SPS, spontaneous casual-style monologue. There are a greater expectation of linguistic adeptness and an accordingly greater cost associated with error, so planning and self-monitoring take a high priority and repairs are avoided. Similar considerations go into writing and editing in the register of Newspapers (dedicated in large part to presenting objective fact) to a greater degree, on average, than in Books and Magazines, which include more writing for entertainment. Variability in patterns of discourse correlates with change in the level of formality at least in part due to these basic factors. We will address why this is manifested specifically in differences in the frequency of multiple clause structures in the discussion later in this section.

Comparing these two channels, we see a tendency for the multiple clause linkage structure to appear more often in speech than in writing. This is conceivably attributable to differences in the same set of basic factors: planning, monitoring and repair. As for the great range of difference in tendencies across registers, we surmise that the most crucial factor is to be found in the requirements of the speech event of casual-style monologue.

The reasoning with regard to speech is as follows: continuous speech in a conversational setting involves the simultaneous processing of information, syntax, phonology and articulation (or “incremental production” [Levelt 1989, 25]), which puts limits on the degree to which planning and self-monitoring can be carried out. Furthermore, incremental production requires that repairs be effected within the flow of speech. However, the conversational setting allows the speaker to 1) monitor the listener, 2) receive and respond to feedback, and 3) use suprasegmental resources, kinesics and other affective resources to supplement vocabulary and syntax. If the speaker detects a failure in the communicative process, the speech situation allows repairs in the flow of speech to be made as necessary with

little or no cost: disfluencies (Cutler 1982; Levelt 1989; Gilquin, De Cock 2013) are a natural part of speech. In addition, the fact that speaker and listener share the same temporal and spatial location allows a speaker to assume a great deal about common knowledge and orientation. The listener is expected to 1) cooperate by accommodating errors and presuppositions, 2) actively search for relevance or steps of inference when information is presented without logical connectives, and 3) give constant feedback about attention and comprehension. Refusal to cooperate in the production of a speech event can trigger implicatures or incur great social costs. Add to this the fact that in a casual-style monologue a speaker is expected to 1) hold the floor for an extended period of time and 2) exhibit cultural and linguistic adeptness. For the question we are addressing, these are the most relevant specific factors.

The factors in the speech situation manifest themselves in the lexis and the syntax of the texts produced in speech and the multiple clause linkage structure is one such manifestation. Restrictions on planning increase the difficulty of ordering the presentation of information and selecting appropriate logical connectives, but the cooperation of the listener can be enlisted to accommodate a lack of overt logical connection and seek relevance in the discourse. In such a case, the discourse could just as well be presented as a series of finite sentences without any connectives (a format which, for convenience, we will call 'parataxis'). However, in face to face interaction, the sentence-final pitch lowering and pause that typically accompany finite sentence boundaries can be interpreted as a turn-taking cue or can trigger other implicatures, while the nature of monologue requires the speaker to skilfully hold the floor and actively advance the exposition of ideas. One solution is the use of multiple clause linkage structures, frequently including semantically under specified clause boundary morphology. In fact, it has been shown that in the SPS, filled pauses (in general, linguistic devices often used to gain time for macro-planning during extended turns) are longer if they follow coordinating clause linkages (Watanabe et al. 2015) and shorter if they follow EOS and adjunct clause linkages. We take this as indirect indication that the multiple clause linkage structure is used as a floor-holding device in spontaneous monologues.

To illustrate with some concrete data, consider the first few clauses of example (1) reproduced below:

7. *watashi ga sundeita tokoro wa danchi: no nikai deshite* |te| / (F *e:to*) *sono mae wa ōkina* (F *e:*) *Meiji: dōro ga hashitte itandesu keredomo* |keredomo| / *danchi to dōro no aida niwa kō danchi no niwa mitaina kanjide* |de| / (F *e:to*) *dōro no temae ni ki ga takusan haete ita node* |node| / (F *e:to*) (F *ma*) *tori ga* (D *tsutsu*) *tobidashita to shitemo* |temo| / … (CSJ:S02M0076)

> Where I used to live was the second floor of an apartment, / and in front of that ran the great Meiji Avenue, / but taking the form of, like, this kind of apartment garden between the apartment and the street, / a lot of trees were growing on this side of the street, / so it was like, well, even if the bird rushed out of the room /…

The first few lines of the narrative take the following abstract form: *topic + background information* |te| *background information* |keredomo| *topic +* [*manner adjunct* |de|] *background information* |node| [*hypothetical condition* |temo|]… Clause boundary *te* is syntactically coordinating and in (7) it links two clauses that each give background information and are logically independent of each other (although the scope of the initial topic appears to extend over both of these clauses). While *te* (with flat intonation and reduced vowel) can be used to set up a realis condition (such as logical condition, causal or temporal priority, etc.) or to form manner adjuncts, it is not used as a logical connective or adjunct connective here. The clause boundary *keredomo*, which is syntactically coordinating and normally used to mark a concessive relationship with a following clause, ends the second clause, but it too is not used as a logical connective here. Instead, it links to another (complex) topic/comment clause containing yet more background information. A subordinate clause with the clause boundary *de,* which modifies the predicate *haete ita* (were growing), immediately follows the third clause's topic *danchi to dōro no aida niwa* (as for in between the apartment and the avenue). The clause that this predicate heads has an unambiguous coordinating clause boundary *node* that is used as a logical connective ('so') to link to a clause embedding a hypothetical concessive condition.

In short, the first three clauses are independent of each other, linked with semantically underspecified clause boundary morphology – which is not used to indicate any logical connection – but all mutually relevant in setting a scene in a narrative. To enumerate background circumstances in conversation, parataxis is sometimes used together with a pattern of rising final intonation and vowel elongation (similar to that used in reading off items in a list). But the non-finite nature of the clause boundaries *te* and *keredomo* that we see at the beginning of (7) is particularly well-suited for monologue because 1) the use of clause linkage strongly implies relevance between clauses, and 2) the nonfinite nature of the linkages helps the speaker hold the floor by signaling that the speech is not finished.

Let us contrast this to formal writing styles, where clause linkage of the type seen in (7) appears much less frequently and finite sentences are often overtly marked with initial conjunctive expressions that specify the relation with the preceding sentence. Why do we not find large numbers of sentences with underspecified non-finite clauses in formal writing? Writing has none of the online time constraints imposed by incremental production

in speech. There is time to plan and self-monitor. Furthermore, neither errors nor repairs can be made in the flow of writing without cost. Moreover, there is no opportunity to monitor the reader for comprehension and attention and little basis for assuming shared knowledge and orientation, so writing needs to be organised to explicitly set out presuppositions and to overtly mark logical steps and relevance. The reader can give no feedback and faces no immediate social cost for refusing to cooperate. This sets a sharp limit on the degree to which a writer can expect a reader to accommodate loose composition, accordingly making coherence and cohesion a necessity, at least in writing for purposes other than entertainment. But most importantly, in writing there is no need to signal that an utterance is not complete in order to gain time to process a subsequent utterance, because holding the floor is not an issue.

So with regard to writing for practical purposes, the combination of factors outlined above imposes the need for an efficient order of exposition, explicit steps of inference and the avoidance of ambiguity. With regard to the last point, note that the use of simple finite clauses 1) clarifies the scope of topics, quantification, negation, and modality, and 2) limits the candidates for adjunct modification, both of which reduce syntactic ambiguity. Furthermore, there is no necessity to hold the floor, and accordingly less need for non-finite linkage of independent clauses. And finally, especially with regard to some of the most commonly used clause linkages in speech such as *te* and *keredomo*, note that in writing there is no possibility to distinguish between their use as underspecified coordinating linkages (which, in speech, are frequently articulated with rising intonation plus elongated final vowel) and their use as adjunct linkages (which, in speech, are frequently articulated with flat intonation and reduced final vowel). We anticipate that the use of intonation phrases, kinesis, gaze etc. may play a big role in the difference between the structures of speech and writing. Learning not to rely on these resources is part of learning how to write.

It is worth noting that there is a well-established proscription against the use of multiple clause linkage structures within Japanese prescriptive grammar. Pupils in elementary school are apt to write compositions including long sentences with multiple clause linkage, apparently carrying their habits of speech over to another channel. Their teachers typically instruct them to split these long sentences into groups of simple sentences, adding conjunctives to clarify the relationships between them. The rationale is that lengthy sentences give to readers the impression of sloppiness and pointlessness and, thus, short sentences with conjunctives are preferable.

We do not intend our observations about the exigencies of writing and formality to be a justification for the conventions of prescriptive grammar. While multiple clause linkage structures are exploited and facilitated in speech in ways that do not apply to writing, in theory there is nothing preventing their use in writing, including formal prose. Logical conjunctives

can appear just as easily at the beginnings of coordinated clauses as they can in sentence-initial positions; syntactic ambiguities made possible by extended structures can be avoided by careful planning and self-monitoring; clauses with saturated argument structures can be linked with little risk of violating constraints on coordination. In fact, there are historical reasons to believe that the avoidance of multiple clause linkage structures in prose is more a convention of contemporary formal style than a natural consequence of high levels of planning, self-monitoring and avoidance of repair. We will return to this point in section 4.4.

4.2 The Dynamic Production of the Multiple Clause Linkage Structure

Kondo (2005) proposes a model for the process of producing multiple clause linkage structures to account for their frequent occurrence in the prose of Early Middle Japanese. In his model, a main clause modified by a preceding subordinate clause is rewritten to become another subordinate clause, which in turn modifies a following clause.[5] This rewriting process is driven by what he calls "dynamic rewriting rules". According to this framework, the multiple clause linkage structure can be understood as the result of repeated applications of the dynamic rewriting rule.

Kondo (2005) exemplifies the stages of dynamic rewriting using an example from the fifth chapter of *The Tales of Ise*. We will apply his reasoning to an extended example here. As the structure develops from (8a) to (8f), the portion that constituted the main clause is dynamically rewritten into a subordinate clause that modifies the subsequent clause. This captures the process by which a chain of clauses becomes longer. The underlined parts in (8) below are main clauses that are rewritten into subordinate clauses in the next stage.

8. a) [*hito sigekumo aranedo*] + *tabi kasanari keri*
   b) [*hito sigekumo aranedo*] + *tabi kasanari kereba*] + *aruji kikituku*
   c) [[*hito sigekumo aranedo*] + *tabi kasanari kereba*] + *aruji kikitukete*] + *sono kayohiji ni yogoto ni hito wo suetu*
   d) [[[*hito sigekumo aranedo*] + *tabi kasanari kereba*] + *aruji kikitukete*] + *sono kayohiji ni yogoto ni hito wo suete*] + *mamorasekeri*
   e) [[[[*hito sigekumo aranedo*] + *tabi kasanari kereba*] + *aruji kikitukete*] + *sono kayohiji ni yogoto ni hito wo suete*] + *mamorasekereba*] + *ikedomo e awadu*

5 Again, the issue of whether clauses thus concatenated are in a relation of subordination or coordination, and whether they are semantically dependent or independent, needs to be addressed with a detailed analysis.

f) [[[[[*hito sigekumo aranedo*] + *tabi kasanari kereba*] + *aruji kikitukete*] + *sono kayohiji ni yogoto ni hito wo suete*] + *mamorasekereba*] + *ikedomo e awade*] + *kaerikeri*
Although there weren't many people, because it kept happening, the householder heard about it, and because he placed someone on that route every night, making them keep guard, (the protagonist) being unable to meet (her) – go as he might – went back (home).

It is no great leap to suppose that the multiple clause linkage structure for spoken language is attributable to a similar production process. The question we need to address is why so many multiple clause linkage structures should be found in a written genre. Kondo (2005) cites Sakakura (1975) who suggests that the nature of Classical Japanese literature as rooted in oral tradition may be a factor bearing on the prevalence of extended clause linkages structures in that register. The earliest attested writing of full texts in Native Japanese (i.e. *wabun* as opposed to *kanbun*) records songs from oral tradition and original songs. Some of the earliest examples of Native Japanese prose also record tales (*monogatari*) from oral tradition or texts meant to be recited to an audience. Many of the texts of Early Middle Japanese that Kondo analyses are also called *monogatari* and some (including *The Tales of Ise*) have their origins in oral tradition. Finally, while many of the texts are diaries, the practice for reproducing and disseminating these texts was for one person to read them aloud while another person transcribed, and this may have had an influence on the form that the discourse takes.

Kondo's findings prompt us to sample records of Native Japanese at various points in the history of the language. While at this point we cannot offer a compelling explanation for the distribution of this pattern of clause linkage, in Early Middle Japanese, a wider historical view suggests that the pattern was the norm for most of the history of the Japanese language.

4.3 Application to Old Japanese

Next, we will compare the distribution of clause boundaries between Contemporary Japanese and Old Japanese as a case study extending the same analytic techniques to a wider range of text. Applying the same procedure of clause boundary analysis to a corpus of Old Japanese in the 7th and 8th centuries, we will examine whether the same tendency of clause linkage structure can be found between Contemporary and Old Japanese.

We use the Oxford Corpus of Old Japanese (OCOJ), a highly organised and annotated corpus of Old Japanese (Frellesvig et al. 2010). The ongoing corpus building project is hosted at the Faculty of Oriental Studies (University of Oxford). The OCOJ contains almost all of the extant main texts

from the Old Japanese period, including seven poetic collections and two non-poetic collections, with a total of approximately 111,000 words. The main features of the OCOJ are that the texts in the corpus are annotated to a high level of specificity, with full lemmatisation, morphological information, syntactic constituency, and core grammatical and semantic roles.

We examined a non-poetic text, *Shoku Nihongi Senmyō*, which contains a total of 14,306 words in 512 sentences. A total of 3,121 clause boundary labels were manually annotated to the text. (9) shows an example of the annotated text.

9. *koko wo motite* |te| *nori no pumi ni nosetaru wo ato tosite* |te| *nori no mani mani nagaku topoku ima wo pazimete* |te| *tugitugi ni tamapariyukamu* |adnominal| *mono zo to* |quote| *pepito iti tamapaku to noritamapu opomikoto wo kikitamapeyo to* |quote| *noritamapu* |EOS|(*Shoku Nihongi Senmyō* 2, 707)
   It says, listen to the great words that tell the fact that hereby, with what is written in law as precedent, as something that shall be received for a very long and distant time from now according to the law, I bestow (upon you) five thousand families.

Figure 3 shows the result of the comparison between SPS, Books and *Senmyō* text. The frequencies were normalised to 20,000 words in each register.

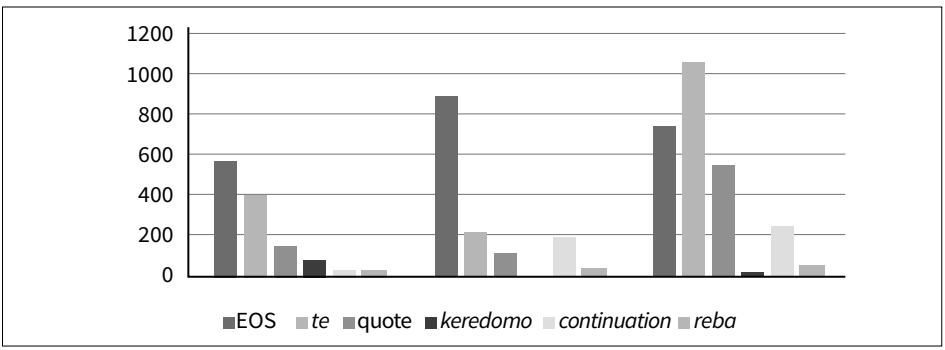

Figure 3. Frequency of Clause Boundaries (per 200,000 words)

As figure 3 shows, the frequency of *te*-clauses in *Senmyō* is much higher than that in SPS or Books. This indicates that the multiple clause linkage structure with *te*-clause occurs much more frequently in the particular genre of Old Japanese that the *Senmyō* represents (as exemplified in (9)).

The most frequent patterns of clause linkage in the *Senmyō* were analysed and compared with those in SPS and Books. Table 7 shows the result.

Table 7. Frequent Clause Linkage Patterns (SPS, Books, *Senmyō*)

SPS		Books		*Senmyō*	
26.2%	EOS	45.1%	EOS	21.3%	EOS
5.9%	*te* _ EOS	7.8%	*te* _ EOS	18.4%	*te* _ EOS
3.6%	quote _ EOS	6.4%	cont. _ EOS	8.0%	quote _ EOS
1.8%	*toyu* _ EOS	3.5%	quote _ EOS	5.5%	*te* _ *te* _ EOS
1.7%	*te* _ quote _ EOS	2.4%	*ga* _ EOS	2.9%	*te* _ quote _ quote _ EOS
1.6%	*de* _ EOS	1.8%	*to* _ EOS	2.7%	quote _ quote _ EOS
1.4%	*keredomo* _ EOS	1.8%	*de* _ EOS	2.5%	cont. _ EOS
1.4%	*to* _ EOS	1.6%	*reba* _ EOS	2.5%	*suruni* _ EOS
1.4%	*te* _ *te* _ EOS	1.2%	*toyu* _ EOS	2.1%	*te* _ quote _ EOS
1.2%	*ga* _ EOS	1.2%	*kara* _ EOS	2.1%	*te* _ *te* _ *te* _ EOS

The patterns "*te* _ EOS" and "quote _ EOS" are ranked at 2nd and 3rd place in *Senmyō*, which are common to the three registers. These patterns may have constituted the basic clause linkage structure from the era of Old Japanese, although the ratios of these patterns in *Senmyō* are much higher than those of the contemporary spoken and written Japanese. Some of the patterns of clause linkage structure in *Senmyō* may be attributable to the fact that it was a register written to be recited aloud in proclamations as the quoted words of current and former emperors.

While Kondo (2005) showed how the multiple clause linkage structure is frequently observed in the Early Middle Japanese texts, we show that such structures can also be found in Old Japanese. The question of which eras and registers gave rise to what type of clause linkage structures is an issue that should be clarified by compiling a large-scaled diachronic written corpus from a variety of different time periods. While such projects are only in the planning stages at this point, we will offer some concrete evidence of what they may bring to light in the future by examining a small excerpt of Japanese text sampled from an era only slightly removed from our own.

4.4 A Question of Style

From the viewpoint of the prescriptive grammar of contemporary Japanese, highly complex clause linkage structures should be avoided in prose. Nevertheless, we noted that multiple clause linkage structure frequently occurred in the prose of Early Middle Japanese and at even earlier stages as well. This prompts us to ask at what historical point the multiple clause linkage structure came to be regarded as something to be avoided in writing. (10) is an example of a newspaper article published in 1908.

10. *Tsurumizai ni hakken seraretaru kaiketsu ga oanasama to tataerarete / meishinja no sankei ōki yoshi o ichihayaku hōjitaruni, / kiji no eikyō wa keihin densha kabu ni oyoboshi, / hongetsu jōjun*

rokujū san shi en narishi mono / tonde / nanajū en go roku jissen ni bōtō shitaruga, / kowa oanasama sankeisha no higoto ni zōka shi / ichinichi sūsen nin no hitode arite / keihin densha no shūnyū ichijirushiku / zōka shitaru yori / tomi ni ninki no kōkyō o kuwaeshi mono naruga, / dō kaiketsu wa sude ni kisai seshi gotoku / kanagawaken keisatsusho yori meishinja o yūchi suru gotoki setsubi wo tekkyo seshimerareshini, / matamoya sono kiji ga arawaruru to dōji ni, / nijū yokka no shijō ni oite / gazen san en go jissen no bōraku o miru ni itareri to. /

When the news was reported first (here) that, the mysterious hole that was discovered in Tsurumizai being revered as 'the honorable hole' / the pilgrimages made (there) by worshippers are numerous / the influence of the article having affected the stock value of the Keihin Railroad Company, / while it was a value of 63-4 yen at the beginning of this month, / jumping / it rose rapidly to 70 yen and 50-60 sen, / but as for this, the people who make pilgrimage to 'the honourable hole' increasing daily / there being several thousand trips per day / it is a case where, added to the Keihin Railroad Company's earnings' notably / increasing / is its sudden favourable state of popularity, / but as for the selfsame mysterious hole, as has previously been reported, / installations for attracting worshippers having been removed by the Kanagawa Police Department / at the publication of the article in question, / positioned in the market of the 24th / (the value) suddenly came to see a fall of 3 yen and 50 sen, (so it is said) that… /

This article consists of one sentence fragment with a highly complex multiple clause linkage structure. The attestation of such a text in 1908 indicates that multiple clause linkage structures were a part of standard Japanese prose at least at the end of the Meiji era. Radical changes in written styles were happening throughout the Meiji era, but at some point after the Meiji era the multiple clause linkage structure evidently came to be avoided in formal written registers of Standard Japanese. In order to clarify the dynamic state of style-shift in Japanese history, we need to examine more corpora of written Japanese in the early 20th century. The investigation of all these issues remains as work for the future.

5 Concluding Remarks

This paper focuses on the issue of the multiple clause linkage structure. Quantitative and qualitative analyses and descriptions of this structure were conducted using two large-scale contemporary Japanese corpora, the CSJ and the BCCWJ. By identifying and examining multiple clause linkage structures, we shed light on the question of what kinds of settings give

rise to what type of clause linkage structures through what processes. The "dynamic rewriting rule" proposed by Kondo (2005) was discussed as a model for the production of multiple clause linkage structures, and their distribution in written texts of Early Middle Japanese was noted. Some common patterns of multiple clause linkage structures occurring in Old Japanese prose were identified and compared to patterns in written and spoken contemporary Japanese. Finally, a late Meiji era text was examined to suggest that the 'non-normative' status of the multiple clause linkage structure in prose is a more recent development than we initially expected.

Since the turn of the millennium, various corpora of contemporary Japanese have gradually been developed. Their widespread use has quickly prepared a foundation that makes it possible to quantitatively observe and describe linguistic phenomena that heretofore have been conventionally considered to be 'non-normative'. Even though contrasts and comparisons of spoken and written language have been conducted for a long time, by using large-scale corpora including a variety of registers, it has now become possible to conduct quantitative analyses that were unimaginable previously. In order to better understand the processes that bear on the form of the Japanese language, we encourage the creation of more various speech and written corpora including data across as many registers and historical periods as possible.

Acknowledgements

This study is supported by JSPS (Japan Society for the Promotion of Science) in two ways: through the *Program for Advancing Strategic International Networks to Accelerate the Circulation of Talented Researchers*, which is titled *Innovative Japanese Studies through International Cooperation: The Fostering of Young Researchers by Cooperation with Overseas Institutes of Japanese Studies* at Kobe University, and through the Grant-in-Aid for Scientific Research (no. 24520523). Some of the research for this study was also made possible by the Hakuho Foundation's 10th International Japan Research Fellowship, and the research grant from the Society of Japanese Language / Japanese Literature and Culture, Senshu University.

Bibiliography

Cutler, Anne (ed.) (1982). *Slips of the Tongue and Language Production*. Amsterdam: Mouton.

Frellesvig, Bjarke et al. (2010). "Verb Semantics and Argument Realization in Pre-modern Japanese. A Preliminary Study of Compound Verbs in Old Japanese". *Gengo Kenkyū*, 138, 25-65.

Gilquin, Gaëtanelle; De Cock, Sylvie (eds.) (2013). *Errors and Disfluencies in Spoken Corpora*. Amsterdam: John Benjamins Publishing.

Haiman, John; Thompson, Sandra A. (eds.) (1993). *Clause Combining in Grammar and Discourse*. Amsterdam: John Benjamins Publishing.

Hasegawa, Yōko (1996). *A Study of Japanese Clause Linkage. The Connective TE in Japanese*. Stanford; Kyōto: CSLI; Kurosio Publishers.

Kondō Yasuhiro (2005). "Heian jidai go no fukushisetsu no setsu rensa kozo ni tsuite" (On the Clause Linkage Structure of Subordinate Clauses in Heian Period). *Kokugo to Kokubungaku*, 82(11), 114-24.

Levelt, Willem J.M. (1989). *Speaking: From Intention to Articulation*. Cambridge: MIT Press.

Maekawa et al. (2014). "Balanced Corpus of Contemporary Written Japanese". *Language Resources and Evaluation*, 48, 345-71.

Maruyama et al. (2004). "Nihongo setsu kyōkai kenshutsu puroguramu CBAP no kaihatsu to hyōka" (Development and Evaluation of Japanese Clause Boundaries Annotation Program). *Shizen gengo shori* (Journal of Natural Language Processing), 11(3), 39-68.

NINJAL, National Institute for Japanese Language and Linguistics (2006). *Nihongo hanashi kotoba kōpasu no kōchikuhō* (Construction of The Corpus of Spontaneous Japanese) [online]. URL `http://pj.ninjal.ac.jp/corpus_center/csj/k-report-f/CSJ_rep.pdf` (2017-04-10).

Olson, Michael L. (1981). *Barai Clause Juncture. Towards a Functional Theory of Inter Clausal Relations* [PhD Dissertation]. Canberra: The Australian National University.

Sakakura Atsuyoshi (1975). *Bunshō to hyōgen* (Discourse and Expression). Tōkyō: Kadokawa Shoten.

Tsunoda, Tasaku (2013). "Five Levels in Clause Linkage. Theoretical and Methodological Preliminaries". Tsunoda, Tasaku (ed.), *Five Levels in Clause Linkage*, vol. 1, 17-44.

Tyler, Joseph (2012). "Rising Pitch, Continuation, and the Hierarchical Structure of Discourse". *University of Pennsylvania Working Papers in Linguistics,* 20(1), 36.

Van Valin, J.R.D. (1984). "A Typology of Syntactic Relations in Clause Linkage". *Proceedings of the Tenth Annual Meeting of the Berkeley Linguistics Society*, 542-58.

Watanabe, Michiko et al. (2015). "The Relationship Between Preceding Clause Type, Subsequent Clause Length and the Duration of Silent and Filled Pauses at Clause Boundaries in Japanese Monologues". *Workshop Disfluency in Spontaneous Speech* (Edinburgh, 8-9 August 2015).

Yuasa, Etsuyo; Sadock, Jerry M. (2002). "Pseudo-subordination. A Mismatch Between Syntax and Semantics". *Journal of Linguistics*, 38, 87-111.