University of Ljubljana
FACULTY OF ARTS

Irena Srdanović and Andrej Bekeš (eds.)

The Japanese Language from an Empirical Perspective

Corpus-based studies and studies on discourse

2019

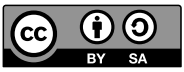
The Japanese Language from an Empirical Perspective:
Corpus-based studies and studies on discourse

Editors: Andrej Bekeš, Irena Srdanović

Reviewers: Kristina Hmeljak Sangawa, Terry Joyce, Heiko Narrog, Masaki Ono, Prashant Pardeshi, Vera Sheinman, Darinka Verdonik
Proofreading: Krešimir Vunić

Layout: Aleš Cimprič

Cover photo: Andrej Bekeš



Published by: Znanstvena založba Filozofske fakultete Univerze v Ljubljani
(Ljubljana University Press, Faculty of Arts)

Issued by: Department of Asian Studies

For the publisher: Roman Kuhar, Dean of the Faculty of Arts, University of Ljubljana

Printed by: Birografika Bori d. o. o.

Ljubljana, 2019

First edition

Number of copies printed: 200

Price: 21,90 EUR

The authors acknowledge the financial support from the Slovenian Research Agency (research core funding No. P6-0243(A).

The publication was supported by Javna Agencija za raziskovalno dejavnost Republike Slovenije (Slovenian Research Agency)

First e-edition. Publication is available free of charge on https://e-knjige.ff.uni-lj.si/
DOI: 10.4312/9789610602170

Kataložna zapisa o publikaciji (CIP) pripravili v
Narodni in univerzitetni knjižnici v Ljubljani

Tiskana knjiga
COBISS.SI-ID=288040192
ISBN 978-961-237-886-8

E-knjiga
COBISS.SI-ID= 300516096
ISBN 978-961-06-0217-0 (pdf)

Contents

III Research of corpora applied to Japanese language education

IV Corpus-based diachronic research

Preface

Language is a multifarious phenomenon, and so is language research. Language certainly does have its individual aspects, and yet, as linguists of many different persuasions have constantly pointed out, it is essentially a social phenomenon. Therefore, relying on introspection alone will not lead one any great distance towards a solid understanding of linguistic phenomena; the research of empirical data cannot be avoided.

Concerning the empirical research of Japanese, it has a long tradition, which can be traced back to early vocabulary studies, which includes the collections of *waka* texts and the philological work of the *kokugaku* tradition. In the 1950's and 1960's, what was then known as the National Language Research Institute (NLRI, *Kokuritsu kokugo kenkyūsho*) was the focal point of empirical research, its primary motivation being language planning and language education. The ever-relevant work of Hayashi Shirō and Minami Fujio have their origins in this experience. Another empirical tradition includes early written discourse studies, as proposed by Tokieda Motoki as *bunshōron* and continued by his disciples. Yet another important tradition, from the same period and embracing a thorough empirical approach to language in a social context is represented by *Gengo kenkyūkai* (Language research society). Kudō Hiroshi's seminal research on adverbs (Kudō 2000) must be mentioned among their relatively recent achievements, which the editors here have found particularly inspiring.

At a somewhat later date, from the late 1960's to the end of the 1990's, the emphasis was primarily on theory centred approaches, which were not always keen on empirical verification. At the same time, a systematic sample-based analyses of the lexicon continued with the use of index cards and other techniques, with the first collections of computerized data appearing in the 1970's, such as those produced by the NLRI. Also, since 1980's, discursive studies and sociolinguistic studies were again in the ascent, with a focus on work on text and discourse, including conversation analysis.

In the first decade of the new millennium, corpus building and research received a new impetus with the development of large-scale corpora, mainly CSJ (Corpus of Spontaneous Japanese) and BCCWJ (Balanced Corpus of Contemporary Written Japanese). The second decade of the new millennium, with arrival of even larger scale Japanese web corpora (e.g. JpWaC, Tsukuba web corpus, JpTenTen), and the wide accessibility of these corpora over the Web, proved to be a major turning point in the empirical research of the Japanese language.

This book was inspired by the special satellite sessions accompanying the 14th International Conference of the European Association of Japanese Studies and some of the presentations from these sessions proved seminal for future collaboration and were

also an inspiration for a book which would cover various research achievements within an empirical approach to Japanese language research. The purpose of these sessions was to draw the attention of researchers in Europe towards the "paradigm shift" which occurred in the empirical research of the Japanese language in Japan, inspired to great degree by the easy accessibility of the aforementioned corpora and widespread perception of the importance of language study in context (as in discursive studies) which we believed required further elaboration resulting in a monograph for the wider community.

This book presents empirical methodologies and insights into the field of spoken and written discourse, in syntax, lexis, in corpus-based research and its applications to Japanese language education, and an exploration of the differences across time and register in diachronic Japanese language corpora. As such, it is divided into four parts, which are presented below with a brief overview of each chapter.

Part I, *Analysis of spoken and written discourse*, consists of four chapters. The first chapter, by Sakuma Mayumi, deals with the question of the appropriate units for the analysis of Japanese written and spoken discourse. Since Tokieda proposed his written discourse constituents (*bunshō no seibun*), there has been a lively discussion of what the appropriate units actually are, as well as the criteria for their identification. At a later date, Minami and others extended this discussion towards spoken discourse. With this research as her point of departure, Sakuma argues for *dan* (grammatico-semantic paragraph) as a communicative unit of spoken and written discourse. *Dan* is a unit, and is coherent on the basis of topic unity, it can also be multiple, and is realized as *bundan* (grammatico-semantic written paragraph) in written discourse and *wadan* (grammatico-semantic spoken paragraph) in spoken discourse.

The second chapter, written by Takasaki Midori, focuses on the text-organizing function of a certain type of lexical items, "text-organizing words" and their cohesive function in the text. Takasaki argues that "text-organizing words", although abstract (yet not as abstract as formal nouns), are employed in order to organize a text as a series of "semantic segments". They contribute, on the basis of their cohesive properties, to the coherence of a text. The relationship between the text-organizing function of "text-organizing words" and cohesion, Takasaki claims, is that of realization, the former being realized by the later.

The third chapter, written by Polly Szatrowski, investigates how Japanese and American participants track references to unfamiliar food at taster lunches. In her analysis she investigates (1) What aspects of the food do participants use as resources to create references to unfamiliar food?, (2) What patterns in reference tracking can be observed through conversation?, (3) How do participants' choices of similar or different referring expressions influence their assessment and categorization of the food in question and their relationship to each other? An interesting outcome of this analysis could be that referring expressions for less familiar foods continued to be modified throughout

the discussion of the food item. This result suggests the ephemeral and fluid nature of the referential categories we use when dealing with the world we live in, in opposition to our default static notion of language, especially in an educational context.

The final chapter of Part I, by Sunakawa Yuriko, deals with Japanese cleft sentences. There are two types of Japanese cleft sentences: WA-clefts and GA-clefts. The predicate of a WA-cleft can be either a noun or a subordinate clause, whereas the predicate of GA-clefts is restricted to a noun. While the predicate noun in both types of clefts tends not to be accompanied by a case particle, this tendency is much stronger in the case of GA-clefts. Sunakawa argues that the above characteristics are not syntactic restrictions but the preferred patterns of use of cleft sentences in discourse, claiming that (1) Japanese cleft sentences have two types of discourse function, namely 'focus-presentational function' and 'prominence-presentational function', and (2) that the above-mentioned grammatical characteristics of WA-clefts and GA-clefts can be explained by their discourse functions.

Part II, *Corpus-based research on discourse variety and lexis*, consists of two chapters. The first one, written by Andrej Bekeš, deals with suppositional adverbs as discriminators of Japanese corpora according to oral and written discourse varieties. Considering modal expressions as a speaker's/writer's signals for the nature of a particular linguistic exchange or as a trace of such linguistic exchange, Bekeš argues that since modal adverbs are easier to identify than sentence-final modal expressions, they may serve to discriminate between different varieties of Japanese oral and written discourse in corpora. Focusing on suppositional adverbs, a subset of modal adverbs, Bekeš analyses their distribution in several written and spoken corpora, belonging to different genres. Cluster analysis shows that the distribution of suppositional adverbs in analysed corpora varied according to discourse type. Differences in distribution within corpora belonging to same clusters are accountable by the difference in the degree of formality or different rhetoric strategies. Thus, indeed, it seems that suppositional adverbs discriminate corpora according to discourse type.

The second chapter of Part II, by Irena Srdanović, focuses on the Japanese *i*-adjectives, mainly in their role as modifiers of nouns. Empirical methods of corpus linguistics and employing the latest language resources and lexical profiling tools have been put to use here. The study confirms the distribution of *i*-adjectives by pointing out the most prominent adjectives as well as the most productive adjectival suffixes. Furthermore, Srdanović singles out adjectives with lexical constraints in syntactic patterns of certain adjectives and provides suggestions that a division should be made between the types of attributive roles of adjectives based on complexity and the varieties of the patterns discovered. This research is based on two corpora of different sizes and demonstrates how lexical constraints need to be observed in larger data collections in order to obtain results that are more reliable.

Part III, *Research of corpora applied to Japanese language education* consists of two chapters. The first is by Jae-ho Lee and Hasebe Yoichiro, who propose a method for measuring the readability of Japanese texts. Its originality is in its use of corpora consisting of textbooks for learners of Japanese as a foreign language, modelling the corpora using six-levels of difficulty and developing measuring formula appropriate for teachers or learners of Japanese. This research has led to an application of its results for a web-based system that can provide support to teachers of Japanese when preparing reading materials appropriate to a student's level.

The second chapter of Part III, by Kikuko Nishina, Bor Hodošček, Yagi Yutaka and Abekawa Takeshi, also deals with the application of corpora in Japanese language learning and teaching. This research presents Nutmeg, a writing support system for Japanese language learners, whose main feature is to identify mistakes in learner writing while also being register-aware. This can be achieved by using a number of Japanese corpora in various registers and classifying learners' expressions based on their frequency distribution across the corpora. This paper examines adverbs within Japanese academic register and evaluates the classification results of the system.

Part IV*Corpus-based diachronic research*, consists of two chapters, both dealing with the development and usage of diachronic corpora in the Japanese language. The first chapter, written by Ogiso Toshinobu, quantitatively analyses stylistic differences across time and register in old Japanese texts. Ogiso begins with an overview of the construction of the Corpus of Historical Japanese (CHJ) at the National Institute for Japanese Language and Linguistics, also covering information on their level of annotation. As historical materials in Japanese are limited, it is a challenge to determine if these discovered characteristics are the result of diachronic linguistic change or a matter of genre differences. The significance of this research is in examining the characteristics of old Japanese texts by use of multiple methods while comparing them to the various text genres of contemporary written Japanese.

The second chapter in part IV, by Maruyama Takehiko, also discusses the limitations of diachronic corpora as unavoidable when analyzing resources for the study of old language, but from the viewpoint of the possibilities of compiling a diachronic speech corpus of Japanese. This research presents several analyses by using three different recorded resources of old spoken Japanese: intonation patterns and grammatical expressions, auxiliary verbs and sentence-final particles, directing our attention to several interesting findings about spoken Japanese, such as rapid rising intonation. This research comes to the conclusion that more (and different) recorded resources are necessary for a more adequate diachronic speech corpus in the future.

Irena Srdanović and Andrej Bekeš

I

Analysis of spoken and written discourse

1 Units for the analysis of Japanese written text and spoken discourse[1]

SAKUMA Mayumi
Waseda University

Abstract
Written text and spoken discourse, which use letters and sounds, respectively, as the medium for communication, are the largest and most concrete linguistic units. Both, as the sole actual forms of Japanese communication, are complete coherent wholes which dynamically unify linguistic behaviour.

Issues related to the units of written text and of spoken discourse are both old and new. With respect to written text units, the question of which written text constituents (sentences, sentence sequences, paragraphs, *bundan* 'written grammatico-semantic paragraphs', etc.) are appropriate has been debated since Tokieda (1950:289). Similarly, with regard to spoken discourse, Minami (1997:295-356) investigated criteria for the identification of "written text (*bunshō*)", "conversation (*kaiwa*)" and "discourse (*danwa*)" units, and Hayashi (1998:394-396) argued for the necessity of "qualitative units" which he called "communication units".

Investigation of the similarities and differences between "written text" and "spoken discourse" as effective analytic units for the comprehensive description of linguistic behaviour is an issue that cannot be avoided in "written text/spoken discourse theory" in Japanese linguistics. In this paper, I explore the potential for *bundan* (groups of utterances in written texts) and *wadan* (groups of utterances in spoken discourse) to be effective analytic units of written text/spoken discourse focusing on their "unifying function" and "multiple structure".

Keywords: Analytic units (*bunseki tan'i*), Japanese discourse analysis (*bunshō-danwaron*), grammatico-semantic written paragraph (*bundan*), grammatico-semantic spoken paragraph (*wadan*), coherency function (*tōkatsu kinō*)

1 This study is extensive elaboration of my original text, Sakuma (2006).

1 Introduction

Written texts and spoken discourses (*bunshō-danwa* 文章・談話) are the largest and most concrete units of written and spoken language, respectively. I will use the term "linguistic unit" (*gengo tan'i* 言語単位) to refer to a dynamic unit (discrete whole) of verbal and nonverbal action in Japanese communication.

Tokieda, the founder of Japanese discourse analysis (*bunshōron* 文章論), proposed a "qualitative view of units" (*shitsuteki tan'ikan* 質的単位観)[2] based on his unique theory of "language as process" (*gengo kateisetsu* 言語過程説) within the framework of Japanese traditional linguistics (*kokugogaku* 国語学) (1950: 15-17; 1960: 9-11). He claims that the "qualitative view of language" is a view of units that presumes that a whole (*ichi zentai* 一全体), understood as qualitatively unified whole (*shitsuteki tōitsutai* 質的統一体), is not the ultimate end of analysis, but is rather already given at the beginning of research.

In Tokieda's (1950: 15-17) view, the "basis for establishing the field of Japanese text study (*bunshō kenkyū* 文章研究) lies in the fact that among the three "units of language," i.e., "word" (*go* 語), "sentence" (*bun* 文), and "written text" (*bunshō* 文章), the written text differs from the other two units, because it is "a whole with a unified structure".

The problem of units in Japanese written text and spoken discourse is both old and new. Since Tokieda (1950:289), there has been continuous debate concerning units in Japanese written texts, specifically on which "written text constituents" (*bunshō no seibun* 文章の成分) (e.g., sentence, sequence of sentences (*renbun* 連文), formal (indented) paragraph (*danraku* 段落), grammatico-semantic written paragraph (*bundan* 文段), etc.), are appropriate. Although scholars have differed in their views on language, grammar and the position of text study, they all have agreed that the written text is a unit beyond the sentence.

Regarding the units of Japanese spoken discourse, Minami (1997:295-356) addresses the question of which criteria are applicable for identifying units, such as a written text, a conversation, and discourse. Following Minami's view of discourse[3] as a unit intermediate between sentence and conversation, Hayashi (1998:394-396) pointed out that in addition to units that are intermediate between sentence and written text, i.e., sentence clusters (*bunkai* 文塊) and formal (indented) paragraphs, it is necessary to have "qualitative units" (*shitsuteki tan'i* 質的単位), which he called "communication units" (*komyunikēshon tan'i* コミュニケーション単位), particularly in spoken discourse.

---

2 Tokieda (1950) contrasts his theory with Saussure's "structural view of language" and "atomistic view of units".

3 See Minami (1997: 297) and Minami (1997: 337-355). After 1970, Minami established "a (spoken) discourse" (*danwa* 談話) as a unit of "conversation" (*kaiwa* 会話), comparing it to "something like an indented paragraph of a written text" (Minami 1997: 297). However, he later renamed it as a "coherent unity of a conversation" (*kaiwa no matomari* 会話のまとまり) (Minami 1997: 337-355).

On the other hand, there is also a view that questions the very existence of structure and units in spoken discourse of oral communication[4]. Nonetheless, when positioning written text/spoken discourse analysis in Japanese linguistics (*nihongogaku* 日本語学), there is no question that it is an unavoidable and an important challenge to compare and contrast the units in spoken discourse, which have entered a participants' memory the moment they are uttered, and the units of a written text which are fixed as strings of characters.

Thus, beginning with the premise that written texts and spoken discourse themselves are units of verbal communication, it is possible to refine their analysis and description by establishing the multiple levels (*dankai* 段階) and elements (*yōso* 要素) involved in the process of communicating linguistic information. One could say that it is in fact self-evident that some "part" (*bubun* 部分) and "process" (*katei* 過程) must exist to support the "whole" (*zentai* 全体) / "completion" (*kanketsu* 完結) of the highest level units, i.e., written texts / spoken discourses, which are composed of lower level units such as sentences, words etc.

Based on the assumption that we establish linguistic units according to their usefulness for the analysis and goals of the research, in the remainder of this paper I will examine the constituent elements of written texts and spoken discourses and discuss their similarities and differences. The goal will be to ascertain what units are useful for analyzing the organization of written text and spoken discourse.

2 Sentence and grammatico-semantic written paragraph as the units of Japanese written text

While Tokieda (1950:289) did not consider individual sentences to be constituents of "Japanese written texts", he did consider "paragraphs" (*bunsetsu* 文節, *danraku* 段落, *bundan* 文段), and "chapters" (*shō* 章, *hen* 篇) as constituents. However, he did not provide a detailed definition regarding these constituents.

Nagano (1972/1986), systematizing "grammatical Japanese text analysis" (*bunpōronteki bunshōron* 文法論的文章論), proposed "sentence" and "formal (indented) paragraph" (*danraku* 段落) as units. In contrast, Ichikawa (1978), taking the view of "general Japanese text analysis" (*hanbunshōron* 汎文章論), proposed "sentence" and "grammatico-semantic written paragraph" (*bundan* 文段) as units. Like Ichikawa's "grammatico-semantic written paragraph", Tsukahara (1966) does not consider "formal (indented) paragraphs" (in his terminology "rhetorical paragraphs" (*shūjiteki danraku* 修辞的段落)), to be a constituent

4 See Nomura (2002: 110). A similar opinion was also expressed at the Symposium of the Society for Japanese Linguistics held in spring in 2006 (Session A "(Japanese) written texts/ spoken discourse").

unit of Japanese written texts, but rather he considers "logical paragraph" (*ronriteki danraku* 論理的段落) to be the constituent unit. Furthermore, by dividing paragraphs into "basic paragraphs" (*kihon danraku* 基本段落) and "paragraph clusters" (*danraku rengō* 段落連合), he admits the possibility that a complex sentence might consist of several "paragraphs". Nagata (1995), on the basis of " sentence sequence theory" (*renbunron* 連文論) and Japanese written text analysis, proposes "word", "sentence" and "paragraph" (*danraku* 段落) as units, basing his understanding of paragraph on Tsukahara's definition.

There is also a view that does not recognize the "formal (indented) paragraph" as a "linguistic unit" based on objective criteria, considering it more as something belonging to the sphere of punctuation rules. Furthermore, regarding grammatico-semantic written paragraphs with multiple structure (*jūsō kōzō* 重層構造) deriving from "coherent organized units based on big and small topics" (*daishō no wadai no matomari* 大小の話題のまとまり), opinions disagree as to which relative division into units, at what level, and with what amount/extent of content should be taken as a basis for constituent units of Japanese written text and discourse. The existence itself of formal indicators/criteria for grammatico-semantic written paragraph has also been questioned.

2.1 Analysis of 'cohesion between sentences' (*bun no tsunagari* 文のつながり) based on sentence as a unit

Nagano's (1972/1986) "theory of cohesion" (*rensetsuron* 連接論), "theory of continuity" (*rensaron* 連鎖論), and "theory of coherency" (*tōkatsuron* 統括論) share a view that takes sentence as the basic unit for analyzing the structure of Japanese written texts. This point of view considers the structure of a Japanese written text, defined as a unified body consisting of connected sentences (*bun no renzoku tōitsutai* 文の連続統一体), to be relations between individual sentences and between formal (indented) paragraphs. These relations are viewed as "conjunctive relations between sentences" (*bun no rensetsu kankei* 文の連接関係) based on the cohesion between individual sentences or formal (indented) paragraphs, and, furthermore, as "continuity relations between sentences" (*bun no rensa kankei* 文の連鎖関係), based on chains of subjects (*shugo* 主語), predications (*chinjutsu* 陳述)[5], and principal words and phrases (*shuyōgoku* 主要語句). In my opinion, this view of Japanese written text is being too microscopic and a mere application of the results of grammar research to written texts.

A "sentence sequence" (*renbun* 連文) is a body of semantically connected sentences. The smallest sequence of sentences is a pair of two adjacent sentences, and the largest coincides with the whole written text. In addition, a "complex sentence" (*fukubun* 複文), consisting of several clauses, can be identified as a grammatico-semantic written

5 The usage here follows Heiko Narrog (2009) *Modality in Japanese: The layered structure of the clause and hierarchies of functional categories*, Amsterdam: John Benjamins, translating *chinjutsu* 陳述 as "predication" (translator's comment).

paragraph, based on the fact that the function of the adverbial predicate forms (*ren'yōkei* 連用形) that make up these sequences corresponds to that of connectives. Thus, complex sentences can be considered to be units akin to a sentence sequence.

"Sentence sequence theory in the narrow sense" (*kyogi renbunron* 狭義連文論) is concerned with the semantic connections between adjacent sentences, while "sentence sequence theory in the wide sense" (*kōgi renbunron* 広義連文論) takes into account the whole written text. In other words, "cohesion between sentences" is the precondition for a "coherent unity of sentences". The grammatico-semantic written paragraph is a unit composed of connected sentences which constitute a written text. In contrast, sentence sequences are merely a part of the grammatico-semantic written paragraph, that is, a body of connected sentences expressing a fragment of some topic. This limits the extent to which a sentence sequence can be a constituent unit of a written text.

2.2 Analysis of 'coherent unity of sentences' (*bun no matomari* 文のまとまり) based on grammatico-semantic written paragraph as a unit

"Grammatico-semantic written paragraphs", occupying the middle ground between sentence and written text, can be embedded to larger units, that is, "semantic paragraph sequences" (*rendan* 連段). These "large grammatico-semantic written paragraphs" (*daibundan* 大文段), which are made of several grammatico-semantic written paragraphs express a larger coherent unity of semantically related themes. In the final analysis of a written text, grammatico-semantic written paragraphs and semantic paragraph sequences, mutually related through coherency (*tōkatsu* 統括), establish the largest multiple structure of the written text, on the basis of the coherent unity of topics.

In Sakuma (2003:91-119), I made a distinction between "topic/core sentences" (*chūshinbun* 中心文), sentences that express the principal information of a grammatico-semantic written paragraph in the most straightforward way, and "thesis sentences" (*shudaibun* 主題文), sentences that express the theme of the whole written text. It is usually the case that a written text consists of several semantic paragraph sequences with a unified theme (*shudai* 主題) and that a grammatico-semantic written paragraph consists of several sentences with a unified topic (*wadai* 話題). Here, by "topic" I mean an expression of the principal content in a grammatico-semantic written paragraph. Topic sentences and thesis sentences both possess the coherency function (*tōkatsu kinō* 統括機能), imparting coherent unity to the large and small topics in a written text.

Topic sentences and thesis sentences impart a relative strength of coherency corresponding to the scope and frequency of the topics in grammatico-semantic written paragraphs, as well as to the text-developing function (*bunshō tenkai kinō* 文章展開機能) of these topics. Furthermore, the thesis sentence of the theme/core paragraph (*chūshin dan* 中心段), the grammatico-semantic written paragraph with the largest strength of

coherency (*tōkatsuryoku* 統括力) in the written text, organizes and completes the whole written text.

The coherency function of the grammatico-semantic written paragraph performs the following roles:

(i) topic presentation (*wadai teiji* 話題提示)
(ii) conclusion (*ketsuron hyōmei* 結論表明)
(iii) issue raising (*mondai teiki* 問題提起)
(iv) introduction of the problem/ issue to be solved (*kadai dōnyū* 課題導入)
(v) connection with the preceding context and introduction of the following context (*shōzen kigo* 承前起後)
(vi) introduction (*maeoki* 前置き)
(vii) appending (*atozuke* 後付け)

In addition, both topic/core sentences and theme/core paragraphs can appear in a grammatico-semantic written paragraph and in the written texts respectively, in one of the following six positions:

1) beginning (*saisho* 最初)
2) end (*saigo* 最後)
3) beginning and end (*saisho to saigo* 最初と最後)
4) middle (*chūkan* 中間)
5) several dispersed positions (*fukusū bunsan* 複数分散)
6) implicit (*senzai* 潜在)

Furthermore, the position of topic/core sentences and theme/core paragraphs may differ depending on the type of written text, on their position and frequency within grammatico-semantic written paragraphs, and on their text-developing function in the written text. The "topic presentation" (*wadai teiji* 話題提示), the "introduction of a problem/issue to be solved" (*kadai dōnyū* 課題導入) and the "introduction" (maeoki 前置き) itself tend to appear at the beginning, while the expression of the "conclusion" (*ketsuron hyōmei* 結論表明), the "issue raising" (*mondai teiki* 問題提起), and the "appending" (*atozuke* 後付け) tend to appear at the end. Finally, the "connection with the preceding context and introduction of the following context" (*shōzen kigo* 承前起後) tend to appear in the middle or in the grammatico-semantic written paragraphs that are constituted by a single sentence.

3 Utterances and grammatico-semantic spoken paragraphs as units of Japanese spoken discourse

Japanese spoken discourse is made up of "utterances" (*hatsuwa* 発話), subunits of Japanese spoken discourse, that is, spoken linguistic units of various sizes. It is nonetheless reasonable to think that the structure of Japanese spoken discourse can be described to a considerable degree from the point of view of the analysis of Japanese written text.

3.1 Analysis of 'topic sequences' with utterances as units

Sugito (1987: 83) analyzes the "transfer/continuity of utterances" (*hatsuwa no uketsugi* 発話のうけつぎ) in data from round table discussions. He defines an "utterance" as follows:

> This [an utterance] is a unit, an internally consistent chunk of a continuum of spoken language by a single participant (also including laughter and short back channeling). Each chunk is delimited by continuum of spoken language (same as above) produced by other co-participant[s] and by pauses [gaps in time], and is counted as a separate unit. (Example omitted). This unit, i.e., utterance, may be shorter than what is considered to be a sentence in grammar, [and often] corresponds to a phrase (*bunsetsu* 文節, *ku* 句) or a word or just a part of an interrupted expression. Yet sometimes it may be longer, appearing to be a sequence of two or more sentences. The "utterance" is a unit which may appear variously in long or short form. If one regards units as necessarily uniform/homogeneous, it is true that it is difficult to call such an internally consistent chunk a unit. Nonetheless, for the purpose of following the verbal exchange between participants in Japanese spoken discourse, the utterance provides a rather explicit clue regarding the actual divisions in such verbal exchange.
>
> (Sugito 1987: 83, underline by Sakuma)

If an "utterance" with no fixed length is a unit lacking a "homogeneous" size, then a "grammatico-semantic spoken paragraph", defined as a content-based relative division, shares a similar characteristic, i.e., a lack of "homogeneity" of its length due to the variable complexity of its content. However, is it not that this property characterizes the communication units of all verbal behaviour? It can be stated that in this respect "grammatico-semantic written paragraphs" in written text are similar communication units.

Szatrowski (1993) identified *wadan* (i.e., grammatico-semantic spoken paragraph, which she reinterpreted as 'stages') in invitation conversations based on the differences in the participants' goals and the interaction between "utterance functions", and described the overall structure (*zentaiteki kōzō* 全体的構造) of telephone conversations.

Her analysis provided empirical evidence for the limitations of analyzing Japanese conversation structure based only on adjacency pairs and utterance sequences.

3.2 Analysis of 'coherent unity of topics' with grammatico-semantic spoken paragraphs as units

Minami (2005: 537), assuming that "linguistic units are set up in linguistic analysis and description methodology as the basic elements constituting language", proposes units for each of the five realms of language:

> (5) [Units] based on spoken discourse (written text) and related to verbal behaviour: sentences (and utterances equivalent to sentences), as well as various coherently organized discourse units belonging to written texts and spoken discourse (i.e., formal [indented] paragraphs, grammatico-semantic spoken paragraphs (*wadan* 話段), etc.), and in addition, also coherently organized units of communication behaviour including both verbal and nonverbal expressions also belong here.
>
> (Minami 2005: 537, underline by Sakuma)

Here, "spoken discourse (written text)" refers to the largest linguistic unit in the same way as "written texts/spoken discourses" does in this paper. However, in particular the units Minami has put in parenthesis (), that is "utterance" and "formal (indented) paragraph", "grammatico-semantic spoken paragraph" (*wadan*), etc., require special attention. It is particularly relevant for the present research that Minami (1997), who has been consistently investigating "units of Japanese spoken discourse", at this point for the first time proposes a unit called "grammatico-semantic spoken paragraph".

Sakuma (1987) was the first to propose the term "grammatico-semantic spoken paragraph" (*wadan*) for a constituent of Japanese spoken discourse, corresponding to "grammatico-semantic written paragraph" (*bundan*), a unit of Japanese written texts. In addition, Sakuma (1992; 2003:91) proposes and defines the "grammatico-semantic paragraph" (*dan*), as a group of one or more sentences or utterances which in principle form a "communicative unit". Verification of the six to eight criteria proposed by Minami (1997) as clues for the identification of the units of Japanese spoken discourse remains the most urgent task in Japanese discourse analysis (*bunshō danwaron* 文章・談話論).

As units of Japanese spoken discourse, neither "utterances" nor "grammatico-semantic spoken paragraphs" can be included in the framework of particular formal units such as "words", "phrases" (*ku*), "clauses" (*setsu*), "sentences", etc. As coherently organized units of spoken/sound/vocalized expression and semantic content, they come into existence fluidly during the communication process of spoken discourse. "Grammatico-semantic spoken paragraphs", like "grammatico-semantic written paragraphs", possessing

an internal multiple structure imparted by the coherency function. The coherency function itself originates in coherently organized units based on topics, supports the macrostructure of Japanese spoken discourse and is thus a dynamic unit of linguistic behaviour, that is deeply involved in the process of spoken communication.

Sugito (1984) considers the units of Japanese spoken discourse to be entities which fulfill the conditions of "necessity, validity/effectiveness (*yūkōsei* 有効性) and sufficiency (*jusokusei* 充足性) based on his research goals and the analytic viewpoint ", and proposes the following four "basic characteristics of linguistic units": parallelism (*heiretsusei* 並列性), possibility of combination (*ketsugōsei* 結合性), "multilayeredness" (*jūsōsei* 重層性) and "exhaustiveness" (*mōrasei* 網羅性). However, it is my claim that among these characteristics, "multilayeredness" in particular seems to reveal the essence of the units of Japanese spoken discourse.

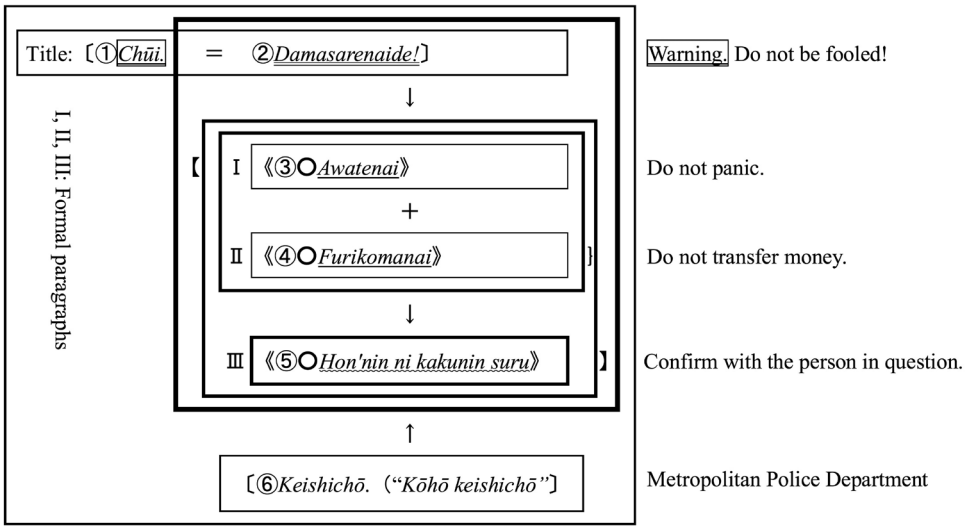
4 "The grammatico-semantic paragraph" dan as a unit of Japanese written text and spoken discourse

Following Minami's (2005; 1997) "coherently organized communicative behaviour units, based on linguistic and nonlinguistic expression", valid analytic units reflecting various aspects of communication in Japanese possess several clues that constitute criteria for their identification, as well as linguistic expressions, which function as formal markers.

If the "grammatico-semantic paragraph" (*dan* 段) is used as a general term which includes both "grammatico-semantic written paragraphs" of Japanese written text and "semantic spoken paragraphs" of Japanese spoken discourse, then, in addition to topic, there are various other criteria for the identification of units. These include criteria based on factors such as participants, scene, communicative function, attitudes of expression, etc. These criteria reveal elements and levels which are indispensable for the analysis and description of the structure and function of "Japanese written text and spoken discourse".

Example (1) below is a short public information text which, including the title, consists of 6 sentences and 3 formal (indented) paragraphs (I, II, III). Its text structure consists of an enumeration of specific measures to take in response to a warning. Coherency of the 3 formal (indented) paragraphs, realized in the body of the text as the sentences ③, ④ and ⑤ respectively, is imparted by the thesis sentence ②.

(1)



“*Kōhō keishichō*” *16 gō,* 2005. 1. 16. (Metropolitan Police Department public information No. 16, 16. Jan. 2005.), underline by Sakuma, with *dan* units (in boxes) and conjunctive relations among them. For the explanation of symbols =,↓, +,↑, and for the annotated original text, see Appendix 1. Brackets are explained in the footnote 5 below.

Sentences ① and ② represent the Opening section (*kaishibu* 開始部) and ⑥ the Closing section (*shūryōbu* 終了部) of the text, displaying its overall structure. Sentence ② is part of the title and conveys information by addressing the reader with an expression of prohibition (*kinshi hyōgen* 禁止表現). Although the main text ③～⑤ consists of three formal (indented) paragraphs, from the point of view of content, they are all part of a grammatico-semantic written paragraph with sentence ⑤ as the topic/core sentence. It is difficult to recognize formal (indented) paragraphs, employed for visual effect, as intrinsic units of the text. On the other hand, this grammatico-semantic written paragraph goes beyond the limits of the main text ③～⑤. Sentence ② from the title imparts coherency to the whole text, while the predication of sentence ⑤, i.e., an imperative expression, imparts coherency to the expressions of prohibition in sentences ②～④. Based on the properties of ⑤, it is possible to identify the multiple structure of the whole grammatico-semantic large written paragraph ①～⑥.

Example (2) is the Opening section of a spoken discourse, a public lecture, and since it is a monologue, the whole text (including quotations), is one utterance, with a multiple structure (imparted by several grammatico-semantic spoken paragraphs belonging to different dimensions). Furthermore, it contains complex sentences that are

made up of several grammatico-semantic spoken paragraphs. In addition, it also has formal markers that indicate its multiple structures based on the coherency function of big and small grammatico-semantic spoken paragraphs.

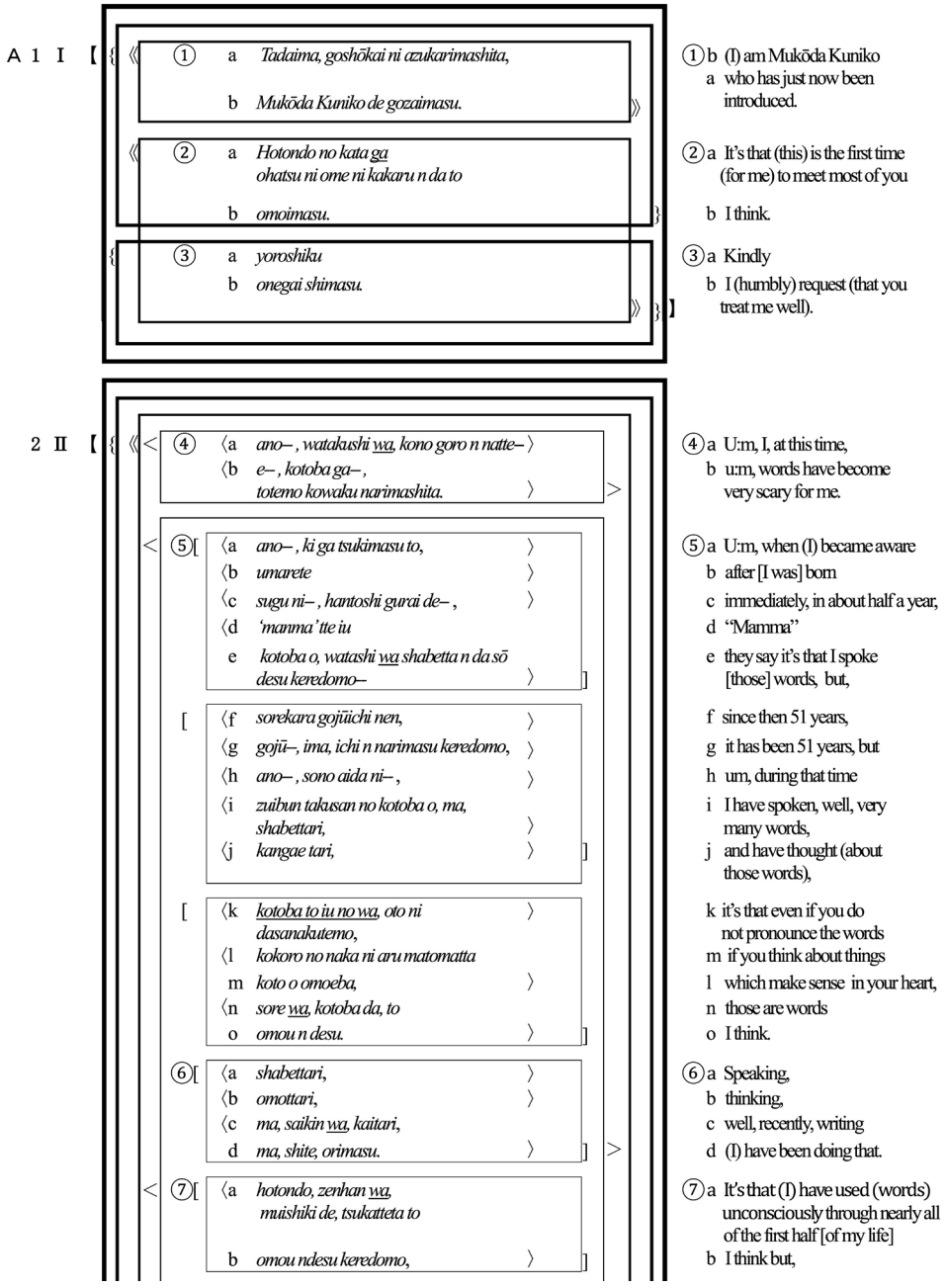
Example (2) below is divided into 4 grammatico-semantic spoken paragraphs. The super large grammatico-semantic spoken paragraph A, the semantic paragraph cluster consisting of grammatico-semantic paragraphs I, II, III, corresponds to the Opening section of this lecture discourse.

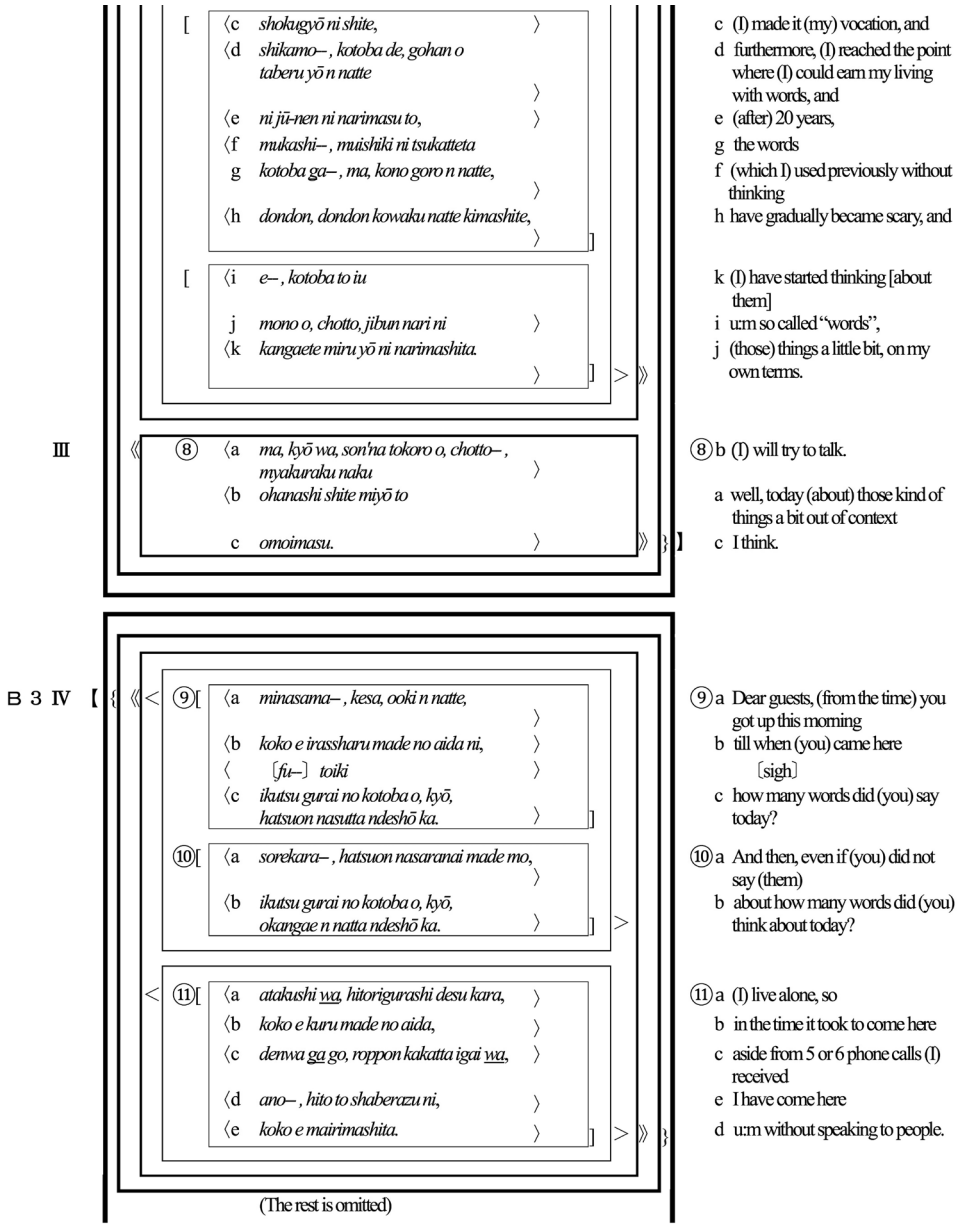
Grammatico-semantic paragraph IV opens a new topic realized in the text by the repetition of "*ikutsu gurai no kotoba o, kyō*—" (how many words did (you) say today?) in ⑨c and ⑩b. It is part of the super large grammatico-semantic spoken paragraph B (which belongs to the Development section (*tenkaibu* 展開部)). B begins by addressing the audience with the topic expression "*minasama*--" (Dear guests--), and then presenting its content in sentences ⑨ "*hatsuon nasutta ndeshō ka*" (did (you) say?) and ⑩ "*o kangae n natta ndeshō ka*" (did (you) think about), further impressing the audience with two repeated interrogative expressions.

In the omitted part after sentence ⑪, the lecturer provides answers using first person topic expressions as in ⑪a "*atakushi* WA" (I) and "*atashi* WA" (I), the lecturer proceeds with the theme of vocabulary in everyday use.

Grammatico-semantic paragraph I consists of the Opening a self-introduction with greetings in the Opening section and grammatico-semantic paragraph II introduces the topic touching upon the theme of the lecture, i.e., *"kotoba ga kowai"* (the words are scary). The large grammatico-semantic spoken paragraph I develops into grammatico-semantic spoken paragraph II. After grammatico-semantic paragraph II, topic expressions concerning '*kotoba* (words)' are repeated. Although first person topic expressions, such as "*watakushi* WA" (I), are ellipted in the grammatico-semantic paragraph I, and sentences ① and ②, they are expressed explicitly in grammatico-semantic paragraph II sentence ④a "*watakushi* WA" (I) and ⑤c "*watashi* WA" (I). The reason they are expressed explicitly is because after a different topic expression has been introduced in sentence ② "*Hotondo no kata* GA (most of you [present here]) a new topic expression, "*kotoba* GA" (the words) has been introduced.

(2)


A 1 I
2 II
① a Tadaima, goshōkai ni azukarimashita,
b Mukōda Kuniko de gozaimasu.
② a Hotondo no kata ga ohatsu ni ome ni kakaru n da to
b omoimasu.
③ a yoroshiku
b onegai shimasu.
④ 〈a ano–, watakushi wa, kono goro n natte–〉
〈b e–, kotoba ga–, totemo kowaku narimashita. 〉
⑤ 〈a ano–, ki ga tsukimasu to, 〉
〈b umarete 〉
〈c sugu ni–, hantoshi gurai de–, 〉
〈d 'manma' tte iu
e kotoba o, watashi wa shabetta n da sō desu keredomo– 〉
〈f sorekara gojūichi nen, 〉
〈g gojū–, ima, ichi n narimasu keredomo, 〉
〈h ano–, sono aida ni–, 〉
〈i zuibun takusan no kotoba o, ma, shabettari, 〉
〈j kangae tari, 〉
〈k kotoba to iu no wa, oto ni dasanakutemo, 〉
〈l kokoro no naka ni aru matomatta
m koto o omoeba, 〉
〈n sore wa, kotoba da, to
o omou n desu. 〉
⑥ 〈a shabettari, 〉
〈b omottari, 〉
〈c ma, saikin wa, kaitari,
d ma, shite, orimasu. 〉
⑦ 〈a hotondo, zenhan wa, muishiki de, tsukatteta to
b omou ndesu keredomo, 〉
① b (I) am Mukōda Kuniko
a who has just now been introduced.
② a It's that (this) is the first time (for me) to meet most of you
b I think.
③ a Kindly
b I (humbly) request (that you treat me well).
④ a U:m, I, at this time,
b u:m, words have become very scary for me.
⑤ a U:m, when (I) became aware
b after [I was] born
c immediately, in about half a year,
d "Mamma"
e they say it's that I spoke [those] words, but,
f since then 51 years,
g it has been 51 years, but
h um, during that time
i I have spoken, well, very many words,
j and have thought (about those words),
k it's that even if you do not pronounce the words
m if you think about things
l which make sense in your heart,
n those are words
o I think.
⑥ a Speaking,
b thinking,
c well, recently, writing
d (I) have been doing that.
⑦ a It's that (I) have used (words) unconsciously through nearly all of the first half [of my life]
b I think but,


[〈c shokugyō ni shite, 〉
〈d shikamo–, kotoba de, gohan o taberu yō n natte 〉
〈e ni jū-nen ni narimasu to, 〉
〈f mukashi–, muishiki ni tsukatteta
g kotoba ga–, ma, kono goro n natte, 〉
〈h dondon, dondon kowaku natte kimashite, 〉]
[〈i e–, kotoba to iu
j mono o, chotto, jibun nari ni 〉
〈k kangaete miru yō ni narimashita. 〉] > 》
Ⅲ 《 ⑧ 〈a ma, kyō wa, son'na tokoro o, chotto–, myakuraku naku 〉
〈b ohanashi shite miyō to
c omoimasu. 〉 》 }]
B 3 Ⅳ 【 { 《 < ⑨[〈a minasama–, kesa, ooki n natte, 〉
〈b koko e irassharu made no aida ni, 〉
〈 〔fu–〕 toiki 〉
〈c ikutsu gurai no kotoba o, kyō, hatsuon nasutta ndeshō ka. 〉]
⑩[〈a sorekara–, hatsuon nasaranai made mo, 〉
〈b ikutsu gurai no kotoba o, kyō, okangae n natta ndeshō ka. 〉] >
< ⑪[〈a atakushi wa, hitorigurashi desu kara, 〉
〈b koko e kuru made no aida, 〉
〈c denwa ga go, roppon kakatta igai wa, 〉
〈d ano–, hito to shaberazu ni, 〉
〈e koko e mairimashita. 〉] > 》 }
(The rest is omitted)
⑧ b (I) will try to talk.
a well, today (about) those kind of things a bit out of context
c I think.
c (I) made it (my) vocation, and
d furthermore, (I) reached the point where (I) could earn my living with words, and
e (after) 20 years,
g the words
f (which I) used previously without thinking
h have gradually became scary, and
k (I) have started thinking [about them]
i u:m so called "words",
j (those) things a little bit, on my own terms.
⑨ a Dear guests, (from the time) you got up this morning
b till when (you) came here
〔sigh〕
c how many words did (you) say today?
⑩ a And then, even if (you) did not say (them)
b about how many words did (you) think about today?
⑪ a (I) live alone, so
b in the time it took to come here
c aside from 5 or 6 phone calls (I) received
e I have come here
d u:m without speaking to people.

Mukōda Kuniko (向田邦子) '*Kotoba ga kowai*' (the words are scary) "*Shinchō kasetto kōen* (Shincho casette lectures)" 1991, Tokyo: Shinchōsha (underline and symbols added by Sakuma). A ~ B, 1 ～ 3 are "large grammatico-semantic spoken paragraphs" (*daiwadan* 大話段), I ~IV are grammatico-semantic spoken paragraphs[6], ①～⑪ are sentences, and a ～ o are clauses in respective sentences. Annotated original text is shown in Appendix 2.

In this lecture, several formal markers within a single utterance to indicate the multiple structure of grammatico-semantic spoken paragraphs are used. In addition, several formal markers can be observed that hint at the multiple structure created by the coherency function of grammatico-semantic spoken paragraphs, which belong to different dimensions.

Grammatico-semantic paragraph II consists of three small grammatico-semantic paragraphs (*shōwadan* 小話段), i.e., sentence ④, sentences ⑤ and ⑥, and sentence ⑦. In sentences ④ and ⑦, expressions related to the theme of the text (such as "*kono goro n natte*--" (it is in these days...), "*kotoba* GA --" (words), "*kowaku narimashita / natte kimashite*" (became very scary / (have gradually became scary) are repeated.

Grammatico-semantic paragraph II is made of sentences ④ through ⑦. Sentence ④, the topic/core sentence (中心文) in the grammatico-semantic paragraph II, is described in more detail in the sequence of sentences ⑤～⑦.

Because there were different new inserted topic expressions in the second half of the compound sentence ⑤ (i.e., ⑤k "*kotoba* TOIUNOWA" (words), and ⑤n "*sore* WA" (that)), the clause ⑤i "*shabettari*" (talk and...) from the first half of ⑤ is repeated in ⑥a, and reworded in ⑥b "*omottari*" (think and/thinking...), ⑥c "*saikin* WA, *kaitari*" (well, recently, I write), and ⑥d "*ma, shite, orimasu*" (well, doing).

Grammatico-semantic paragraph III begins with sentence ⑧ which hints in advance about a change in the topic (*wadai tenkan* 話題転換), with the filler "*ma*" (well). Subsequently, the overall theme of the lecture is presented with the topic expression ⑧a "*kyō* WA ***sonna*** *tokoro* O" (today, **such** points), and with the predications ⑧b "*ohanashi shite miyō* TO" (would like to talk) and ⑧c "*omoimasu*" (I think).

Grammatico-semantic paragraph III consists of sentence ⑧, the large topic/core sentence (*daichūshinbun* 大中心文). By imparting coherency to the preceding grammatico-semantic spoken paragraph II consisting of four sentences ④～⑦, the large grammatico-semantic spoken paragraph 2. In addition, this large grammatico-semantic spoken paragraph 2 derives coherency from large grammatico-semantic spoken paragraph 3 in the Development section (*tenkaibu* 展開部). The large

6 Expressions in brackets 【{《<[〈〉]>》}】 are grammatico-semantic spoken paragraphs, formed by larger and smaller coherently organized units based on theme. Ordering of brackets shows the hierarchy of layers from large to small. Brackets are visualized by enclosures.

grammatico-semantic spoken paragraph 3 includes grammatico-semantic paragraph IV, and forms the complex multiple structure of grammatico-semantic paragraphs over the whole discourse of this lecture. The fact that this multiple structure all results from the coherency relations among several sentences and grammatico-semantic spoken paragraphs within a single utterance is an important characteristic of the units of Japanese spoken discourse in monologue data.

5 Conclusions

If we think about the units of analysis of Japanese written text and spoken discourse from the point of view of the actually realized forms in communication, then the importance of "grammatico-semantic paragraph units beyond the sentence" as units whose coherency function organizes topics, becomes apparent. In other words, it becomes clear that the "grammatico-semantic paragraph" (*dan*) is the very unit for conveying linguistic information.

The medium for conveying linguistic information is not only sound and writing but also nonverbal expressions including gestures, visual images, etc. Thus, means of communication diversify in complex ways, and, when attempting to exchange the information content in a more effective way, communicative behaviour is accomplished by adapting to various stages and elements of the multiple structure of grammatico-semantic paragraphs in Japanese written texts and spoken discourse. In establishing valid units for the analysis and description of the entire process of expressing and understanding Japanese written text and spoken discourse, from its opening to its closing, the task faced by Japanese discourse analysis is to elucidate the dynamics of linguistic performance.

Valid units of verbal or written communication are important as foundation for detailed analysis of linguistic exchange, as is exemplified in the conjunctive relations analysis of example (1) in the Appendix 1.

There is still a need to strive to establish even more valid units of analysis: that will enable our understanding of both the processing of various kinds of nonverbal information and of the differences in the scale of all kinds of communicated expressions. These include "complex written texts and spoken discourses" (*fukubunshō/danwa* 複文章・談話), "hybrid written text and spoken discourse" (*kongōbunshō/danwa* 混合文章・談話), and "simple written text and spoken discourse" (*tanbunshō/danwa* 単文章・談話).

Acknowledgements

I would like to thank Andrej Bekeš for inviting me to join this project, for his useful comments, and for translating my text from Japanese into English. I would also like to thank Polly Szatrowski for her valuable comments during the translation process. Of course, the responsibility for the final text remains mine.

Literature

Hayashi S. (1998) *Bunshōron no kiso mondai* (Basic issues in text analysis). Tokyo: Sanseidō.

Ichikawa T.(1978) *Kokugo kyōiku no tame no bunshōron gaisetsu* (Outline of text theory for pedagogy of Japanese as a national language) Tokyo: Kyōiku shuppan.

Minami F. (1997) *Gendai nihongo kenkyū* (Studies of modern Japanese). Tokyo: Sanseidō.

Minami F. (2005) 'Gengo no tan'i' (Units of language), Nihongo kyōiku gakkai (ed.), *Shinpan nihongo kyōiku jiten* (Dictionary of Japanese language pedagogy - New edition), 537-538. Tokyo: Taishūkan shoten.

Nagano M. (1972) *Bunshōron shōsetsu* (Detailed text analysis). Tokyo: Asakura shoten.

Nagano M.(1986) *Bunshōron sōsetsu* (General remarks on text analysis). Tokyo: Asakura shoten.

Nomura M. (2000) *Nihongo no tekusuto: kankei, kōka, yōsō* (Japanese text: relations, effects, appearance). Tokyo: Hitsuji shobō.

Nagata H. (1995) *Kokugo bunshōron* (Text analysis of Japanese). Osaka: Izumi shoin.

Sakuma M. (1987) 'Bundan nintei no ichi kijun (I): teidai hyōgen no tōkatsu (A criterion for recognition of grammatico-semantic written paragraphs (I): coherency of thematizing expressions', *Bungei, gengo kenkyū - gengo hen* 11: 89-135.

Sakuma M. (1992) 'Bunshō to bun: dan no bunmyaku no tōkatsu' (Text and sentence: coherency of the paragraph context). *Nihongogaku* 11 - 14: 41-48.

Sakuma M. (2003) 'Bunshō-danwa ni okeru dan no tōkatsu kinō' (Coherency function of grammatico-semantic paragraphs in text and/or spoken discourse), Kitahara Yasuo (editorial supervision), Sakuma Mayumi (ed.) *Asakura nihongo kōza 7 bunshō-danwa*, 91-119. Tokyo: Asakura shoten.

Sakuma M. (2006) 'Bunsho-danwa no bunseki tan'i (Units for the analysis of Japanese written texts and/or spoken discourse) *Gekkan Gengo*. Vol.35, No.10: 65-67. Tokyo: Taishūkan shoten.

Sakuma M. (2012) 'Bunsho-danwa no bunseki tan'i (Units for the analysis of Japanese written texts and/or spoken discourse). *Gengo serekushon* (Selection from 'Gengo') Pt. 1, 93-100. (republication of Sakuma 2006) Tokyo: Taishūkan shoten.

Sugito S.(1984) 'Danwa no tan'i ni tsuite' (On the units of discourse), *Gengo seikatsu* 393: 34-41.
Sugito S. (1987) 'Hatsuwa no uketsugi (Inheriting the utterance)' Ch.2.2, National Institute for Japanese Language and Linguistics (ed.) *Danwa kōdō no shosō* (Aspects of discursive actions), 68-106. Tokyo: Sanseidō.
Szatrowski, P. (1993) *Nihongo no danwa no kōzō bunseki* (Analysis of Japanese conversation structure). Tokyo: Kuroshio shuppan.
Tokieda M. (1950) *Nihon bunpō kōgo hen* (Japanese grammar: spoken language). Tokyo: Iwanami shoten.
Tokieda M. (1960) *Bunshō kenkyū josetsu* (Introduction to text study). Tokyo: Yamada shoin.
Tsukahara T. (1966) 'Ronri-teki danraku to shūji-teki danraku' (Logical paragraphs and rhetorical paragraphs), *Hyōgen kenkyū* 4: 1-9.

Appendix 1

Original text in example (1)

①注意②だまされないで！

Ⅰ③○慌てない

Ⅱ④○振り込まない

Ⅲ⑤○本人に確認する

⑥警視庁

（『広報けいしちょう』16号(2005. 1. 16)（下線、筆者付す。）

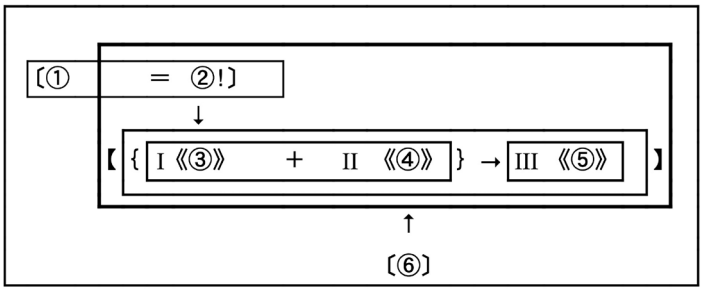


Figure 1. Conjunctive relations between units in Example (1)

Explanation of symbols (cf. Ichikawa 1978: 89-93):

1. ①→② causal/illative (*junsetsugata*)
2. ①Z② adversative (*gyakusetsugata*)
3. ①＋② additive (tenkagata)
4. ①⇔② contrastive (*taihigata*)
5. ①＝② illustrative (*dōretsugata*)
6. ①←② causative/supplementing (*hosokugata*)
7. ①↓② transitional (*tenkangata*)
8. ①－② sequential (*rensagata*)

Appendix 2

Original text in example (2)

A１I【{《①a ただ今、ご紹介にあずかりました、b向田邦子でございます。》
《②a ほとんどの方がお初にお目にかかるんだとb思います。}
{③a よろしくb お願いします。》}】

２II【{《＜④〈a あのー、わたくしは、**この頃んなってー**、〉〈b えー、言葉がー、とても怖くなりました。〉＞＜⑤［〈a あのー、気が付きますと、〉〈b 生まれて〉〈c すぐにー、半年ぐらいでー、〉〈d「マンマ」っていうe 言葉を、わたしはしゃべったんだそうですけれども、〉］［〈f それから五十一年、〉〈g 五十一（ごじゅういち）、今、一（いち）んなりますけれども、〉〈h あのー、その間にー、〉〈i ずいぶん沢山の言葉を、ま、しゃべったり、〉〈j 考えたり、〉］［〈k 言葉というのは、音に出さなくても、〉〈l 心の中であるまとまったmことを思えば、〉〈n それは、言葉だ、とo思うんです。〉］［⑥〈a しゃべったり、〉〈b 思ったり、〉〈c ま、最近は、書いたり、d ま、して、おります。〉］＞＜⑦［〈a ほとんど、前半は、無意識で、使ってたとb思うんですけれども、〉］［〈c 職業にして、〉〈d しかもー、言葉で、ご飯を食べるようんなって、〉

〈e 二十年になりますとー、〉〈f 昔ー、無意識に使ってた g 言葉がー、ま、この頃んなって、〉〈h どんどん、どんどん怖くなってきまして、〉〕〔〈i えー、言葉というｊものを、ちょっと、自分なりに〉〈k 考えてみるようになりました。〉〕＞》

Ⅲ 《⑧〈a ま、今日は、そんなところを、ちょっとー、脈絡なく〉〈b お話ししてみようとｃ思います。〉》〕】

B３Ⅳ【{《＜〔⑨〈a 皆様ー、今朝、お起きんなって、〉〈b ここへいらっしゃるまでの間に、〉〔「フッー」吐息〕〈c いくつぐらいの言葉を、今日、発音なすったんで しょうか。〉〕⑩〔〈a それからー、発音なさらないまでも、〉〈b いくつぐらいの言葉を、今日、お考えんなったんでしょうか。〉〕＞＜〔⑪〈a あたくしは、独り暮らしですから、〉〈b ここへ来るまでの間、〉〈c 電話が五、六本かかった以外は、〉〈d あのー、人としゃべらずに、〉〈e ここへ参りました。〉〕＞》〕} （以下、省略）

（向田邦子「言葉が怖い」『新潮カセット講演』1991 新潮社）

（各種の下線・記号は筆者付す。A～B、１～３、Ⅰ～Ⅳは話段、①～⑪は文、ａ～ｏは節の番号を示す。下線＿＿は時間表現、＿＿に□囲みは提題表現、＿＿と＿＿は大小の中心文を示す。）

要旨（Abstract in Japanese）

「日本語の文章・談話における分析単位」

佐久間まゆみ（早稲田大学）

「文章」と「談話」は、それぞれ、文字と音声を伝達媒体とする最大かつ最も具体的な言語単位である。いずれも、日本語のコミュニケーションの唯一の実現形態として、言語行動の動態的なまとまりを表す完結統一体とされる。

文章・談話の単位に関する課題は古くて新しい。文章の単位は、時枝(1950:289)以来、「文章の成分」として「文」「連文」「段落」「文段」等のいずれが「文章の成分」として妥当かが論議され、談話の単位も、南(1997:295-356)が「文章」「会話」「談話」の単位認定の手がかりを問い、林(1998:394-396)も「コミュニケーション単位」という「質的単位」が必要だとしている。

日本語学の「文章・談話論」における言語行動を包括的に記述する有効な分析単位として、「文章」と「談話」の異同の検討が不可避の課題である。文章における文のまとまりからなる「文段」、談話における発話のまとまりからなる「話段」の「統括機能」の「多重構造」を中心に、文章・談話の有効な分析単位の可能性を探る。

2 Lexical cohesion and text-organizing function in the Japanese text: A Japanese text linguistics proposal

TAKASAKI Midori
Ochanomizu University

Abstract
This paper discusses the concepts of "text-organizing words" and "cohesion" and reports results of how they are used in some Japanese texts. These concepts are a part of the larger group of concepts of 'textuality' that establish a text as a text. Text-organizing words divide a stream of text, according to which they have the function of structuring the text (or a subsection thereof). Cohesion brings semantic consistency to the text (or a subsection thereof) by forms of language having relationships with each other. The relationship between text-organizing function and cohesion, in short, will be such that the former is realized by the latter. Here I bring forth the concept of "semantic segments" as a kind of work unit that semantically organizes the text.

Keywords: vocabulary, text-organization, cohesion, segments, demonstratives

1 Introduction

"Text" here is used as a term that refers to a certain body of written language — writing that has been written for the purpose of a literary work, the news, an advertisement, criticism and explanation or expression of opinion, etc.[1] The term "text" is used when such a body of written language as this is taken up as the subject of language study. Textuality is not the simple accumulation of words and sentences, but, rather, refers to a property that establishes text as text and enables the conveyance of its contents and intention to the reader. I will consider lexical cohesion and the text-organizing functions that are related to the establishment of textuality.

First of all, I think that the following five conditions are necessary for textuality to be established in a certain body of language.

1. The text has attributes that distinguish it from other things outside of itself. Its unity and completeness are its crucial attributes.
2. The text as an independent document exists in relationship with other texts outside of itself (that is to say, it possesses intertextuality).

1 Regarding the range and genre of the written language, see Ichikawa (1978:36-37) and Takasaki and Tachikawa (2010:175-179).

3. The inside of the text is semantically connected by explicit verbal signals (=cohesion) and semantic segments are formed.
4. The text forms consistency by generating a multilayered structure of semantic segments[2] inside and it can be one whole for the outside.
5. Textuality is acknowledged by readers. The readers understand the dynamism of development with linear and temporal properties within the text, experience the existence of cohesion and the formation of semantic segments, and can recognize the consistency of the text when they come to the end of the text.

Let's take a book as an example and consider its "textuality." The unity of the book is, for example, shown by the title, author's name, table of contents, and headings as well as the body of the text. The textuality is defined by this unity of the book, which is closed off from external entities.

While reading, formation of semantic segments is helped by cohesion. They are correlated with each other, reiterated, and completed with clues of text-organizing words. Clusters of semantic segments appear coherently and consistently throughout the text. The text is finished when this dynamism is physically cut off by the end of the book.

The reason why such textuality is possible is that individual linguistic forms having grammatical function and lexical meaning are concerned in cohesion and text-organization. From another perspective, all the linguistic forms including the word can be said to have function and characteristics shown in text. Even a smaller unit such as a character is related to the cohesion of text.

In other words, concerning logographic *kanji* characters, for example, in a sentence about university students finding employment, a Sino-Japanese word *shoku* 職 'job' is taken up from the word *shūshoku* 就職 'finding an employment,' and becomes a part of Sino-Japanese words such as *shokugyō* 職業 'occupation,' *shokushu* 職種 'type of job,' and *rishoku-ritsu* 離職率 'rate of quitting a job.' Further, those Sino-Japanese words become a part of compounded words such as *shūshoku katsudō* 就職活動 'job hunting' and *shokugyō sentaku* 職業選択 'career choice.' Reading a newspaper article or an editorial carefully, we can find more than a few phenomena of these alignments and realignments involved in the formation of context.

Therefore, it is significant to approach Japanese text linguistics as it is explained below.

2 Semantic segments are discussed later. Cf. p35.

2 What is "Japanese text linguistics"? — On the actual situation of words' behaviour in text

In studies of language where the written discourse is taken up as the maximum unit of language, the material or object of the study is referred to as "text". The studies are usually concerned with its formation, structure, organization, context formation, development, cohesion, consistency, expression, style, etc. Such studies are referred to as *bunshoron* 文章論 'theory of written text' within Japanese linguistics, or as "text linguistics" in English.

This paper extends the scope of language study which deals with such "text," by observing behaviour at the level of lexis, grammar, and orthography in the whole text as its subject of study, and proposes a relationship between this behaviour and "text linguistics" as mentioned before. I want to propose this approach as a possible methodology of "Japanese text linguistics."

Takasaki (2011) stated: "Concerning 'theory of written text' I want to focus more attention on differences of approach in analyzing the objects in comparison with usual approaches in lexicology and grammar rather than focusing on enlarging the size of units of analysis (word→sentence→passage)." The same is true even when 'theory of written text' becomes 'text analysis.' That is to say, text analysis should document general *tendencies* rather than strict *rules*, identify *behaviours* rather than *functions*, and emphasize a method of *qualitative* analysis over *quantitative* analysis. These differences in approach are crucial to my method of text analysis. They could provide more effective methodology for lexicology, grammar, and orthography. Note that the term 'behaviour' above refers to a flexible way of working according to circumstances that is not as rigid as theoretical notion of 'function.' The term 'behaviour' will be used hereafter in text analysis, whereas it would be often called "function" in grammar.

Now, the 'behaviour' of words in text is considered below from the viewpoints of text-organizing words and lexical cohesion, which is based on the results of text analysis in Takasaki (1976, 1985, etc.).

3 On lexical function in a sentence: from the viewpoint of "text-organizing function"

Takasaki (2013) examined what kind of function words have in a sentence from the standpoint of text-organizing function and cohesion. Using a corpus[3] of introductory

3 The corpus used was *Gakujutsu Nyūmon-sho Kōpasu* 学術入門書コーパス 'Corpus of Introductory Science Textbooks' made in the project '*Bunshō ni okeru Goi no Bunpu to Bunshō Kōzō* 文章における語彙の分布と文章構造' 'Distribution of the Vocabulary in the Sentence and Sentence Structure' by National Institute for Japanese Language and Linguistics (Project Leader: Makoto Yamazaki). The following 4 types, 976 pages, and 194000 characters were used

science textbooks as material, I examined some examples and observed how words functioned in an actual text and how they built up that text.

As a result, some tendencies were observed as described below.

1. The most important words that undertake text-organization are nouns. Sino-Japanese words, which tend towards a higher level of abstraction compared with other categories of words, undertake much of this task.
2. The lexical cohesion of a text contributes greatly to the unity of semantic segments throughout the text.
3. There are some relationships of cohesion between the text-organizing words and the words inside the semantic segments that are combined with them.
4. Demonstratives contribute to signalling of text-organization in many cases.
5. Relative abstractness of text-organizing words actually observed and cohesion of words does not always reflect the system that is provided theoretically in lexicology, such as synonyms, superordinate or subordinate relationships. Rather, there are many temporary cases where they are affected by context, which surely guarantee originality and a one-time-only nature of the text.

Items 1-5 will be explained in the next sub-sections. To begin with, basic concepts of "text-organizing function," "segments," and "lexical cohesion" will be briefly stated below.

3.1 On "text-organizing function," "segments," and "lexical cohesion"

Concerning text-organizing function, McCarthy (1991:75) used the term "discourse-organizing words" for words whose job is to organize and structure the argument, rather than to answer for its content or field. Taking inspiration from the term 'discourse-organizing words,' in this paper, I will use the term 'text-organizing function' for a function that gives organization and structure to text. Takasaki (2011) simply used McCarthy's term "discourse-organizing words." However, this paper refers to the concept of "discourse-organizing words" as 'text-organizing words,' and to the concept of "discourse organizing function" as 'text-organizing function,' so as to clearly indicate that it is specifically written works that are under consideration. There are various theories and opinions about the terms "discourse" and "text," so I will adopt a simple method of explanation here.

---

from the corpus: *Seiji-gaku Nyūmon* 政治学入門 'An Introduction to Political Science,' Abe, H., Iwanami Textbooks; *Nippon Gaikō-shi Kōgi* 日本外交史講義 'Lecture on the Japanese Diplomatic History,' Inoue, T., Iwanami Textbooks; *Amerika no Keizai* アメリカの経済 'The economy in America' 2nd ed., Haruta, M. and Suzuki, N., Iwanami Textbooks; *Keihō Genron* 刑法原論 'A Basic Principle of Criminal Law,' Naitō, K., Iwanami Textbooks.

McCarthy (1991) classified different types of words as 'grammar words' and 'lexical words[4]' and considered "discourse-organizing words" as words having a function intermediate between the two, which was noteworthy for purposes of text analysis.

Examples of such words are: 'issue,' 'problem,' and 'dilemma,' which, in the words of McCarthy (1991: 74-75) "... stand in place of segments of text just as pronouns can; a segment may be a sentence, several sentences or a whole paragraph, or more."

That is to say, the range which the word indicates —what part of the contents of the text does 'issue' point at? Or, what and what does "dilemma" refer to?— becomes a 'segment.' And, some of the discourse-organizing words give us indications of the larger text-patterns the author has chosen, and build up expectations concerning the shape of the whole discourse (McCarthy 1991:74-75).

McCarthy (1991)'s phrase "just as pronouns can" suggests that language forms which become text-organizing words have such simple forms and meanings as to substitute and represent concrete things. Their level of abstraction and generality are considered relatively high compared with most other categories of vocabulary, assuming that formal nouns such as *mono* もの 'things' and *koto* こと 'matters' are the forms of the highest level of abstraction in meaning.

'Segment' refers to the content of the text which is integrated on the basis of such text-organizing words. However, by 'segment' this paper does not mean customary divisions such as a paragraph, passage, or some large or small portion of simple linguistic forms. Instead, the segment refers to a 'unit of meaning,' in other words, semantic unity is given to the part of text that was chosen in accordance with the instruction of a particular text-organizing words.

Hence, this paper refers to such segments as 'semantic segments.' Semantic segments are considered to possess certain verbal signals, through which it will be possible to concretely divide the internal parts of the text and pick them out. The clues could be the relationships of cohesion that exist within the set of text-organizing words and segments, or demonstratives, modifiers, and determiners that are referred to as text-organizing words. Semantic segments somewhat resemble the linguistics concept of double articulation. They are lower-level semantic units which come together to form meaning in the text. Also, semantic segments could be mutually piled up, included in each other, and capable of combination.

Lexical cohesion is observed in co-texts within the text. Firstly, text-organizing words and vocabulary within segments have cohesive relationships. Secondly, synonymous rewording and reiteration in words within segments are also regarded as lexical cohesion.

---

4 McCarthy (1991: 74) stated: "This distinction also appears sometimes as *function* words versus *content* words, or *empty* words versus *full* words. The distinction is a useful one: it enables us to separate off those words which belong to *closed systems* in the language and which carry grammatical meaning, from those that belong to *open systems* and which belong [sic] to the major word classes of noun, verb, adjective and adverb."

Halliday and Hasan (1976:8) stated that:

> Cohesion is a semantic relation between an element in the text and some other element that is crucial to the interpretation of it.

According to Halliday and Hasan (1976), 'lexical cohesion' in the linguistic system is represented by 'reiteration[5] (identity of lexical reference)' and 'collocation (similarity of lexical environment),' while 'grammatical cohesion' is represented by 'reference, substitution, ellipsis, and conjunction.'

The sequence of a text as a whole is segmented formally by divisions of paragraph and sentence, which are also regarded as text-organizing means. However, what I want to consider here is the case where text-organizing function emerges in the relationship of vocabulary and text. Such a way of thinking is often seen in previous studies that observe the division of meaning and content in the text by focusing on cohesion (lexical cohesion and grammatical cohesion) and the function of conjunction.

Each semantic segment is indicated by a semantic or contextual break in the text. A part of a sentence, a part of a paragraph, a few sentences, or a few paragraphs can be chosen as a segment. Or, it could be obtained by extracting a specific proposition and topic that emerge from the interplay of text-organizing words and the context. This reminds us of the viewpoint that "a text, after all, is not a unit of form but of meaning (Halliday and Hasan 1989:94)." A text is constituted by semantic segments, combination of semantic segments, and the correlation of inclusive relations, so that the intention of the text is realized. In order to read and understand deeply the text of an extended work of scientific prose, it is necessary to create large and small semantic segments based on some keywords, and make them correspond and relate to each other. And sketching the plot with these keywords is more efficient than summarizing what the writer wants to say in every paragraph.

3.2 The most important words that undertake text-organization

As described previously, "*1. The most important words that contribute to text-organization are nouns. Sino-Japanese words, which tend towards a higher level of abstraction compared with other categories of words, undertake much of this task.*"

Takasaki (2013) pointed out some aspects of words such as *gen'in* 原因 'cause,' *mondai* 問題 'problem,' *ten* 点 'point,' and *ugoki* 動き 'motion.' Even a single word of this type occurring in a much larger body of text can assume a text-organizing function for semantic segments together with various kinds of support and intervention from the

5 "Reiteration" is the repetition of a lexical item; synonym; superordinate; general word (nouns having a general referent such as people, stuff, and move); and personal reference. "Collocation" means "to share the same lexical environment," and two lexical items that tend to occur in the similar context (Halliday and Hasan 1976).

context. Takasaki (2013) also pointed out that plenty of iteration and relating words and phrases inside such semantic segments contribute to cohesion, and that semantic segments can be identified by such cohesion. The combination of these semantic segments attains the purpose of the text.

Takasaki (1988), where newspaper editorials were used as materials, sums up the following points: Many text-organizing words were nouns. Sino-Japanese words made of two Chinese characters were used abundantly. Chinese characters have meanings. Sino-Japanese words made of two Chinese characters can be combined to form a nonce word, can become separated into individual Chinese characters. Moreover, the separated individual Chinese character can form another Sino-Japanese word through compounding with additional Chinese characters. Such dynamic usage of Chinese characters contributes to the formation of the context.

Demonstratives often play an auxiliary role for text-organization. Noticing this, Takasaki (1988) examined if nouns with demonstratives are involved in text-organization by corresponding semantic segments in text. Such nouns with demonstratives were extracted from editorial columns in *Asahi*, *Mainichi*, and *Yomiuri* newspapers during August 1-31, 1987, grouped by meaning, and listed below. This categorization is based on Takasaki (1988, etc.).

Terms pertaining to thought and logic:

ikikata 行き方 'a way to go,' *ishiki* 意識 'conscience,' *omoi* 思い 'thought,' *kangaekata* 考え方 'way of thinking,' *kanten* 観点 'viewpoint,' *kitai* 期待 'expectation,' *kimochi* 気持ち 'feeling,' *gimon* 疑問 'question,' *keikaku* 計画 'plan,' *keiken* 経験 'experience,' *ketchaku* 決着 'settlement,' *kettei* 決定 'decision,' *kokoromi* 試み 'trial,' *jikaku* 自覚 'awareness,' *shuhō* 手法 'technique,' *jōhō* 情報 'information,' *seisaku sentaku* 政策選択 'choice of policy,' *tēma* テーマ 'theme,' *tenbō* 展望 'prospects,' *nanmon* 難問 'difficult problem,' *ninshiki* 認識 'understanding, recognition,' *hairyo* 配慮 'consideration,' *hassō* 発想 'idea,' *hansei* 反省 'reflection,' *handan* 判断 'judgment,' *hōsaku* 方策 'means,' *hōshiki* 方式 'procedures,' *hōshin* 方針 'policy, course,' *mondai* 問題 'problem,' *yosoku* 予測 'prediction,' *rinen* 理念 'principle,' *rei* 例 'example,' and *ronri* 論理 'logic.'

Terms pertaining to language:

kankoku 勧告 'advice,' *giron* 議論 'argument,' *kugen* 苦言 'frank advice,' *koe* 声 'voice,' *kotoba* 言葉 'words,' *shuchō* 主張 'claim,' and *hihyō* 批評 'review.'

Terms pertaining to time:

katei 過程 'processes,' *aida* 間 'intervals,' *kiun* 機運 'mood,' *kikai* 機会 'opportunity,' *sai* 際 'in case of,' *jiki* 時期 'period,' *jiten* 時点 'point in time,' *toki* 時 'time,' and *baai* 場合 'case.'

Terms pertaining to spatial relations:
kakudo 角度 'angle,' *kyokumen* 局面 'aspect,' *kuiki* 区域 'area,' *naka* 中 'in,' *chiiki* 地域 'area,' *ten* 点 'point,' *bubun* 部分 'part,' *bun'ya* 分野 'area,' and *men* 面 'aspect.'

Terms pertaining to conditions:
genjō 現状 'present conditions,' *jōkyō* 状況 'situation,' *jōsei* 情勢 'state of affairs,' *jōtai* 状態 'state, circumstances,' *taisei* 態勢 'condition, attitude' and *tachiba* 立場 'standpoint.'

Terms pertaining to situations:
koto こと 'matters,' *genjitsu* 現実 'actuality,' *genshō* 現象 'phenomenon,' *jiken* 事件 'case,' *jijitsu* 事実 'fact,' and *jitai* 事態 'situation.'

Terms pertaining to quantity:
ketsuraku 欠落 'omission,' *sa* 差 'difference,' *suijun* 水準 'level,' *sūryō* 数量 'amount,' *teido* 程度 'degree,' *ninzū* 人数 'number of people,' and *hiritsu* 比率 'ratio.'

Terms pertaining to abstract relationships:
kekka 結果 'result,' *gyappu* ギャップ 'gap,' *jirenma* ジレンマ 'dilemma,' *jōken* 条件 'condition,' *baratsuki* バラつき 'unevenness,' and *mokuhyō* 目標 'goal.'

Terms pertaining to processes:
akujunkan 悪循環 'vicious circle,' *ikisatsu* いきさつ 'sequence of events,' *ugoki* 動き 'motion,' *undō* 運動 'exercise,' *kōyō* 高揚 'uplift,' *gōrika* 合理化 'rationalization,' *tenkan* 転換 'switch,' *nobi* 伸び 'growth,' and *henka* 変化 'change.'

These words are considered to function more or less as text-organizing words. There are loan-words *gyappu* ギャップ 'gap' and *jirenma* ジレンマ 'dilemma,' native Japanese words *koto* こと 'matters' and *nobi* 伸び 'growth,' and Sino-Japanese words consisting of the single Chinese character *ten* 点 'point' and the single Chinese character *men* 面 'surface' in the list above. The largest number is Sino-Japanese words made of two Chinese characters such as *mondai* 問題 'problem' and *hōshin* 方針 'policy, course.' What is interesting is that some of the above text-organizing words are common to the ones pointed out in Takasaki (2013), which showed the results of the investigation into introductory science textbooks. Namely, they are *mondai* 問題 'problem,' *ten* 点 'point,' *ugoki* 動き 'motion,' and *gyappu* ギャップ 'gap,' etc. According to Kim (2012), loan-words have increasingly come into their own as basic words recently. Hence, the number

of loan-words which are concerned with text-organization may well be in the process of increasing today.

In examples from works such as editorial columns and introductory science textbooks, nouns come after demonstratives and are more concerned in text-organization than are other parts of speech. Nouns are used to summarize the previous context plainly, to increase the degree of abstraction, and to recapture the whole text. In addition to it, another reason could be that nouns have flexibility to be brought into later development as attributive and predicative modifiers, or as a subject and theme. Requirements of text-organizing words are considered to be the following: their contextual flexibility is high (cf. Takasaki 1976); their semantic level of abstraction is relatively high; and they are used quite frequently. In addition, they are not so much "lexical words" as "grammar words," as McCarthy calls them.

In this connection, a word becoming a grammar word through the process of grammaticalization is deeply involved in the existence of a text in various ways. Grammaticalization as a phenomenon can only occur over the course of a text. Also, a requirement of grammaticalization, that is, extensive and frequent use, is naturally satisfied during frequent use of such lexical items in various texts.

The process of grammaticalization shows that the use in the concrete meaning and the use in the formal meaning coexist and that, although having width of multiple meanings, the use in the formal meaning gradually becomes dominant in the course of time. Grammaticalization occurs as textual phenomenon because the development of the text is superposed with the process whereby superordinate words bundle up and generalize subordinate words, and because the logical development to arrive at one abstraction from numerous concrete things is in accord with our natural thought process in reading editorial columns and introductory science textbooks.

As for parts of speech, nouns and noun-like phrases accounted for most, and adjectives, adjectival verbs, and verbs accounted for few.

For instance, as for verbs with postpositional particles and auxiliary verbs, there are examples such as *Naze kō natta ka* なぜこうなったか 'Why did it become this way?' *Konoyōni mite kuru to* このようにみてくると 'When I look (at it) in this manner,' *Sō de aru nara* そうであるなら 'If it is so,' and *Sō suru koto de* そうすることで 'In doing so.' They are considered to be text-organizing words because they have semantic segments which correspond to (or combine with) *naru* なる 'become,' *miru* みる 'look,' *aru* ある 'is,' and *suru* する 'do' beforehand. However, as far as editorial columns and introductory science textbooks are concerned, verbs are limited qualitatively and quantitatively compared with nouns. It seems to be uncommon that verbs actively participate in the development of such forms as naming, metaphor, interpretation, and opinion.

In the cases of adjectives (*i*-adjectives) or adjectival verbs (*na*-adjectives), for example,

Tōdai ikaken wa, nyūin kanja no rassa-netsu kansen o kakunin shi nagara, hōkoku ga yonkagetsu chikaku okure ta. Senmon-ka ga densenbyō ni taishi, kono yō ni rūzu de ii no daro u ka.

東大医科研は、入院患者のラッサ熱感染を確認しながら、報告が四か月近く遅れた。専門家が伝染病に対し、このようにルーズでいいのだろうか。

(The Tokyo University Institute for Medical Sciences confirmed the inpatient's infection with Lassa fever. However, the report was nearly four months late. Is it acceptable for an expert to be so lax in response to an epidemic?)

(*Rassa-netsu ga Nippon ni jōriku shita.*『ラッサ熱』が日本に上陸した 'Lassa fever struck Japan,' the editorial column, *Yomiuri* newspaper, August 17, 1987; English translation by Takasaki.)

An adjectival phrase, as used in the example above, is not quite so abstract as a verb, and often reflects aspects of the writer's viewpoint such as evaluation and interpretation. However, such examples are also quite limited in frequency and scope compared with nouns.

Considering McCarthy's method of intermediary positioning between lexical word and grammar word, it is naturally possible that some words assume the role of lexical word while others assume the role of grammar word, and still others assume an ambiguous interpretation. For exact text-organization, we need a lot of words that have become attenuated in meaning while still preserving their lexical meanings, and yet have not quite finished becoming grammar words either.

In other words, text-organization does not so much mean that a specific word independently and exclusively takes on all the work but that, the word functions according to its *contextual* meaning above and beyond its ordinary lexical meaning. Also, it means that, for specifying semantic segments that constitutively present text, the word does a selective and designated work with demonstratives and modifiers in some cases.

A certain interest to such phenomenon is shown from the standpoint of lexicology. Based on Takasaki (2011), Saito (2011) takes a viewpoint of "what establishes the association with word and sentence" and stated about "functionality of the meaning of a word" as follows:

> 'Functionality of the meaning of a word' means the following: there are cases where a word, with its meaning, necessarily performs a certain function in a passage, or it is consequently made to perform a special function from the relation with the content of the passage. Some examples of the former are 'discourse-organizing words' and 'proper nouns,' etc., which Takasaki mentioned. Some examples of the latter are 'keywords,' 'theme,' and 'title,' etc. What is important is that the function of the former is based on abstract meanings

> of specific Chinese characters, often independent from the context. On the other hand, the function of the latter is defined by its relation to the context. In this sense, the former is more interesting than the latter in lexicology (Saito 2011:271; English translation by Takasaki).

Saito (2011) additionally pointed out that words that serve as text-organizing words have inherent, specific characteristics.

3.3 Japanese *ko*, *so*, *a*, and *do* demonstratives contribute to signalling of text-organization in many cases

Since demonstratives are strongly coupled to the other parts of the sentences, McCarthy and Halliday see dependence there and get them into grammatical cohesion, which will be related to the following: text-organizing words are often accompanied by demonstratives after all. Takasaki (2013) named this the "text-organizing auxiliary function" of demonstratives.

As stated before, typical "text-organizing words" are considered to be lexical words with attenuated meaning overtly presented and accompanied with demonstratives "like pronouns."

Although demonstratives are not a required element, they do have a text-organizing auxiliary function. Therefore, text-organizing words' function is more conspicuous when demonstratives are attached to them. Demonstratives are categorized into "grammar words" (functional words) as a "closed system" in McCarthy (1991). In this case, it can be said that grammatical words help lexical words to show their functional aspect rather than their lexical meanings, and draw them towards grammaticalization. In other words, looking for text-organizing words by a corpus search, Japanese *ko* こ 'this', *so* そ 'that', *a* あ 'that', and *do* ど 'which' demonstratives could be clue words of the search. Since they form specific strings of *hiragana* characters, they are easily found and observed in corpus.

That is to say, typical "text-organizing words" are lexical words with attenuated meaning which are accompanied with demonstratives and overtly presented "like pronouns," as stated before. Although demonstratives are not a required element, they have grammatical cohesion themselves, similar to pronouns. Therefore, terms functioning as text-organizing words are more conspicuous when demonstratives are attached to them.

This being the case, let's begin with the question of what kind of function demonstratives have in text? Takasaki (1990a) showed some viewpoints of the study concerning function of demonstratives in sentence and discourse, and pointed out that a demonstrative sometimes performs not only a work of indication but also, in a larger range than discourse and consecutive sentences, the following works: I summarize them

as A-E and F below. Note that "demonstrative phrase[6]" refers to a combination of words and phrases such as "a demonstrative + α" like *kō-shita jōkyō* こうした状況 'such situation.' In this case, a large part of "α" is noun and noun phrase. That is to say, it corresponds to a "text-organizing word" as referred to in this paper.

A. A demonstrative phrase which indicates a wide range will greatly affect the structure of the whole sentence (Takasaki 1990a: 40).
B. Both the unifying function of anaphora and the notifying function of cataphora, which are contrasting works performed by demonstrative phrases, play an important role in sentence structure (Takasaki 1990a: 41).
C. It plays an important role in specifying the contents and range indicated by words such as *ketsuron* 結論 'a conclusion' and *wadai* 話題 'a topic' which follow demonstratives in demonstrative phrases (ex. *kono yō na ketsuron* このような結論 'such a conclusion' and *sonna wadai* そんな話題 'such a topic'), which correspond to α in "a demonstrative + α" (Takasaki 1990a: 44).
D. There are demonstrative expressions whose indications are not recognized nor thought of by the listener, such as *sōda* そうだ of 'a spur-of-the-moment idea' and *sōda, sōda* そーだ、そーだof 'making agreeable responses in the spoken language' (Takasaki 1990a: 43).
E. In written language, a writer will be aware of the readers and use demonstratives of *a*-series so that mutual understanding is realized in text (Takasaki 1990a: 38).

These functions as above were pointed out in Takasaki (1990a). Though I did not mention it in Takasaki (1990a), I would like to further add the following "F" on *do*-series as a function of demonstratives in text:

F. Demonstratives from the *do*-series which appear at the beginning of text give notice beforehand of the theme of the subsequent development, and their questioning power lasts, pending all the while, by means of the cohesion of words and phrases, until segments on the theme are brought to a conclusion.

For example, Takasaki (2013) showed the following sentence from Chapter 4, *Shakai Shūdan to Seiji* 社会集団と政治 'A Social Group and Politics' in *Seiji-gaku Nyūmon* 政治学入門 'Introduction to Political Science': after having stated the need of the appointment of women, *Sono tame ni wa, gutaiteki ni dono yō na hōsaku ga kangaerareru de arou ka.* そのためには、具体的にどのような方策が考えられるであろうか。'To that end, what kinds of plans are thought about concretely?' Then the content of 'plans' is described, and it follows that: *Waga kuni de wa mokka no tokoro kō-shita hōsaku ga*

6 Takasaki (1990a) originally used the term *shiji hyōgen* 指示表現 'demonstrative expression,' not *shiji goku* 指示語句 'demonstrative phrases.' However, both of these two terms refer to the same contents. This paper uses the term *shiji goku* 指示語句 'demonstrative phrases' in accord with Takasaki (1988).

torareru mikomi wa usui. 我が国では、目下のところこうした方策がとられる見込みは薄い。'For the time being, there is not much likelihood that such plans will be realized in our country.' Semantic segments received by the phrase ***kōshita hōsaku*** こうした方策 'such plans' become unified above.

【*do*-series demonstratives＋～interrogative word：*ka*】has an aspect of expression working towards the reader. It is noteworthy that it has the function of backward segmentation, opposite of such words as *kono yō na* このような 'like this.' In this case, it means to segment the part after the description of the 'plan,' and it announces and guarantees in advance that they will certainly be referred to afterwards. Phrases of indefinite *do*-series demonstratives have such a powerful text-organizing function that they become pending all the while until the indefinite part becomes a definite part with the conclusion of segmentation and correspondence. Also the following "preface" forms extensive segments concerning text-organization, and has consistency that we can see into the structure of the whole text there: *Honsho wa～ga dono yō ni～shita ka o kaimei shita mono de aru* 本書は～がどのように…したかを解明したものである 'This book elucidated how...'

Based on the above observation, it is clear that the function which demonstratives show in text is based on an inherent property and function of demonstratives, and on the differences among *ko, so, a,* and *do* demonstratives.

3.4 Cohesion of words

As mentioned at the beginning of Section 3 about the tendencies of text-organizing words, "*3. There are some relationships of cohesion between text-organizing words and the words inside the semantic segments that are combined with them.*" Furthermore, "*5. Relative abstractness of text-organizing words actually observed and cohesion of words does not always reflect the system that is provided theoretically in lexicology, such as synonyms, superordinate or subordinate relationships. Rather, there are many temporary cases where they are affected by context, which surely guarantee originality and once and for all characteristics of the text.*"

For example, Takasaki (2013) gave the following column of *9. 11 Tero no shōgeki* テロの衝撃 'Shock of September 11th Terrorism.' The following is what the column says about the shock of terrorism:

> *Keizaiteki eikyō ni kagitte mo… hamon wa chōki ni wataru. Koko de wa chokugo no keizai mondai o shōkai suru. Mottomo chokusetsuteki na dageki o uke ta no wa kōkū un›yu de aru ga, tōsho no unkō teishi, saikai go mo keibi kyōka ni yoru jūtai ya ryokō tebikae ni yoru ryokaku no genshō nado ni yori…ryokō gyōkai ga dai dageki o uke, kouri uriage mo ichiji ōkiku ochikon da. Hoken gaisha wa kyogaku no shiharai mondai ni chokumen shi, seizōgyō de wa…keiki no kakō wa kono jiken de ketteiteki ni natta to itte yoi.*

> 経済的影響にかぎっても（中略）波紋は長期にわたる。ここでは、直後の経済問題を紹介する。最も直接的な打撃を受けたのは航空運輸であるが、当初の運行停止、再開後も警備強化による渋滞や旅行手控えによる旅客の減少などにより（中略）旅行業界が大打撃を受け、小売売上げも一時大きく落込んだ。保険会社は巨額の支払い問題に直面し、製造業では（中略）景気の下降はこの事件で決定的になったといってよい。
>
> 'Even just limited to economic influence, the ripple lasts for a long term. Here I introduce economic problems immediately after the event. It is air transportation that has received the most direct blow. Their operations were halted at first. Even after the operations were restarted, congestion occurred because the security was reinforced and passengers decreased because they cut down on travelling. … Travel industry suffered great damage and retail sales significantly dropped for a while. Insurance companies faced the problem of a large amount of payment. As for manufacturing industry… it can be said that the drop of the economy became decisive because of this incident (English translation by Takasaki).'

And a long description in this editorial still continues. When the word *mondai* 問題 'problem' appeared in the phrase *keizai mondai* 経済問題 'economic problem,' a previous notice of stating the content of that *mondai* 問題 'problem' comes next, and that range is segmented as *keizai mondai* 経済問題 'economic problem.'

Inside the segmented part are words such as *dageki* 打撃 'blow,' *jūtai* 渋滞 'delay,' *genshō* 減少 'decrease,' *shōgai* 障害 'obstacle,' *dai dageki* 大打撃 'severely wounding,' *ochikonda* 落ち込んだ 'dropped,' *kon'nan* 困難 'difficulty,' *jakuten* 弱点 'weak point,' *todokōri* 滞り 'stagnation,' and *kakō* 下降 'decline' as a clue of that segmentation. And words with the negative meaning, whose superordinate concept is "*mondai* 問題 'problem' = undesirable state (judging from economy)," enter into temporary cohesive relationships within the text. This is not a lexicological relationship, however. Strictly speaking, it is not meant to refer to a later sentence or paragraph, but to imply the meaning of "undesirable state (judging from the economy)" in the relevant semantic segment. In other words, text-organization is shown as semantic segments based on the choice of the *meaning*, not form, of temporary cohesive relationships.

Such "signal words" are empirically known. Alternatively, it can be usage, not the words themselves. For instance, let us focus on the word *mondai* 問題 'problem' in the above column. The meanings of *mondai* 問題 'problem' that are described first in the dictionary are: "a question to find an answer, a question to require an answer and teaching, or a question" (*Nihon kokugo daijiten* 日本国語大辞典; English translation by Takasaki), and the meanings described second are: "criticism and a debate, or a matter

to be studied, a matter to be settled," "a matter to be kept in mind, notable point" (*Nihon kokugo daijiten* 日本国語大辞典; English translation by Takasaki). In the above column, the more abstract meanings of the word *mondai* 問題 'problem' contribute to the formation of context as text-organizing words. The word *mondai* 問題 'problem' cannot always be said to work as a text-organizing word. Text-organizing words are semantically chosen in a specific text.

In short, semantic segments are not formed based on the lexicological relations of words. Rather, cohesive relations are observed in a range segmented by text-organizing words. The text-organizing words could be superordinate words, and the subordinate words could also yield cohesive relations.

Actually, the relations among words within text can be freer and more creative, having a one-time-only nature and unexpectedness each time, in contrast with the more fixed relations of synonymity or coordinate, superordinate, and subordinate relations found in lexical semantics.

4 Summary: Works of lexical items in text

A word provides various meanings to text; from autonomous words (typically, proper nouns) in text to abstract words (typically, formal nouns, formal verbs, and formal adjectives, cf. Takasaki 1976) that cannot have autonomous meanings because their meanings are determined by the context. In the concept of 'polysemy' in lexicology, there is presupposition that the meaning of a word is not monolithic, but, rather determined by the context. Such dictionaries as *Kihongo jiten* 基本語辞典 'Dictionary of basic words' and *Ruigigo jiten* 類義語辞典 'Dictionary of Synonyms' have various examples of word usages from actual texts. As much as various examples are taken for their meanings to be explained, the meaning division becomes detailed and incomprehensible. The "central meaning" of a word will only be a reworded meaning of the word after all. There exists a rule for the order of the meanings in the Japanese dictionaries —primary meaning, secondary meaning, and so on. In my opinion, this order is intrinsically connected with the function of lexical items in the text.. This is based on my own experience of compiling a Japanese dictionary (*Sanseido gendai shin kokugo jiten* 三省堂現代新国語辞典, 4th edition).

Observing a real text, text-organizing words, as used in this paper, can be said to have occurred as a result of continuous usage in the following way: they intuitively choose appropriate components from the existing words within the constraints of the context, select a lexical meaning, or function in correspondence with text-organization based on a metaphorical idea.

In fact, the word *mondai* 問題 'problem' mentioned before was used in the secondary meaning. Also, the words *ten* 点 'point' and *shisei* 姿勢 'attitude' that were discussed

in Takasaki (2013) were used in the secondary meaning. However, the word *gen'in* 原因 'cause' was used in the primary meaning. In Japanese, the loan-word *apurōchi* アプローチ 'approach' (c.f. Takasaki 2012) has the following primary meanings: 'research the subject in a study, or its method, methodology; they are mainly used in a social science.' And the secondary meanings are: 'the path which leads to a specific place or building from the entrance or gateway to the site; ski jumping, running long jump, the high jump-, golf-'etc. It is considered that the abstract meanings of the word approach were brought into Japan earlier than concrete meanings. So the abstract meanings came first in the dictionary and concrete meanings came second. Secondary meanings do not always become text-organizing words.

A word, inflected and accompanied by an auxiliary word for reasons of the sentence structure, functions in a sentence structure. Likewise, the meaning of a word is put to practical use with various senses to contribute to constitution of context, or it is accompanied by modifiers to determine its sense.

Thus, behaviour of lexical words occurring in text is such that we realize the following point from Nomura (2003).

> The grammar, like a vocabulary item, is a "sign" of the conventional relation between form and meaning, and can be said to exist to express a meaning. The differences between a vocabulary item and the grammar only reside in the differences of degree of complexity of the form of the sign or degree of abstractness of the meaning of the sign. Vocabulary and the grammar are continuous and should not be divided in two as having totally different characters, as has been conventionally done (Nomura 2003:55; English translation by Takasaki).

In addition, another point that I want to pay attention to is the following:

> Text linguistics, being deeply related with the corpus linguistics, focuses on structuring lexical items by text-organizing functions (Ishii 2011:287; English translation by Takasaki).

Ishii (2011:287) states that the "text-organizing function" of a word means "reiteration" in Halliday and Hasan (1976) or the function of "discourse-organizing words" in McCarthy (1991), and continued that:

> Reiteration is shown in some strategies. Important lexical items such as synonymous words, superordinate and subordinate words is involved in such strategies. It is considered that words in such lexical relations are expressed with functioning of reiteration in text. A group of words in such lexical relations is considered to be prepared for reiteration or functioning of the text-organization ... (Ishii 2011:287; English translation by Takasaki).

Furthermore, Ishii (2011) cited the following statement in McCarthy (1991: 67)[7]:

> … synonyms are not just ways of understanding new words when they crop up in class, nor are they some abstract notion for the organisation of lexicons and thesauri, but they are there to be used, just as any other linguistic device, in the creation of natural discourse.

Ishii (2011) goes on to say, "Here is an answer from text linguistics to the question of why vocabulary is shaped and structured like that (English translation by Takasaki)."

Therefore, taking these statements as our point of departure, we in the field of text linguistics can consider that text causes a word to have the power of organizing the text itself by continuous creation and characterization of a meaning of the word while giving function at the same time.

Concerning lexical cohesion, it was made clear that words support textuality by being repeated in text (Takasaki 1986, 1990b, 2007, etc); some words are coherent with having lexical relations, temporary relations, and relationships based on the world knowledge; and they form semantic segments from small to large. In other words, the text-organizing function of vocabulary does not simply mean that a word as text-organizing word works with combination of segments, but that a word's cohesion via reiteration (such as a tautology and rewording by lexically superordinate words, subordinate words, synonymous words, or words with the same meaning) organizes the whole text or segments that organize text. Of course, there can be not only lexical relations, but also temporary relations of cohesion limited to the specific text.

Such phenomena can be used as a standard for making segments. Moreover, in a long text such as an introductory academic textbook, it can be observed that some technical terms both appear repeatedly in the text as a whole and are reiterated as well. For example, the word *seitō* 政党 'a political party' is used 337 times over the course of the text without any sense of disproportion in *Seiji-gaku Nyūmon* 政治学入門 'Introduction to Political Science.' The word *reisen* 冷戦 'cold war' is used 193 times in *Nippon Gaikō-shi Kōgi* 日本外交史講義 'Lecture on the History of Diplomacy in Japan,' the word *shijō* 市場 'market' is used 206 times in *Amerika no Keizai* アメリカの経済 'Economy of America,' and the word *keihō* 刑法 'criminal law' is used 588 times in *Keihō Genron* 刑法原論 'Basic Principles of Criminal Law.' In addition, non-technical terms such as *mondai* 問題 'problem,' *gensoku* 原則 'principle,' *keikō* 傾向 'tendency,' *jōkyō* 状況 'situation,' and *henka* 変化 'change' are frequently used as text-organizing words, with specific senses each time, and sometimes form a long chain of cohesion by repetition of the same word in the whole text as a result.

---

7 Ishii (2011) quoted McCarthy from Andō and Katō's 1995 Japanese translation (see bibliography); however, the English from McCarthy's 1991 original is instead supplied here for the reader's convenience.

It is observed that text develops with words that are not particularly abstract, having relations and being combined with segments to become text-organizing words. It can be said that lexical cohesion itself is deeply connected with text-organizing function.

5 Conclusion

In this paper some of the functions of vocabulary in a sentence are observed. It is considered that a meaning of the word is grammatically restricted and determined in text; it is ambiguous between a lexical autonomous meaning and the contextual meaning that received contextual interference; and it comes to have text-organizing function by itself. We can even see concrete words, such as *shisei* 姿勢 'posture,' *ugoki* 動き 'motion,' or *chōryū* 潮流 'trend,' combine with big segments in editorial column. We can also see an aspect that typical common nouns, such as *jōkyō* 情況 'situation,' *henka* 変化 'change,' and *gensoku* 原則 'principle,' whose degrees of abstraction are relatively high compared with more concrete nouns such as *ringo* りんご 'apple' and *sora* 空 'sky,' are frequently used as text-organizing words necessary for text development in introductory science textbooks.

Furthermore, many Sino-Japanese words represent text-organizing words. It has been said that only native Japanese words can serve as postpositional particles, auxiliary verbs, adverbs, conjunctions, and interjections, etc. which have strong functional aspects in a sentence structure. However, some Sino-Japanese words seem to tend towards forming a group of functional words that bring out function rather than meaning. It can also be said that Japanese writers are rapidly making fuller use of Sino-Japanese words.

Textuality makes a text an entity with a meaning, not simply the set of its constituent words, and indicates an aspect that a word from a vocabulary system is rearranged so that an intention can be conveyed. Therefore, the text is a field where a word exhibits its functional aspect. And the word functions so that the meaning of text is exactly conveyed. What bears textuality in a text is not any single feature of the text on its own; various cohesive relations and organizational clues are prepared and working together, indeed realizing each other, within the text. The organization of text is more complex than mere sentence structure. The units of various scales are combined and incorporated like a nest of boxes to effect realization of the meaning of the text for the purpose of conveying it as much as possible to the reader.

The text has a large quantity of language, which is unidirectional, linear, and timewise. Such characteristics are quite troublesome. However, there exists function for concisely grasping the large quantity of language inside the text. This function has bidirectional, planar, and consequential characteristics, and text-organizing function and cohesiveness perform such function.

Thus, any language form including words has functions and characteristics that are particularly brought out in text. With the corpora being steadily improved, the actual state of language forms in real text will be easily confirmed. Japanese linguistics is trying to confirm what has traditionally been said by using corpora. We want to make further observations of such behaviour in the whole text in the fields of lexicology, grammar, and orthography. That is to say, we want to continue pursuing the methodology of "Japanese text linguistics."

Acknowledgment

I am grateful to Prof. Andrej Bekeš who suggested me to contribute to this book, and made valuable suggestions.

Also, I would like to thank Irena Srdanović for numerous valuable suggestions and comments at the drafting stage of this chapter.

I thank to Mitsuko Takahashi and Laura E. Johnson for their help in English translation. I am, of course, entirely responsible for any faults that remain.

Literature

Halliday, M. A. K. and Hasan, R. (1976) *Cohesion in English.* London and New York: Longman. (Trans. by Andō, S. et.al. (1997) *Tekusuto wa Dono yō ni Kōsei sa reru ka : Gengo no Kessokusei*. Tokyo: Hituzi Syobo.)

Halliday, M. A. K. and Hasan, R. (1989) *Language, Context, and Text: Aspects of Language in a Social-Semiotic Perspective.* Oxford: Oxford University Press. (Trans. by Kakei, H. (1991) *Kinō Bunpō no Susume*. Tokyo: Taishūkan Shoten.)

Ishii, M. (2011) *Rinsetsu Shobun'ya no Goi Kenkyū to 'Korekara no Goiron.'* In: Saito, M. and Ishii, M. (eds.) *Korekara no Goiron.* 275-291, Tokyo: Hituzi Syobo.

Kanaoka, T. (1963) *Shudai to Kōsei.* In: Morioka, K. et al (eds.) *Kōza Gendaigo Dai 3 kan Dokkai to Kanshō.* 36-56, Tokyo: Meiji Shoin.

Kawakami, S. (ed.) (1996) *Ninchi Gengogaku no Kiso* (*An Introduction to Cognitive Linguistics*). Tokyo: Kenkyū-sha.

Kim, E. (2012) *Gairaigo no Kihongo-ka.* In: Jin'nai, M., et al. (eds.) *Gairaigo Kenkyū no Shin- tenkai.* 29-45, Tokyo: Ohfu.

McCarthy, M. (1991) *Discourse Analysis for Language Teachers.* Cambridge: Cambridge University Press. (Trans by Andō, S. and Katō, K. *Gogaku Kyōshi no Tame no Danwa Bunseki.* (1995) Tokyo: Taishūkan Shoten.)

Nomura, M. (2003) *Ninchi Gengogaku no Shiteki, Rironteki Haikei.* In: Tsuji, Y. (ed.) *Ninchi Gengogaku he no Shōtai.* 17-62, Tokyo: Taishūkan Shoten.

Saito, M. (2011) *Nihongogaku, Gengogaku no Sho-bun'ya to Korekara no Goiron.* In: Saitō, M. and Ishii, M. (eds.) *Korekara no Goiron.* 255-274, Tokyo: Hituzi Syobo.

Takasaki, M. (1976) *Keiyōshi, Keiyōdōshi no Kijutsuteki Kenkyū: Ogai no Roku Sakuhin o Taishō to shite. Kokubun* 45:73-84. Ochanomizu University.

Takasaki, M. (1985) *Bunshō ni okeru Hanpuku Gokuteki Oyobi Kanren Goku no Kinō ni tsuite. Bunkyō Kokubun* 14:26-41. Bunkyō University.

Takasaki, M. (1986) *Bunshō no Gokuteki Kōzō. Kokubun* 64:47-57. Ochanomizu University

Takasaki, M. (1988) *Bunshō Tenkai ni okeru "Shiji Goku" no Kinō. Kokubungaku Gengo to Bungei* 103:67-88.

Takasaki, M. (1990a) *Shiji Hyōgen.* In: Teramura, H. et al. (eds.) *Kēsu Sutadi Nihongo no Bunshō: Danwa.* 34-45, Tokyo: Ohfu.

Takasaki, M. (1990b) *Hanpuku to Shōryaku no Hyōgen.* In: Teramura, H. et al (eds.) *Kēsu Sutadi Nihongo no Bunshō: Danwa.* 46-57, Tokyo: Ohfu

Takasaki, M. et al. (eds.) (2007) *Nihongo Zuihitsu Tekusuto no Shosō*. Tokyo: Hituzi Syobo.

Takasaki, M. and Tachikawa, K. (eds.) (2008) *Koko kara Hajimaru Bunshō: Danwa*. Tokyo: Hituzi Syobo.

Takasaki, M. and Tachikawa, K. (eds.) (2010) *Gaidobukku: Bunshō, Danwa*. Tokyo: Hituzi Syobo.

Takasaki, M. (2011) *Bunshōron, Buntairon to Goi.* In: Saito, M. and Ishii, M. (eds.) *Korekara no Goi-ron.* 113-124, Tokyo: Hituzi Syobo.

Takasaki, M. (2012) *Tekusuto no Kessokusei ni Azukaru Goi to Sono Kinō ni tsuite. Dai 2 kai Kōpasu Nihongogaku Wākushoppu Yokōshū.* 7-14, Tokyo: National Institute for Japanese Language and Linguistics.

Takasaki, M. (2013) *Bunshōchū no Goi no Kinō ni tsuite: "Tekusuto Kōsei Kinō" to iu Kanten kara.* In: Yamazaki, M. (Project leader) *Tekisuto ni okeru Goi no Bunpu to Bunshō Kōzō Seika Hōkoku-sho* 12-6:41-66. *A Report of Collaborative Investigation*, Tokyo: National Institute for Japanese Language and Linguistics.

Tanaka, S. and Fukaya, M. (1998) *Imidzukeron no Tenkai: Jōkyō Hensei, Kotoba, Kaiwa.* Tokyo: Kinokuniya Shoten.

Dictionaries:

Bunrui goi-hyō zōho kaitei-ban. (2004) National Institute for Japanese Language and Linguistics. Tokyo: Dainippon Tosho.

Gendai shin-kokugo jiten, 4th edition. (2011) Tokyo: Sanseido.

Nihon kokugo daijiten, 2nd edition. (2001) Tokyo: Shogakukan.

要旨（Abstract in Japanese）

「語彙的結束性とテクスト構成の機能－－文章論からの提言－－」

髙﨑みどり（お茶の水女子大学）

「テクスト構成語」という概念と「結束性」という概念を使って、日本語テクストではどのようにそれらが現れるのか、いくつかのテクストで観察してみた結果を報告する。これらの概念はテクストをテクストとして成立させている「テクスト性」に関するいくつかの概念の一部である。テクスト構成語はテクストの流れを区切り、それによってテクスト（あるいはテクストの一部）を構造化する機能を有する。結束性は言語形式同士が関係しあうことにより、テクスト（あるいはテクストの一部）に意味的な一貫性をもたらす。両者の関係は、端的に言えば、テクスト構成の機能は結束性によって実現するということになろう。また、ここではテクストを意味的に構成する一種の作業的単位として“意味分節”という概念を設けることとした。

3 Tracking references to unfamiliar food in Japanese Taster Lunches: Negotiating agreement while adapting language to food

Polly SZATROWSKI
University of Minnesota

Abstract
In this paper I investigate how Japanese participants track references to unfamiliar food in the interaction at Taster Lunches. The analysis investigates (1) What aspects of the food do participants use as resources to create references for unfamiliar food?, (2) What patterns in reference tracking can be observed through the conversation?, (3) How do participants' choices of similar or different referring expressions influence their assessment and categorization of the food and their relationships with one other?

Patterns in the use of nouns/noun phrases to refer to unfamiliar food showed that participants tended to use demonstrative pronouns initially, and subsequently used more specific references including features of color, shape, texture, flavor, and combinations. This reflects participants' multi-sensory experience of food. While the referring expressions for more familiar food were settled referring expressions, references for less familiar foods were monitored and modified throughout the discussion of the food item. Choice of referring expression also influenced participants' assessment and categorization of the food. Participants' repetition of expressions that other participants used to describe, assess and categorize the food in subsequent non-predicate (Minami 1974, 1993, 1997) referring expressions suggested their agreement on food descriptions and categorization, and contributed to the stability of the referring expression.

Results indicate ways in which participants adapt language to unfamiliar food in the process of negotiating food references based on their multi-sensory experience, knowledge, assessment and categorization of the food in the talk-in-interaction. This study also contributes to research on contextualized social and cognitive activity, language and food, and cross-cultural understanding.

Keywords: referring expressions, unfamiliar food, demonstrative, agreement, sensory experience

1 Introduction

In this paper I investigate how Japanese participants track references to unfamiliar food in the interaction at Taster Lunches (Szatrowski 2011, 2013, 2014a,b,c,d, 2015a,b,

2017). When encountering unfamiliar food participants are faced with the need to adapt language to refer to the food while describing, assessing and categorizing the food. The data for this study come from videotaped conversations of 13 Japanese triads and 10 American English triads, each eating and commenting on three courses containing three to four foods from Japan, America, and Senegal, respectively.

In my analysis I investigate the following questions: (1) What aspects of the food do the Taster Lunch participants use as resources to create references for unfamiliar food?, (2) What patterns in reference tracking can be observed through the conversation?, (3) How do participants' choices of similar or different referring expressions influence their assessment and categorization of the food and their relationships with one other?

In this study I use the terms "referring expression" to refer to a candidate name or category of a food or drink in the Taster Lunch, and include in my analysis *jutsubuteki na yōso* 'predicate elements' (verbal, adjectival, nominal + copula predicates) as well as *jutsubuteki na yōso igai no seibun* 'non-predicate components' (N+*wa* (TOPIC), N+*ga* (SUBJECT), and N+other case particles, e.g., N+*ni* (INDIRECT OBJECT), N+*o* (DIRECT OBJECT), etc.) (Minami 1974, 1993, 1997).[1] My inclusion of predicate elements and consideration of the referring expressions used by a multiple of participants differs from previous research on referring expressions that tended to focus on non-predicate components used by a single speaker, the narrator, in narratives about a film, animation, etc. (Clancy 1980, Watanabe 2009, 2010, and others).

2 Previous research

2.1 Research on knowledge and language in conversation

Research on knowledge and language in conversation has focused on the relative epistemic states of the participants. Labov & Fanshel (1977) distinguished between A-events (known to A, but not to B) and B events (known to B, but not to A). Kamio (1994) demonstrated how Japanese speakers use modal forms and final particles to distinguish between knowledge in the speaker's territory of information, knowledge in the hearer's territory of information, and knowledge in both the speaker's and hearer's territory of information to the same or varying degrees. Heritage (2012:4) proposed the notion of "epistemic status" to refer to "relative epistemic access to a domain or territory of information as stratified between interactants such that they occupy different positions

1 See Szatrowski (2007) for an English summary of Minami's hierachical model of Japanese sentence structure. Regarding predicate elements, I consider nouns used in nominal +copula predicates, and descriptions using verbs (e.g., *bunibuni-shite iru* 'is jellylike'), adjectives (*amai* 'sweet') that may be used later in non-predicate references. I also include N+Z (noun+zero particle) as a non-predicate component.

on an epistemic gradient (more knowledgeable [K+] or less knowledgeable [K-]), which itself may vary in slope from shallow to deep". Koike (2014:172) defined a "'knowing participant' as a participant who has more access to the information in question at a given moment in interaction, and an 'unknowing participant' as a participant who has no or less access to the information in question vis-à-vis the knowing participant." She demonstrated how Japanese knowing and unknowing participants achieved mutual understanding of a food X that the knowing participant had eaten in the past by using food categories, comparing similarities (simile) and contrasting differences between food X and known foods, and creating new food categories. My research contributes to this research by focusing on food and drink unknown to all three participants at the Taster Lunch. In particular, I investigate how they construct references to the food through negotiations of agreement on possible categories based on their sensory experiences during the Taster Lunch.

2.2 Research on the relation between language and food in Japanese conversation

Previous research on the relation between language and food in Japanese conversation focused on verbal and non-verbal assessments in television cooking shows (Szatrowski 2009), verbal and nonverbal behavior at Taster Lunches between three women under 30 (Szatrowski 2011), the use of modal and evidential forms in talk-in-interaction in Taster Lunches among Japanese native speakers, among American English speakers, and among native and non-native Japanese speakers (Szatrowski 2014a, 2014d, 2015b), the use of so-called "subjective" and "objective" expressions for food assessment (Szatrowski 2013), the relation between food and family at Taster Lunches (Szatrowski 2014c), the use of onomatopoeia in Japanese Taster Lunches (Szatrowski 2015a, 2018), and identification of unfamiliar food (Szatrowski 2016). There has also been research on the use of "pragmemic triggers" and formal expressions to delineate the stages in the process of commensality from the beginning (invitation) to the end of a meal (Beeman 2014), the structural organization of ordering and serving sushi (Kuroshima 2014), food description in Japanese at a pot luck party (Noda 2014), repetition of the punchline of stories about food and restaurants (Karatsu 2014), and the socialization of Japanese children to food-related practices (Burdelski 2014).

3. Analysis

3.1 Data and methodology

The data for this study come from videotaped Taster Lunch conversations of 13 Japanese triads and 10 American English triads, each eating and commenting on three

courses containing three to four foods from Japan, America, and Senegal, respectively.[2] The triads consisted of three friends in varying gender (FFF, FFM, FMM, MMM) and age (<30 years, >30 years) combinations. In this study, I will analyze a Japanese female triad under 30 (JPN3=FFF). In particular, I will focus on the conversational segments in which the participants are eating a dessert in a bowl called *LAAX* '(white corn) flour pudding with a sweet (yogurt and) milk sauce' in the Senegalese course (the corn flour pudding forms a lump in the middle with the sauce on top). The Senegalese course also included *MAFE* '(chicken in) peanut butter sauce' on Jasmine rice, and *BAFIRA* 'hibiscus juice'.

My methodology was as follows. First, I identified the explicit references used for the *LAAX* and associated them with four perspectives, specifically, visual appearance (shape, quality/substance, color), taste, texture, and smell. Next, I distinguished the predicate elements and non-predicate components (Minami 1974, 1993, 1997). Finally, I analyzed the effect of different referring expressions on the identification and assessment of the unknown food (*LAAX*), and the relation among the participants at the Taster Lunch.

3.1 Referring expressions used in JPN3

In this section I will investigate the patterns in the use of referring expressions for the *LAAX* by three women under 30, Gin (g), Haru (h), and Iku (i) as viewed from left to right on the video. While the references for more familiar food were settled quickly, references for less familiar foods continued to be monitored and modified throughout the discussion of the food item.

As seen in Excerpt 1, there was no negotiation of the reference for *hijiki* 'black seaweed' in the Japanese course.[3] Based on sight, Haru is the first to comment on the *hijiki*. She refers to it as *hijiki* using a non-predicate N+*ni* on first mention in 105h, and Gin refers to *hijiki* in general using a non-predicate N+*tte* (quotative particle) in 107g. They also use non-predicate references for the *kozakana* 'small fish' in 106h (N+*ga*) and 107g (N+Z), respectively. About one minute later Haru and Iku taste the *hijiki*, and Iku says it is delicious in 136i, referring to it as *hijiki* with a non-predicate N+Z, and Haru ellipts the reference when she agrees it is delicious in 137h. In this way participants tended to refer to familiar foods by their name in non-predicate components.

2 See Szatrowski (2014b:27-28) for a description and pictures of the Taster Lunch meal.

3 See the Appendix for the transcription conventions used in this paper. In the data, I put a box around non-predicate referring expressions, underlined categories used in nominal predicates, and put a dotted underline under descriptive elements used in predicates that are later used in non-predicate components.

Excerpt 1:
JAPANESE COURSE-HIJIKI1 'BLACK SEAWEED' 4:42-4:50[4] (Familiar food)

105h 個人的にはひじきにさあ、
Kojinteki ni wa hijiki ni sā,
Personally, in *hijiki*, you know,
106h この小魚が、//混じっているのが、|| ※h,i: bend forward to look at *HIJIKI*
kono kozakana ga, //majitte iru no ga,||
having these small fish mixed in,
107g //ひじきって小魚入ったんだね。||
//Hijiki tte kozakana haitta n da ne.||
//It's that *hijiki* had small fish in it, huh. ||
※g: bends forward to look at *HIJIKI*
108h //ちょっと不思議。||
//chotto fushigi.||
//is a little strange.||
109g //初めて見た。||
//Hajimete mita.||
//(It's) the first time (I)'ve seen (it).||
110h うん、
Un,
Yeah,
111h 私//も初めて見た。|| ※hangs her hair on the right side on her right ear with left hand
Watashi //mo hajimete mita.||
(For) me //too (it's) the first time (I)'ve seen (it).||
…((4:51-5:44 g,h,i talk about the *UDON* 'noodle' broth, and *UDON* 'noodles'))

JAPANESE COURSE-HIJIKI2 'BLACK SEAWEED' 5:45-5:59 (Familiar food)

(2.1) g is eating the *UDON* and h,i are eating the *HIJIKI*
136i ひじきおいしいよ。
Hijiki oishii yo.
(The) *hijiki* is delicious I tell you.
137h あ、おいしい。
Oh, (it)'s delicious.

4 HIJIKI1 means the first section where participants talk about the *hijiki*, and 4:42-4:50 indicates the beginning and end of the excerpt in the video (minutes:seconds).

In contrast, there was more negotiation of the referring expressions used for the *LAAX*. In initial references to unfamiliar food, demonstrative pronouns were the most common non-predicate components used and possible references were given in predicate elements used to negotiate the identity of and categorize the food. Excerpt 2 begins with Haru pointing at the *LAAX* and asking what it is with the demonstrative pronoun *kore* 'this'. In response, based on sight, Iku says she does not know in 454i, and Gin questions whether the *LAAX* is fish in 455g. Then Gin and Iku smell the *LAAX* and simultaneously conclude that it is yogurt. Next, Gin's suggestion that it is fish and yogurt in 458g is met by responses in 460h, 461i, and 462g that question whether it is fish. The excerpt ends with Gin and Iku agreeing that it is yogurt in 463g-465i, and Haru questioning again whether it is fish in 466h while doing a head tilt twist.[5] All the uses of *sakana* 'fish' and *yōguruto* 'yogurt' are predicate elements, and the only non-predicate component used is the demonstrative pronoun *kore* 'this' (N+Z), postposed in 461i and utterance initial in 466h. This was typical of initial references to unfamiliar food; demonstrative pronouns were the most common non-predicate components used initially.

Excerpt 2:
SENEGALESE COURSE-LAAX1 (15:02-15:17) (Unfamiliar food)

453h ˚あと˚これ何だと思う？ ※points at *LAAX* with right index finger
˚Ato˚ kore nan da to omou?
˚Also˚ what do you think this is?
454i わかんない。 ※lifts *LAAX* bowl up to chest height
Wakannai.
(I) don't know
455g 魚？
Sakana?
(Is this) fish?
(2.0) ((g,i:smell the LAAX))
456g //ヨーグルト。||
//Yōguruto.||
//(It's) yogurt.||
457i //なんかヨーグ||ルトだ。
//Nanka yōgu||ruto da.
//Somehow (it's) yogu||rt.
※h:picks up the bowl with both hands and smells the LAAX

5 Szatrowski (2014a:141) defines a "head tilt twist" as tilting one's head to one side while twisting the head in the opposite direction.

458g 魚とヨーグルト？
Sakana to yōguruto?
(Is this) fish and yogurt?
459i //うん。||
//*Un.*||
//Uh.||
460h //さ、||えーーー？
//*Sa,*|| *e*:::?
//Fi, || wha:: :t?
461i～ 魚なの？°これ。°
Sakana na no? °*Kore*.°
Is it that (this) is fish? °This.°
462g 魚じゃないかな、
Sakana ja nai ka na,
I wonder if (it) isn't fish (afterall),
463g なんかヨーグルトだよね。
nanka yōguruto da yo ne.
somehow (it's) yogurt, I tell you, isn't it.
464i うん、
Un–,
Yeah–,
465i ヨーグルト。
Yōguruto.
(it's) yogurt.
466h これ魚、<えー？> ※head tilt R twist L
Kore sakana, <*e*:?>
Is this fish, <wha:t?>

About three minutes later after the participants discuss the *MAFE* with Gin concluding that it contains chicken, Gin initiates another discussion of the *LAAX* in Excerpt 3. Gin begins by referring to the *LAAX* with a non-predicate noun phrase ending in *wa* (the topic particle) in 575g[6] and (drawing from her conclusion that the *MAFE* contains chicken) wonders whether the fact is that the *LAAX* is not meat. Haru denies totally that the *LAAX* has meat in 576h using a postposed demonstrative pronoun *Kore:* 'thi:s' with a final sound stretch.[7] Gin's use of *yōguruto* 'yogurt' in her non-predicate reference in

6 It was common to use a non-predicate nominal reference to refer to the *LAAX* after a shift (Clancy 1980) from talking about another food.

7 Haru's pronunciation of the predicate and postposed *kore*: 'thi:s' in 576h in one intonation unit, with a sound stretch on *kore*: and loud voice over the entire utterance contribute to her strong denial, and exemplify what Ono & Suzuki (1992) refer to as the emotive type of postposing.

575g *kono yōgurutoppoi no wa* 'as for this yogurt-ish one' reflects Gin and Iku's agreement that the *LAAX* contains yogurt in Excerpt 2, although her use of *-ppoi* 'ish' on the end of *yōguruto* 'yogurt' makes it less determinate. Subsequently Gin tries to rationalize why the *LAAX* is not meat by suggesting in 578g that the combination of meat (in the *MAFE*) and fish (in the *LAAX*) would be heavy,[8] pointing at the *MAFE* when she says meat and *LAAX* when she says fish. Then, in 579g Gin questions whether the *LAAX* is a vegetable. Next after Iku and Haru try the *LAAX*, they indicate that they cannot identify it using inexplicit reference, and in 591i–592i Iku accounts this to the fact that the flavor is disguised by the yogurt using a non-predicate N+*ni* component in 592i *yooguruto ni* 'by the yogurt'. Like Gin did in 575g, Iku uses a non-predicate component to refer to the *LAAX* after she and Gin agreed that the *LAAX* had yogurt in it in Excerpt 2. Haru's utterance in 594h suggests that she also agrees that the *LAAX* contains yogurt.

Excerpt 3:
SENEGALESE COURSE-LAAX2 (18:30-19:07) (Unfamiliar food)

575g (1.9)てことはこのヨーグルトっぽいのは肉じゃないってことかな。
(1.9) *Te koto wa kono yōgurutoppoi no wa niku ja nai tte koto ka na.*
(1.9) (From that) I wonder if the fact is that this yogurt-ish one is not meat.

576h～ ・@肉ではないでしょこれー。@・
・*@Niku de wa nai desho kore:.@*・
・@(It)'s probably not meat thi:s one.@・
※Haru bends forward and looks into her bowl with her left hand on side of bowl.

577i //いけいけー。||
//Ike ike:.||
//Go go:.|| ((Iku encourages everyone to eat the *LAAX*.))
※picks up the bowl of *LAAX* with her right hand and spoon with her left hand

578g //さか-魚と||か肉系ー、//じゃない［gてことじゃない？そしたら||さー、
//Saka- sakana to|| ka nikukee:., //ja nai [g te koto ja nai? soshitara|| sa:.,
//Fis- fish or|| meat grou:p, //(it) is not, [gisn't that the case? then|| you kno:w,
[g: raises her left hand from the *LAAX* bowl on the table to upper chest height and in 580g points down twice with her left index finger, first at the *MAFE* in front of her when she says *Niku* 'meat' and second at the *LAAX* a little forward

8 It is interesting to note that Gin uses a non-predicate component in 580g *Niku sakana tte* 'lit. speaking of meat (and) fish' although the participants have not agreed previously that the *LAAX* contains fish. However, this is not a counterexample to the tendency I observed for participants to use non-predicate components after some agreement is reached, because here Gin is speaking about meat and fish in general and giving a reason for the *LAAX* not being fish, rather than referring to the *LAAX* in particular as fish.

from that when she says *sakana* 'fish', associating the *MAFE* with meat and the *LAAX* with fish.]

579i〜 //んーじゃいこうよ。これ。||
//*N*: *ja ikō yo. kore*.||
//Yea:h then let's go I tell you. this ((=*LAAX*)).||

580g 肉魚って濃い$_{g}$］じゃん↑。
Niku sakana tte koi$_{g}$] *jan*↑.
Meat (and) fish (would be) heavy,$_{g}$] (would)n't they↑.

581g （なんか）、//野菜？||
(*Nanka*), //*yasai*?||
(Somehow), //a vegetable?||

582i //何だろ。||
//*Nan daro.*||
//What might (it) be.||

(2.2) ((g: eats *MAFE*; h: drinks *BAFIRA*; i: holding the bowl of *LAAX* in her left hand, puts a spoonful of *LAAX* in her mouth with her right))

583h (2.2)じゃあたしもちょっとこれ一口いってみよう。
(2.2) *Ja atashi mo chotto kore hito-kuchi itte miyō.*
(2.2) Then I too will just try going (with) one bite (of) this ((=*LAAX*)).

584h→i どう？
Dō?
How is (it)?

585i〜 うん、何だろこれ。
Un, *nan daro kore*.
Yeah, what might (it) be this ((=*LAAX*)).

586h ん？
N?
Hm?

587h (1.2)何かわからない。
(1.2) *Nani ka wakaranai.*
(1.2) (I) can't tell what (it) is.

588i うん。
Un.
Yeah.

589i これといった特徴的な味でもない。
Kore to itta tokuchōteki na aji de mo nai.
(It) doesn't have a distinctive flavor (of the sort) that (one could) say (it) is this.

590g (1.4)ああ、そろそろお腹いっぱいに//なってきた。||
(1.4) *Aa, sorosoro onaka ip-pai ni //natte kita.*||
(1.4) O:h, gradually (my) stomach //has come to get|| full.
591i //なんか、ヨー||グルトに=
//*Nanka,* *yō*||*guruto ni*=
//Somehow, || with yoghurt=
592i すべてかき消され//てる@気が||する。@
subete kakikesare //te ru @ki ga || suru.@
everything has been eras//ed @ (I)|| feel.@
593g //{アハハハ}||
//{*a ha ha ha*}||
//{LAUGHTER}||
594h あそれは言えてる//かも。||
A sore wa iete ru //ka mo.||
Oh (you) can say that //maybe.||
595i //{フフッ}||
//{*hu hut*}||
//{LAUGHTER}||

Excerpt 4 is a continuation of Excerpt 3. It begins with Gin eating the *LAAX* for the first time, and Haru referring to the *LAAX* with the non-predicate component (N+*ga*) in 596h and 598h *Kono sā, shi, kono shiroi*:, *katamari ga*:, 'This, you know, whi-, this whi:te lu:mp' and adding that she cannot tell what it is. Unlike the other uses so far of non-predicate references that had at least two people's agreement before using it as a non-predicate component, Haru uses a non-predicate component that refers to the color and shape of the *LAAX* without previous agreement. This suggests that color and shape may be characteristics which do not require agreement, that is, even though the participants do not know what the *LAAX* is, they may assume that knowledge of its shape and color is shared because it comes from visual evidence, and therefore these characteristics do not require agreement. Subsequently, Gin uses a non-predicate demonstrative pronoun (N+Z) *kore* 'this' in 600g to question whether the *LAAX* is *uri* 'gourd'. Haru disagrees in 602h-603h and adds that the non-predicate demonstrative pronoun in 605h *kore jitai wa* 'this itself' (N+*wa*) does not have much flavor. After Gin indicates there are raisins in the *LAAX* in 606g using a non-predicate N+Z for this familiar referent (without previous agreement), Haru clarifies the referent of the demonstrative pronoun (that she used in 60th) in 607h *kono, shi, shiroi katamari no hō* 'the alternative of this, whi, the white lump (as opposed to the sauce)'.

Excerpt 4:
SENEGALESE COURSE-LAAX2 (cont.) (19:08-19:40) (Unfamiliar food)

596h このさあ、
Kono sā,
This, you know,
※g:picks up the bowl with her left hand and eats *LAAX* with spoon in her right hand

597i うん。
Un.
Uh huh.

598h し、この白いー、塊がー、 ※h lifts up some *LAAX* with a spoon in her right hand
shi, kono shiroi:, katamari ga:,
whi, this whi:te lu:mp,

599h (1.5)よくわからん。
(1.5) *yoku wakaran.*
(1.5) (I) can't tell well.

600g 何これ。瓜？何だろ。
Nani kore. Uri? Nan daro.
What is (it) this. Gourd? What might (it) be.

601i {フフ}
{*hu hu*}
{LAUGHTER}

602h °瓜ではない。° .hh ※h shakes her head from left to right twice.
Uri de wa nai.
(It's) not a gourd.

603h (1.2)と思うんだけど。
(1.2) *to omou n da kedo.*
(1.2) it's that (I) think (that) but.

604h (2.9)これ自体は味あんまりないのかなあ。
(2.9) *Kore jitai wa aji anmari nai no ka nā.*
(2.9) (I) wonder if it's that this itself doesn't have much flavor.

605i うん。
Un.
Yeah.

606g レーズン入ってるよ。
Rēzun haitte ru yo.
Raisons are in (it), you know.

607h (1.0)あ、うんあのこの、し、白い塊の方さ。
(1.0) *A, un ano kono, shi, shiroi katamari no hō sa.*
(1.0) Oh, yeah uhm the alternative of this, whi, white lump, you know.
… ((19:27-19:40 Discuss whether the *LAAX* is bread soaked in yogurt.))

In Excerpt 5, a continuation of Excerpt 4, the participants begin to evaluate the *LAAX* and talk about its taste and texture. In 612i Iku evaluates the *LAAX* positively saying she likes it using the non-predicate demonstrative pronoun *kore* 'this' (N+Z). However, Haru and Gin indicate otherwise, commenting that it is (too) sweet, both using non-predicate components (N+*ga*) to refer to the part of the *LAAX* that they find sweet. Specifically Haru repeats her previous reference to color and shape (598h) in 615h *kono shiroi katamari ga* 'this white lump' and Gin uses the mutually agreed upon *yooguruto ga* 'the yogurt' in 616g and 618g.

Excerpt 5:
SENEGALESE COURSE-LAAX2 (cont.) (19:41-20:09) (Unfamiliar food)

612i //でもあたしこれ||全然好きだわ。
//*Demo atashi kore*|| *zenzen suki da wa.*
//But I this|| totally like (it), you know.
613h (2.1)甘い。
(2.1) *Amai.*
(2.1) (It's) sweet.
614g あたしこれちょっと微@妙。@{h}
Atashi kore chotto bi@myō.@ {h}
(For) me this is a bit ques@tionable@.{h}
615h (1.6)°あそっか°、この白い塊が甘いのか°な。°
(1.6) °*A sokka*°, *kono shiroi katamari ga amai no ka* °*na.*°
(1.6) °Oh right°, (I) wonder if it's that this white lump is sweet.
616g いや、ヨーグルトが甘いんだと思うよ。
Iya, yooguruto ga amai n da to omou yo.
Nah, (I) think it's that (the) yoghurt is sweet, I tell you.
617i うん。
Un.
Yeah.
618g (1.9)ヨーグルトがめちゃくちゃ甘い↑。
(1.9) *Yōguruto ga mechamecha amai* ↑.
(1.9) (The) yoghurt is excessively sweet ↑.
619i うん。
Un.
Yeah.

620h 全部甘い<な>。
Zenbu amai <na>.
(It's) all sweet, <isn't it>.
621g {ンフフフフ}
{*N hu hu hu hu*}
{LAUGHTER}
… ((20:01-20:09 Discussion about the *LAAX* not being soggy bread))

In Excerpt 6, a continuation of Excerpt 5, the participants continue to try to identify the *LAAX* and evaluate its texture. Haru asks what it might be, referring to it with non-predicate N+Z components in 626h *kono*, *buttai* 'this, object' and 628h *Buttai X* 'Object X' again using a reference to shape. Then Gin indicates that she finds the texture disgusting describing it with the predicate elements in 630g *bunibuni* 'jellylike' and 631g *zarazara* 'grainy', but Haru and Iku disagree by saying that they like the texture in 634h, 637i and 638h.

Excerpt 6:
SENEGALESE COURSE-LAAX2 (cont.) (20:09-20:27) (Unfamiliar food)

626h～ 何だろこの、物体。
Nan daro kono, buttai.
What might (it) be this, object.
627g うん。
Un.
Yeah.
628h 物体エックス。
Buttai ekkusu.
Object X.
629i {フフ}
{*hu hu*}
{LAUGHTER}
630g (2.0)なんかちょっとぶにぶにしてるのに＝
(2.0) *Nanka chotto bunibuni-shite ru no ni=*
(2.0) Like although (it's) a bit jellylike=
631g ざらざらしててちょっと気@持ち悪い。@
zarazara-shite te chotto ki@mochi warui. @
(it's) grainy and (so) a bit dis@gusting. @
632h でも、//@ごめんー、@||
Demo, //@gomen:,@||
But, //@so:rry:,@||

633g　　//舌触り気持ち||悪くない？
　　//Shitazawari kimochi||waruku nai?
　　//The feeling on your tongue is gro||ss, isn't it?
634h　え？//この||食感@面白くて好きなん//だけど。@||
　　E? //kono|| shokkan @omoshirokute suki na n //da kedo.@||
　　What? //this|| texture @it's that (it) is interesting and (I) like (it) //but.@||
　　※points at *LAAX* with left index finger
635i　　//ええ？||
　　//Ee?||
　　//What?||
636g　　//んーー？||
　　//N::?||
　　//Hm::?||
637i～ あたしも好きだよなんかこの不思//議な食感。||
　　Atashi mo suki da yo nanka kono hushi//gi na shokkan.||
　　I also like it, I tell you somehow this amaz//ing texture. ||
638h　//うんそう得体||の知れない感が。
　　//Un sō tokutai|| no shirenai kan ga.
　　//Yeah (that's) right (this) strange mysterious sense.
　　((lit., sense that you can't know what (it) is)).
639g　　マジで？
　　Maji de?
　　(You) serious?
640g　なんかすごくアウェイな気分。{フフ}
　　Nanka sugoku awei na kibun. {hu hu}
　　Somehow (I have a) extremely "away" feeling (lit., feeling at an away game). {LAUGHTER}
641i　　うん。
　　Un.
　　Yeah.

Excerpt 7 occurs about four minutes later in the conversation after the participants discuss their preferences in the Senegalese course, size and softness of the meat in the *MAFE*, eating utensils, and the course's country of origin. Gin reopens the discussion of the *LAAX* in 764g with the non-predicate *kono … yōguruto no amai no ga* 'this sweet yogurt one' (N+*ga*), using *yōguruto* 'yogurt' and *amai* 'sweet', both aspects that were previously agreed upon in Excerpts 2 and 5, respectively. Subsequently, Haru's use in 768h of the non-predicate (N+Z) *kono shiroi bunibuni-shita no* 'this white jellylike one' combines a pre-nominal demonstrative with color and repeats Gin's description of the texture as

bunibuni 'jellylike' (630g). This suggests that Haru agrees with Gin's previous description. Next, in response to Haru's question in 767h-768h about what things in Japan have texture similar to the *LAAX*, Haru and Gin enumerate similar Japanese tea cakes (*gyūhi* in 769h, *kanten* '(sweet) agar', *suama* 'sweet mochi cake') with predicate nouns (not all shown in Excerpt 7).

Excerpt 7:
SENEGALESE COURSE-LAAX3 (25:02-25:21) (Unfamiliar food)

764g (5.4) が、このなんかヨーロ-、なんかヨーグルトの甘いのが解せぬっ。
(5.4) *ga, kono nanka yōro-, nanka yōguruto no amai no ga gesenut.*
(5.4) this somehow *yōro-*, somehow sweet yogurt one (I) can't comprehend.
765h {フフ}
{*hu hu*}
{LAUGHTER}
766i {フフフ}
{*hu hu hu*}
{LAUGHTER}
767h (2.4) ん、日本にもこうゆうさ、
(2.4) *N, Nihon ni mo koo yuu sa,*
(2.4) Yeah, in Japan too, this kind of you know,
768h この白いぶにぶにしたのなかったっけ？
kono shiroi bunibuni-shita no nakatta kke?
isn't there this white jellylike one?
769h (1.0) 求肥でもないし。
(1.0) *Gyūhi de mo nai shi.*
(1.0) (It)'s not *gyūhi* 'soft skin made of (steamed) refined rice flour and sugar', and.

3.2 Results of the analysis

In response to my first question "What aspects of the food do participants use as resources to create references for unfamiliar food?", initial references tended to be demonstrative pronouns (*kore* 'this') and gradually became more specific including features of color (*shiroi no* 'white one'), shape (*kono buttai* 'this object'), texture (*kono bunibuni-shita no* 'this jellylike one'), flavor (*yōguruto no amai no* 'yoghurty sweet one'), and combinations (*kono shiroi katamari* 'this white lump', *kono shiroi bunibuni-shita no* 'this white jellylike one'). This reflects the participants' multi-sensory experience of the food. Demonstrative pronouns and references involving color and shape were used without previous agreement. In contrast, food categories and features related to flavor and texture tended

to be used in non-predicate components only after agreement between participants had been established in predicate reference. In some cases a participant's use of another participant's previous predicate reference as a non-predicate component indicated their agreement (e.g. 769h).

In regard to the second question "What patterns in reference tracking can be observed through the conversation?", participants referred to the unknown food (*LAAX*) with and without using explicit verbal expressions. For example, they pointed at the unknown foods or drinks with their hands, fingers, and chopsticks and other eating utensils, and ellipted the reference once it was established. Focusing on explicit verbal referring expressions, I found that while the references for more familiar food were settled quickly and often used in non-predicate components from the start, references for less familiar foods started as predicate elements, and were monitored and modified during the discussion of the food item. Participants chose references with varying specificity, repeated, paraphrased or changed their references, chose references that were similar or different from their interlocutors, and used information other participants used to describe, assess and categorize the food in subsequent references to the food. Participants' repetition of their interlocutors' referring expressions and use of the referring expression as a non-predicate component suggested their agreement on descriptions and categorization of the food, and contributed to the stability of the referring expression for that food.

In Figure 1 I show how predicate and non-predicate references were tracked through the first three sections about the *LAAX*. In most cases I give the English translation for the Japanese references, or both the Japanese Romanization and English translation. I use lower case letters to indicate predicate elements, and capital letters for non-predicate components, and put large circles around categories and descriptions negotiated in the predicate. The large arrows indicate categories and descriptions agreed upon in the predicate (lower case letters) that were subsequently used in non-predicate references (capitalized). With the exception of references to color and shape, the large arrows show that all the non-predicate references occurred after they were used and agreed upon in predicates. These include the use of *yooguruto* 'yogurt' that was agreed upon in LAAX1 and subsequently used in a non-predicate component in LAAX2 (four times) and LAAX3 (three times). After agreeing on the *LAAX* being *amai* 'sweet' in LAAX2, this description was subsequently used to refer to the *LAAX* in non-predicate components in LAAX2 (one time) and in LAAX3 (two times). In addition, after the discussion of the texture of the *LAAX* in LAAX2 where Gin referred to it as *bunibuni* 'jellylike' in the predicate, Haru's use of this description in a non-predicate component in *LAAX*3 suggests that she agrees with this characterization of the *LAAX*'s texture.

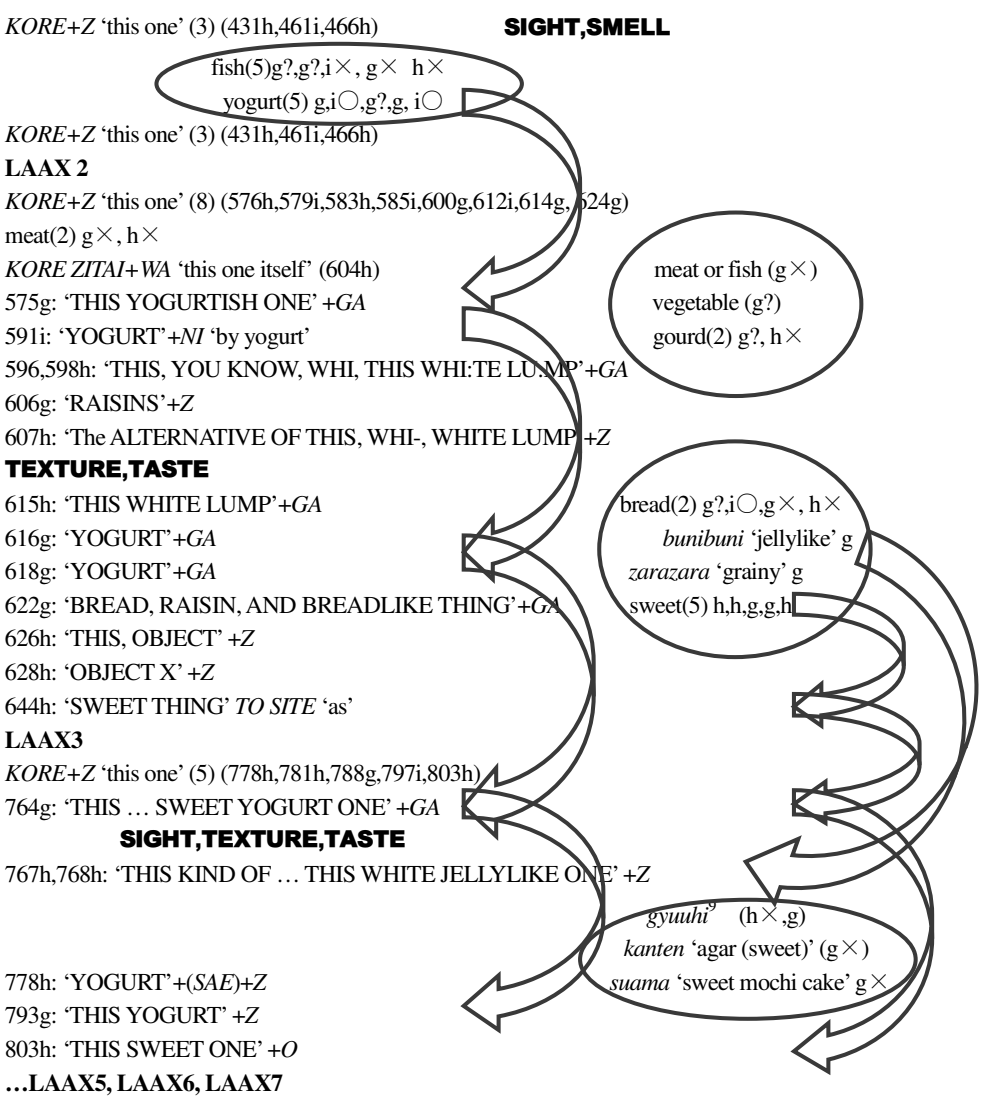
Finally, in regard to my third question "How do participants' choices of similar or different referring expressions influence their assessment and categorization of the

food and their relationships with one other?", participants accepted (with agreements and back channel utterances, repetition and paraphrase, etc.) or rejected (with direct negation, repetition with rising intonation, non-use, etc.) one another's predicate references. Choice of "referring expression" also influenced participants' positive/ negative assessment of the food, and acceptance and use of another's referring expression or description, in particular as a non-predicate component, contributed to the closeness among participants.

4 Conclusion

Results suggest ways in which participants adapt language to unfamiliar food in the process of negotiating food references based on their multi-sensory experience, knowledge, assessment and categorization of the food in the talk-in-interaction. Unlike previous studies that focused on the use of non-predicate references in narratives that tended to be monologic, this study demonstrates that references in talk-in-interaction over food are negotiated (primarily) in the predicate and then used in non-predicate components after establishing agreement. By elucidating the negotiation of assessments, categories, and knowledge in Japanese talk about food, this study aims to contribute to research on contextualized social and cognitive activity, language and food, and cross-cultural understanding.

Figure 1. Referring expressions used in JPN3 (lower case letters=predicate elements, capital letters=non-predicate components; large circle= categories and descriptions negotiated in the predicate, large arrows= categories and descriptions agreed upon in the predicate (lower case letters) that were subsequently used in non-predicate references (capitalized); g=Gin, h=Haru, i=Iku; ?=question,×=disagreement, ○=agreement; *WA*= topic, *GA*=subject, *O*=direct object, *NI*=indirect object, *Z*=no particle; (#)=number of occurrences)



9 'soft skin made of (steamed) refined rice flour and sugar'

Acknowledgments

I would like to thank Professor Emerita Midori Takasaki and Professors Natsuko Hirose, and Midori Kasai (Ochanomizu University), and Assistant Professor Yuko Hoshino (Jumonji University) for their assistance in data collection for this study. I am also grateful to Saori Yamada, Aya Harada, and Masaichi Akiyama for their assistance in collecting and transcribing the data, and I am thankful to the Taster Lunch participants, whose names will remain anonymous. This research was supported by a University of Minnesota Grant-in-Aid of Research, Artistry and Scholarship ("Sensory Evaluation of Food and Cultural Identity in English, Japanese and Wolof") 2009-2011 and a Hakuho Foundation Japanese Language Research Fellowship (2012-2013).

Transcription conventions for Japanese/ romanized Japanese and English[10]

(Chafe 1980; Levinson 1983; Atkinson & Heritage 1984; Szatrowski 1993, 2000, 2004, 2005, 2010, 2013, 2014 a,b,c,d, 2015 a,b)

./。	falling sentence-final intonation.
?/？	rising intonation, not necessarily a question.
,/、	continuing intonation followed by a slight pause.
↑	slight rise in intonation.
.h h	in-breath (.h), out-breath (h); number of 'h's' indicates the length in relation to the length of preceding syllables/mora.
@ @	utterance between the @ @ is said in a laughing voice.
{}	enclose non-linguistic sounds such as laughter, coughing, clicks, etc. Whenever possible the beats and sounds of the laughter are transcribed, e.g., {ハハハ}, {ha ha he }.
{.h}{h}	laughter consisting of an in-breath or out-breath, respectively.
° °	utterance between the ° ° is said in a quieter voice.
・・	utterance between the ・・is said in a louder voice.
< >	indecipherable or slightly audible speech is indicated in < >.
(1.7)	length of pause/silence in seconds, (0.7) indicates a pause of 7-tenths of a second, relative to the speed of the preceding utterance.
// ||	// || indicates where the overlap begins and ends in the present and following utterance.

10 The symbol on the left of the '/' is used in the romanized version of the Japanese and English translation, and the one on the right in the Japanese transcription.

:/— indicates lengthening of the preceding vowel or syllabic nasal in the English and Romanized/ Japanese version of the transcript.
- cut-off of preceding sound.
= a single = sign at the end of an utterance indicates that the next utterance (by the same speaker) continues on to the next line without a pause.
〜 postposing
→i addressee of the utterance; →i indicates that the utterance is addressed to i.
※ explanation of nonlinguistic behavior accompanying an utterance.
(( )) additional explanation of eating and other non-linguistics behavior in pauses between utterances, content of untranscribed talk, etc.

English translation: ( )= words necessary in English, but not reflected in the Japanese.

Back channel utterances and laughter are moved to the right to line up with the end of the previous utterance to which they respond.

Literature

Atkinson, J.M. & Heritage, J. (1984) *Structures of Social Action: Studies in Conversation Analysi*s. Cambridge: Cambridge University Press.

Beeman, W. O. (2014) Negotiating a passage to the meal in four cultures. In: Szatrowski, P. (ed.) *Language and Food: Verbal and Nonverbal Experiences*: 31-52. Amsterdam: John Benjamins.

Burdelski, M. (2014) Early experiences with food: Socializing affect and relationships in Japanese. In: Szatrowski, P. (ed.) *Language and Food: Verbal and Nonverbal Experiences*: 233-255. Amsterdam: John Benjamins.

Chafe, W. L. (ed.) (1980) *The Pear Stories: Cognitive and Linguistic Aspects of Narrative Production*. Norwood, NJ: Ablex Publishing Corporation.

Clancy, P. (1980) Referential choice in English and Japanese narrative discourse. *The Pear Stories: Cognitive and Linguistic Aspects of Narrative Production*: 127-202. Norwood, NJ: Ablex Publishing Corporation.

Heritage, J. (2012) Epistemics in action: Action formation and territories of knowledge. *Research on Language and Social Interaction* 45 (1): 1-29.

Kamio, A. (1994) The theory of territory of information: The case of Japanese. *Journal of Pragmatics* 21 (1): 67-100.

Karatsu, M. (2014) Repetition of words and phrases from the punch lines of Japanese stories about food and restaurants: A group bonding exercise. In: Szatrowski, P. (ed.) *Language and Food: Verbal and Nonverbal Experiences*: 185-207. Amsterdam: John Benjamins.

Koike, C. (2014) Food experiences and categorization in Japanese talk-in-interaction. In: Szatrowski, P. (ed.) *Language and Food: Verbal and Nonverbal Experiences*: 159-183. Amsterdam: John Benjamins.

Kuroshima, S. (2014) The structural organization of ordering and serving sushi. In: Szatrowski, P. (ed.) *Language and Food: Verbal and Nonverbal Experiences*: 53-75. Amsterdam: John Benjamins.

Labov, W. and Fanshel, D. (1977) *Therapeutic Discourse: Psychotherapy as Conversatio*n. New York, NY: Academic Press.

Levinson, S. C. (1983) *Pragmatics*. Cambridge: Cambridge University Press.

Minami, F. (1974) *Gendai Nihongo no Kōzō* [Structure of Modern Japanese]. Tokyo: Taishukan Shoten.

Minami, F. (1993) *Gendai Nihongo Bunpō no Rinkaku* [Outline of Modern Japanese Grammar]. Tokyo: Taishukan Shoten.

Minami, F. (1997) *Gendai Nihongo Kenkyū* [Research on Modern Japanese]. Tokyo: Sanseido.

Noda, M. (2014) It's delicious!: How Japanese speakers describe food at a social event. In: Szatrowski, P. (ed.) *Language and Food: Verbal and Nonverbal Experiences*: 79-102. Amsterdam: John Benjamins.

Ono, T. & Suzuki, R. 1992. Word order variability in Japanese conversations. *Text* 12 (3):429-455.

Szatrowski, P. (1993) *Nihongo no Danwa no Kōzō Bunseki- Kanyū no Danwa no Sutoratejī no Kōsatsu* [Structure of Japanese Conversation--Invitation Strategies]. Tokyo: Kurosio Publishers.

Szatrowski, P. (2000) Kyōdō hatsuwa ni okeru sankasha no tachiba to gengo/ higengo kōdō no kanren ni tsuite [Relation between participant status and verbal/ non-verbal behavior in co-construction]. *Nihongo Kagaku* [Japanese Linguistics] 7: 44-69.

Szatrowski, P., ed. (2004) *Hidden and Open Conflict in Japanese Conversational Interaction*. Tokyo: Kurosio Publishers.

Szatrowski, P. (2005) Danwa to buntai--Kanjō hyōka no dōteki na katei ni tsuite--(Japanese conversation and style--The dynamic process of emotion/ evaluation--). In: Nakamura, A., Nomura, M., Sakuma, M. and Komiya, C. (eds.) *Hyoogen to Buntai* [Expression and Style]: 468-480. Tokyo: Meiji Shoin.

Szatrowski, P. (2007) Subjectivity, perspective and footing in Japanese co-construction. In: Hedberg, N. and Zacharski, R. (eds.) *Topics on the Grammar-Pragmatics Interface: Essays in Honor of Dr. Jeanette K. Gundel*: 313-339. Amsterdam: John Benjamins.

Szatrowski, P. (ed.). (2010) *Storytelling across Japanese Conversational Genr*e. Amsterdam: John Benjamins.

Szatrowski, P. (2011) Shishokukai no gengo/ higengo kōdō ni tsuite--30-sai miman no josei gurūpu o chūshin ni [Verbal and nonverbal behavior in a Taster Lunch-Focusing on a group of women under 30]. *Kyōiku Kenkyū Sentā Kenkyū Nenpō* [Center for Comparative Japanese Studies Annual Bulletin] 7: 281-292. (Ochanomizu University). http://teapot.lib.ocha.ac.jp

Szatrowski, P. (2013) Tabemono o hyōka-suru sai ni mochiirareru "kyakukanteki hyōgen" to "shukanteki hyōgen" ni tsuite [On the use of "subjective" and "objective" expressions for food assessment in Japanese]. *Kokuritsu Kokugo Kenkyūjo Ronbunshū* [NINJAL Research Papers] 5: 95-120. http://doi.org/10.15084/00000506

Szatrowski, Polly. (2014a) Modality and evidentiality in Japanese and American English Taster Lunches: Identifying and assessing an unfamiliar drink. In: Szatrowski, P. (ed.) *Language and Food: Verbal and Nonverbal Experiences*: 131-156. Amsterdam: John Benjamins.

Szatrowski, P. (ed.). (2014b) *Language and Food: Verbal and Nonverbal Experiences*. Amsterdam: John Benjamins.

Szatrowski, P. (2014c) Shishokukai ni okeru tabemono to kazoku to no kankei [On the relation between food and family in Taster Lunches]. Hikaku Nihongaku Kyōiku Kenkyū Sentā Kenkyū Nenpō [Center for Comparative Japanese Studies Annual Bulletin] 10:231-238. http://teapot.lib.ocha.ac.jp

Szatrowski, P. (2014d) Sōgo sayō ni mirareru gengo to bunka no setten--Sutoratejī, danwa no kōsei tan'i, modariti to ebidenshariti ni tsuite--[The intersection of language and culture in interaction: Strategies, discourse units, and modality/ evidentiality]. *Nihon Gengo Bunka Kenkyūkai Ronshū* [Japan Language Culture Research Society Papers] 10:1-17. [The Japan Foundation Japanese-Language Institute, Urawa.] http://www3.grips.ac.jp/~jlc/jlc/ronshu/2014/1szatrowski.pdf

Szatrowski, P. (2015a) Shishokukai ni okeru onomatope [On the use of onomatopoeia in Japanese Taster Lunches]. *Japanese Language Education in Europe 19: The Proceedings of the 18th Japanese Language Symposium in Europe 28-30 August, 2014*: 95-100. Association of Japanese Language Teachers in Europe e.V. [ATE] http://www.eaje.eu/media/0/myfiles/12%20Panel3%283%29_Szatrowski.pdf

Szatrowski, P. (2015b) Nihongo no shishokukai ni okeru modariti to ebidenshariti no mochikata--Nihongo bogowasha to hibogowasha no Amerikajin to no chigai [Use of modality and evidentiality in Japanese Taster Lunches: Differences between Japanese native speakers and nonnative speaking Americans]. In: Abe, J., Iori, I., and Sato, T. (eds.) *Bunpō/ Danwa Kenkyū to Nihongo Kyōiku no Setten* [Intersection of Grammar/ Discourse Research and Japanese Language Education]: 159-177. Tokyo: Kuroshio Publishers.

Szatrowski, P. (2016) Michi no tabemono e no genkyū no sikata--Shishokukai ni okeru dōtei to kyōkan--[Referring to unfamiliar food: Identification and agreement in Japanese taster lunches]. *Kokuritsu Kokugo Kenkyūjo Ronbunshū* [NINJAL Research Papers] 11:93-115. http://doi.org/10.15084/00000843

Szatrowski, P. (2017) Kichi to michi no tabemono o meguru mandara: Shishokukai no kaiwa o rei ni [Mandala (visual schema) revolving around familiar and unfamiliar foods: Examples from Taster Lunches]. In: *Jikan no Nagare to Bunshō no Kumitate: Hayashi Gengogaku no Saikaishaku* [The Dynamic Construction of Discourse over Time: Interpretations of Hayashi Linguistics]: (eds.) I. Iori, K. Ishiguro, and T. Maruyama, pp. 239-268. Tokyo: Hituzi Syobo.

Szatrowski, P. (2018) Sōgo sayō ni yoru onomatope no shiyō--Nyūseihin no shishokukai o rei ni shite--[On the use of onomatopoeia in interaction: Examples from Japanese Dairy Taster Brunches]. *Kokuritsu Kokugo Kenkyūjo Ronbunshū* [NINJAL Research Papers] 16:77-106. http://doi.org/10.15084/00001609

Watanabe, F. (2009) Eigo oyobi nihongo no katari no danwa bunshō ni okeru shijishi [Demonstratives in English and Japanese oral and written narrative discourse]. *Yamagata Daigaku Jinbun Gakubu Kenkyū Nenpō* [Yamagata University, Faculty of Literature & Social Sciences Annual Research Report] 6: 1-13.

Watanabe, F. (2010) Clausal self-repetition and pre-nominal demonstratives in Japanese and English animation narratives. In: Szatrowski, P. (ed.) *Storytelling across Japanese conversational genre*: 147-180. Amsterdam: John Benjamins.

要旨（Abstract in Japanese)

「日本語の試食会における未知の食べ物への言及表現の追跡
ー食べ物に言語を適応しながら同意を得ようと交渉することー」

ポリー・ザトラウスキー（ミネソタ大学）

本研究では、日本語による試食会でどのように未知の食べ物に言及するか、言及表現を用いていく際にどのようなパターンが見られるのか、言及表現は食べ物の評価、範疇化、参加者の人間関係にどのように影響するかを考察する。未知の食べ物の言及表現は、初めに指示代名詞が用いられ、その後、色、形、食感、味、またそれらの組み合わせにより特定化される。これは参加者の複数の身体感覚の体験（multi-sensory experience) を反映している。既知の食べ物は言及表現が早く決まるのに対し、未知の食べ物は、話の間中、言及表現が変化する。また、言及表現の選択はほかの参加者の評価と範疇化に影響を与える。他の参加者が食べ物を描写、評価、範疇化する際に用いた表現を南 (1974,1993) の述部的な要素以外の成分で繰り返すことは、その描写と範疇化に対する同意を示し、言及表現が安定する。本研究は、文脈に応じた社会的・認知的な活動、食べ物と言語、異文化理解の研究に貢献できる。

4 The grammar and discourse functions of Japanese cleft sentences

SUNAKAWA Yuriko
University of Tsukuba

Abstract
There are two types of Japanese cleft sentences: WA-clefts and GA-clefts. These have the following grammatical characteristics.

1) The predicate of a WA-cleft can either be a noun or a subordinate clause, whereas the predicate of GA-clefts is restricted to nouns.
2) Both WA-clefts and GA-clefts show a tendency for the predicate noun not to be accompanied by a *kaku-joshi* (case particle). However, this tendency is much stronger with GA-clefts than WA-clefts.

This paper aims to show that the above characteristics are not syntactic restrictions but preferred patterns of the use of cleft sentences in discourse.

I make the following two claims:

a) Japanese cleft sentences have two types of discourse function, namely 'focus-presentational function' and 'prominence-presentational function.'
b) The above-mentioned grammatical characteristics of WA-clefts and GA- clefts can be explained by their discourse functions.

Keywords: cleft sentence, grammar, discourse, focus-presentational function, prominence-presentational function, topic development, WA and GA

1 Introduction

This paper focuses on the grammatical characteristics and functions of cleft sentences. It is argued that the characteristics of cleft sentences are not syntactic restrictions as has hitherto been claimed, but that these are the result of preferences in patterns of the use of cleft sentences in discourse. The paper thus claims that the grammar that is often regarded as arbitrary can have non-arbitrary characteristics underpinned by particular functions. There are two claims discussed in this paper:

a) Japanese cleft sentences have focus-presentational and prominence-presentational functions.
b) The grammatical characteristics of WA-clefts and GA-clefts can be explained by their discourse functions.

2 Definition of clefts

A cleft sentence may be defined as a copula sentence where the subject is a clause, and the predicate consists of a specific element within the clause, such as[1]:

(1) *Sono toki, fukai mori no oku kara arawareta no wa,*
that time dense forest of depth from appeared NOM TOP
ippiki no kuma datta.
one bear copula-PAST
'Just then, what appeared from the depths of the dense forest was a bear.'

(2) *Sono toki, fukai mori no oku kara arawareta no ga,*
that time dense forest of depth from appeared NOM SUBJ
ippiki no kuma datta.
one bear copula-PAST
'Just then, what appeared from the depths of the dense forest was a bear.'

These sentences share the same propositional meaning as the following sentence:

(3) *Sono toki, fukai mori no oku kara ippiki no kuma ga arawareta.*
that time dense forest of depth from one bear SUBJ appeared
'Just then, a bear appeared from the depths of the dense forest.'

(1) and (2) take the subject of (3) *ippiki no kuma* (a bear) and place it in the predicate position, and use the clause *Sono toki, fukai mori no oku kara arawareta* (Just then, what appeared from the depths of the dense forest) as the subject. In this paper, WA-clefts are defined as sentences of the type shown in (1) that have …*no wa*…*da* and GA-clefts are defined as sentences of the type shown in (2) that have …*no ga*…*da*.

Previous research (including Kumamoto (1989), Sunakawa (1995) and Noda (1996)) has shown that the following differences can be found in the grammatical behaviour of WA and GA-clefts:

a) WA-clefts can take subordinate clauses as their predicate but GA-clefts do not.
b) The predicate noun of WA-clefts can take *kaku-joshi* (hereafter 'case particle')[2] but GA-clefts do not.

---

1 The abbreviations in the glosses are:
NOM (nominalizer), TOP (topic), SUBJ (subject), OBJ (object), LOC (locative), Q (question marker)

2 *kaku-joshi* (case particle) consists of grammatical case particles such as '*ga* (SUBJ)', '*o* (OBJ)' and semantic case particles such as '*kara* (from)', '*made* (until). In this paper, *fukugō-ji* (compound particles) such as '*ni oite* (at)' and '*ni totte* (for)' are included in *kaku-joshi*.

In the next section, the above statements will be examined and it will be explained why a) is entirely plausible while b) requires some modification and further clarification.

3 Grammatical characteristics of WA-clefts and GA-clefts

3.1 Predicate type of WA-clefts and GA-clefts

In this paper, examples are taken from ten magazines, ten essays, three novels and two textbooks. As there is an abundance of examples of typical WA-clefts, and I would like to focus on the analysis of atypical WA-clefts, only a fraction of typical WA-clefts have been used and a search for examples has concentrated on atypical examples of WA-clefts. However, as the occurrence of GA-cleft is rare, I have collected and used all examples of GA-cleft found in the afore-mentioned materials. Table 3.1 shows the predicate type of the collected examples.

Typical type and atypical type of WA-clefts and GA-clefts are as follows:

A) Typical WA-clefts and typical GA-clefts are the ones whose predicate nouns are not accompanied by case particles.

B) Atypical WA-clefts are the ones whose predicate nouns are accompanied by case particles, or the ones whose predicates are subordinate clauses.

C) Atypical GA-clefts are the ones whose predicate nouns are accompanied by case particles.

Table 3.1. *Predicate type of WA-clefts and GA-clefts*

	Typical Type	Atypical Type			
	Noun	Noun + Case particle	Subordinate Clause	Adverbs	Total
WA-clefts	185 (84%)	11 (5%)	21 (10%)	3 (1%)	220 (100%)
GA-clefts	94 (99%)	1 (1%)	0 (0%)	0 (0%)	95 (100%)

As stated above, not all WA-cleft examples of typical types are accounted for here. If all examples are taken into consideration, the frequency of WA-clefts is much greater.

Table 3.1 shows that the predicates of GA-clefts are mostly nouns, whereas WA-clefts have a variety of predicate types.

3.2 Use of subordinate clause as predicate

WA-clefts can use subordinate clauses as their predicate, as shown below:

(4) *Hitozato ni kuma ga arawareta no wa, mori ni shokuryō ga*
settlement LOC bear SUBJ appeared NOM TOP forest LOC food SUBJ
fusoku shite iru tame da.
be short of because copula
'The reason why the bear appeared in the settlement is because there was a shortage of food in the forest.'

In case of GA-clefts, this is not permissible[3]:

(5) **Hitozato ni kuma ga arawareta no ga, mori ni shokuryō*
settlement LOC bear SUBJ appeared NOM SUBJ forest LOC food
ga fusoku shite iru tame da.
SUBJ be short of because copula
'*The reason why the bear appeared in the settlement is because there was a shortage of food in the forest.'

As has been noted in previous research, and also in Table 3.1, subordinate clauses that take phrases such as …*tame* (because…), …*okage* (thanks to…), …*kara* (due to…) can only be used as predicates of WA-clefts and are not permissible in the case of GA-clefts.

The reason why GA-clefts cannot take subordinate clauses as predicates will be discussed in Section 5.2. and 5.3.

3.3 Use of case particles for predicate nouns

Predicate nouns of WA-clefts can take case particles as shown below:

(6) *Sono toki, ippiki no kuma ga arawareta no wa,*
that time one bear SUBJ appeared NOM TOP
fukai mori no oku kara datta.
dense forest of depth from copula-PAST
'Just then, a bear appeared from the depths of the dense forest.'

3 The symbol '*' denotes that the following sentences are not permissible.

As shown in (6), the predicate noun phrase '*fukai mori no oku* (the depths of the dense forest)' is accompanied by a case particle '*kara*.'

On the other hand, a GA-cleft sentence whose predicate noun accompanies a case particle as shown below may not be considered well-formed[4]:

(7) ? *Sono toki, ippiki no kuma ga arawareta no ga,*
that time one bear SUBJ appeared NOM SUBJ
fukai mori no oku kara datta.
dense forest of depth from copula-PAST
'Just then, a bear appeared from the depths of the dense forest.'

However, there are some instances where GA-clefts have their predicate nouns accompanying case particles and yet are still considered well formed.

(8) *Soshite sono kekka, futatabi Kanamaru-Tanabe rain ga migoto ni*
and this result again *Kanamaru-Tanabe* line SUBJ brilliantly
kinō shita no ga, 58-nen 12-gatsu no kaisan-sōsenkyo
functioned NOM SUBJ '58 December of general election after the dissolution
ni oite datta.
at copula-PAST
(*Bungei Shunjū*, Jan 1993)[5]
'And as a result, when the Kanamaru-Tanabe line functioned brilliantly again was at the general election after the dissolution of the government in December '58.'

This example (8) contains the compound particle *ni oite* (at), which functions as a case particle, with the predicate noun *kaisan-sōsenkyo* (general election after dissolution of the government). Also, the following example is not considered ill formed:

(9) *Kanamaru ga Tanabe o mikagitta no ga, masani sono riyū*
Kanamaru SUBJ Tanabe OBJ severed NOM SUBJ very this reason
de datta.
for copula-PAST
'Why Kanamaru severed the relationship with Tanabe was because of this very reason.'

4 The symbol '?' denotes that the following sentences may be permissible but sound unnatural.

5 Indicated in parentheses are the sources of examples. Those that do not show the sources are examples composed by the author.

As shown in Table 3.1, out of the collected 95 examples of GA-clefts, there was only one example (example (8)) that contained a case particle with the predicate noun.

Although in cases of GA-clefts their predicate nouns seldom take case particles, under certain conditions well-formed sentences can be constructed.

On the other hand, WA-clefts have comparatively more examples that accompany case particles with their predicate nouns. But still, the number is limited. On examining examples of WA-clefts, it becomes clear that there are not as many examples of predicate nouns bearing case particles. As shown in Table 3.1, out of the collected 220 examples of WA-clefts, only 11 bore case particles with their predicate nouns, which represent a mere 5% of the total examples. As mentioned in 3.1, not all WA-clefts examples of typical types are accounted for here. If all examples are taken into consideration, the percentage will be much smaller than the 5% quoted here.

It is clear from these findings that not only GA-clefts but also WA-clefts seldom take case particles with their predicate nouns unless certain conditions are met.

Next, let us consider what sorts of conditions are necessary for predicate nouns to carry case particles.

3.4 Conditions for predicate nouns to carry case particles

First let us examine the following examples:

(10) *Jinkō-chinō ni rakkan-shugi ga atta no wa*
artificial intelligence LOC optimism SUBJ existed NOM TOP
1980-nendai made deshita.
1980s until copula-PAST
(*Bungei Shunjū* Jan 1993)
'People were optimistic about artificial intelligence until the 1980s.'

The underlined case particle of the above example can be removed and still remain well-formed:

(11) *Jinkō-chinō ni rakkan-shugi ga atta no wa 1980-nendai deshita.*
'People were optimistic about artificial intelligence in the 1980s.'

The meanings of the two sentences are not the same. In the case of (10), a number of years prior to 1980 are included, while in the case of (11) only years in the 1980s are indicated. As shown here, by removing the case particles, the original meaning cannot be conveyed accurately.

Let us go back to the example of (8), relabelled here as (12).

(12) *Soshite sono kekka, futatabi Kanamaru-Tanabe rain ga migoto ni*
and this result again Kanamaru-Tanabe line SUBJ brilliantly
kinō shita no ga, 58-nen 12-gatsu no kaisan-sōsenkyo
functioned NOM SUBJ '58 December of general election after the dissolution
ni oite datta.
at copula-PAST

(*Bungei Shunjū*, Jan 1993)

'And as a result, when the Kanamaru-Tanabe line functioned brilliantly again was at the general election after the dissolution of the government in December '58.'

The predicate noun *kaisan-sōsenkyo* (general election after dissolution of the government) does not only show the 'time' or 'place' but forms a combined and more abstract concept of 'aspect,' 'scene,' 'situation' and so on. It would appear that the reason why *ni oite* (at) is used in (12) is because, unlike simple 'time' or 'place,' without the case particle, the relationship between the noun *kaisan-sōsenkyo* (general election after dissolution of the government) and the predicate of the subject clause, *kinō shita* (functioned) is difficult to define.

From the above observations, it is possible to postulate that it is necessary to use case particles when the meaning is lost or the relationship between the noun and the predicate of the subject clause becomes vague without them. In any other circumstance, i.e., when the meaning is conveyed without the aid of a case particle, those case particles are usually omitted.

On the other hand, however, there are some cases as shown below where case particles are still used even though the relationship between the noun and the predicate is easily recognisable and the original meanings of the sentence can still be conveyed without using such case particles:

(13) *Odoroita no wa, sono nedan no yasusa ni desu.*
surprised NOM TOP this price of cheapness by copula

(Noda 1996)

'What I was surprised by was its cheapness.'

In this example, the original meaning of the sentence can just as easily be conveyed by removing the case particle *ni* (by) resulting in *odoroita no wa, sono nedan no yasusa desu* (What I was surprised by was its cheapness). Why then, was the case particle preserved? The answer to this question becomes apparent by observing the particular context of this sentence in the example.

This sentence was used in the second paragraph at the beginning of an essay as shown below.

(14) *Mirano shinai ni aru Aritaria no ofisu de kaimotometa passenjā-chiketto ni LIT2276000 to insatsu sarete iru no o mite, boku wa odorokimashita. Mirano-Roma-Tokyo no bizinesu kurasu, katamichi chiketto no nedan desu. Yōroppa no aru toshi made no katamichi chiketto o, Itaria kokunai de katta baai no nedan desu to ii naoshite mo yoi deshō.*

Odoroita no wa sono nedan no yasusa ni desu. 100-rira≒11.55-en to shite, 262,878-en desu. Ga, dentaku o tatakanaku tomo, 20-man-en-dai de aru koto kurai, dare ni datte wakarimasu. Nihon de katta baai, 405,400-en suru koto o shitte ita boku ga odoroita no mo muri arimasen, 142,522-en mo no sa ga aru no desu kara.

(Yasuo Tanaka *Faddish Kogengaku*, p.156.)

'I was surprised to find the price of LIT2276000 printed on the passenger ticket purchased at an Alitalia office in Milan. It was the one-way business class ticket for the route of Milan-Rome-Tokyo. Or, it may be rephrased as the price of a ticket between a city in Europe and Tokyo bought in Italy.

What I was surprised by was the cheapness of the price. Using the approximate rate of Lire 100 = Yen 11.55, it makes 262,878 Yen. Even without using a calculator, anyone can figure out that it is something between 200,000 and 300,000. It is not surprising that I was surprised at the figure, as I had known the price of such tickets in Japan to be 405,400 Yen if bought in Japan. The difference was 142,522 Yen.'

In the two paragraphs shown above, the first paragraph explains the surprise the author of the text felt upon seeing the one-way business class price for the route of Milan-Rome-Tokyo. The WA-cleft sentence in question is at the beginning of the second paragraph and it indicates that the reason why the author was astonished was the cheapness of the ticket. The content of the second paragraph beginning with this WA-cleft is that the tickets bought in Italy are much cheaper than those bought in Japan. The topic of this paragraph is the 'cheapness of tickets' and the same topic is carried on and on to the sixth paragraph. The reason why a case particle is used for the predicate noun in the cleft sentence at the beginning of the second paragraph is to emphasise the 'cheapness of the tickets' and by doing so, the topic 'cheapness of the tickets' becomes more prominent, so that it draws the attention of the listener and can be discussed for a length time in subsequent paragraphs.

Similarly, let us consider the GA-cleft example shown in (15):

(15) *Kanamaru ga Tanabe o mikagitta no ga, masani sono riyū*
Kanamaru SUBJ Tanabe OBJ severed NOM SUBJ very this reason
de datta.
copula-PAST
'Why Kanamaru severed the relationship with Tanabe was because of this very reason.'

In this example, the predicate noun *sono riyū* (the reason) is emphasised by using the adverb *masa ni* (very), and this can be interpreted, by making the noun prominent, that it is made easier for the predicate of the GA-cleft to take on a case particle.

As in these examples, even for GA-clefts, if certain conditions are met, it becomes possible for their predicate nouns to accompany case particles.

As will be discussed in Section 5.3 in detail, predicate nouns in GA-clefts do not usually indicate other cases than the nominative or accusative. Out of 185 WA-clefts and 94 GA-clefts that belong to the typical type (cf. Table 3.1), the ratio of the GA-cleft examples which have predicate nouns other than nominative or accusative is 14%, which is much smaller compared to WA-cleft, where the ratio of predicate nouns other than nominative or accusative is 37%, as shown in Table 3.2.

Table 3.2. *Grammatical or semantic relations of the predicate nouns of WA-clefts and GA-clefts*

	Nominative/Accusative			Others					
	Nominative	Accusative	Total	Locative	Dative	Genitive	Time	Others	Total
WA-clefts	96 (52%)	21 (11%)	117 (63%)	9 (5%)	2 (1%)	1 (1%)	39 (21%)	17 (9%)	68 (37%)
GA-clefts	70 (74%)	11 (12%)	81 (86%)	0 (0%)	2 (2%)	0 (0%)	7 (7%)	4 (4%)	13 (14%)

This is the reason why example (7) given in the previous section, shown as (16) below, appears to be ill-formed. As shown below, the noun in the predicate, *oku* (depth), is ablative: an unusual case to appear in GA-cleft sentences.

(16) ? *Sono toki, ippiki no kuma ga arawareta no ga,*
that time one bear SUBJ appeared NOM SUBJ
fukai mori no oku kara datta.
dense forest of depth from copula-PAST
'Just then, a bear appeared from the depths of the dense forest.'

If this sentence is modified by adding the phrase *nanto* (surprisingly, would you believe) it emphasises the situation, and given more dramatic contexts, it becomes more readily acceptable as a well-formed sentence:

(16') *Sono toki, ippiki no kuma ga arawareta no ga, nanto,*
this time one bear SUBJ appeared NOM SUBJ incredibly
osoroshii majo ga sumu to iu fukai mori no oku kara datta.
wicked witch SUBJ live say dense forest of depth from copula-PAST
'Just then, a bear appeared from, incredibly, the depths of the dense forest where that wicked witch was supposed to live.'

From the above, it is possible to conclude that regardless whether it is a WA or GA cleft sentence, so long as pragmatic conditions in discourse, such as when an important topic has to be emphasised and sustained in subsequent paragraphs or there is a need to emphasise the referent in the development of the discourse, it becomes possible for predicate nouns to accompany case particles.

In the following section, the grammatical similarities and differences in terms of WA-clefts and GA-clefts are discussed.

3.5 Differences and similarities of GA-clefts and WA-clefts

Grammatical similarities in both GA and WA clefts may be summarised as below:

i. Predicate nouns of both cleft types can take case particles but occurrence rate is low.
ii. In order for the predicate nouns of either cleft type to take case particles, the following discourse-pragmatic conditions must be met.[6]
    a) When the relationship between the predicate noun and the predicate in the subject clause becomes unclear due to the absence of the case particle.
    b) When the emphasis is to be added or attention is to be drawn to something.

As for differences, the following may be included:

iii. As predicates of WA-clefts, not only noun phrases but also subordinate clauses are used whereas in GA-clefts, only noun phrases are used.
iv. The tendency of not having case particles in predicate nouns is stronger in the case of GA-clefts.

In the following section, the discourse functions of cleft sentences will be described. By doing so, it will become clear that the differences as listed in (iii) and (iv) are based on the difference in discourse functions of WA and GA-clefts. Furthermore, what used to be considered as restrictions at sentence level are in fact merely the manifestation of the patterns that are favoured in discourse.

6 The examples of clefts that incorporate case particles (11 examples of WA-clefts and 1 example of GA-cleft) collected by the author were found to be all of the a) type. The example in (12) is a b) type example but was borrowed from Noda (1996).

4 Discourse functions of WA-clefts and GA-clefts

Both WA-clefts and GA-clefts can perform the function of focus presentation. On the other hand, there is a GA-cleft unique function of prominence marking. In the following section, these two functions are discussed in turn.

4.1 Focus-presentational function

The focus-presentational function is a function that fills the information gap between the speaker and the listener by providing the information that it lacks in the presupposition. For example, (17) is a sentence that is based on the presupposition and assertion given in (18):

(17) *Sono toki arawareta no wa ippiki no kuma datta.*
this time appeared NOM TOP one bear copula-PAST
'What appeared at that moment was a bear.'

(18) Presupposition: X appeared at that moment
Assertion: X is a bear

As shown above, a WA-cleft is a sentence where a proposition that contains a variable X ('open proposition' in Prince 1986) is the subject and its predicate is the focus, and the function of the cleft sentence is to assign a value to the variable by providing the focus information. In this paper, a function that fills the information gap between the speaker and the listener in communication is termed 'the focus-presentational function.' The focus-presentational function is, in other words, a function that provides information that is lacking in the proposition and by defining 'X is Y' in response to the question of what information X provides.

Next, let us examine the focus-presentational function of GA-cleft:

(19) *Soredewa, Nihon wa dō darō ka. / Rēsen-go no sekai e no taiō*
so Japan TOP how copula Q cold-war-post of world to of response
ni mottomo deokureta no ga Nihon de aru[7].
to most slow in action NOM SUBJ Japan copula
(*Bungei Shunjū* Jan 1993)

'So, how about Japan? / The country that was the slowest in responding to the post cold war world was Japan.'

7 The mark / denotes the end of a paragraph.

In this example, the GA-cleft sentence provides the answer 'being the slowest in responding to the post-cold war world' to the question of 'how about Japan?" The presupposition and the assertion may be summarized as below:

(20) Presupposition: Japan is X.
Assertion: The X is 'the slowest in responding to the post-cold war world.'

In summary, while the focus of WA-clefts is in the predicate, the focus of GA-clefts is in the subject. While the information of WA-clefts is presented in the sequence of 'presupposition → focus,' the information of GA-clefts is presented in the sequence of 'focus → presupposition.' These types of GA-clefts are also used to fill the information gap between the listener and the speaker, and therefore can be considered as sentences that perform a focus-presentational function.

4.2 Prominence-presentational function

Prominence-presentational function may be defined as a function that presents the referent prominently and draws the attention of the listener to it. This function is performed by GA-cleft sentences.

The underlined predicate noun, *ippiki no ōkina kuma* (a big bear), does not indicate the referent that has been conveyed from the previous discourse but the one that appears in this discourse for the first time.[8]

(21) *Watashitachi wa satsuei o akiramete, sono ba o tachisarō to shita.*
we TOP shooting OBJ gave up that place OBJ be about to leave
Sono toki, mori no oku kara arawareta no ga, nanto
that time forest of depth from appeared NOM SUBJ my goodness
ippiki no ōkina kuma datta.
one big bear copula-PAST
'We gave up shooting and were about to leave the place. Just then, what appeared from the depths of the forest was, oh my goodness a big bear.'

As shown in (21), the predicate nouns of the GA-clefts performing the prominence-presentational function indicate the referents in the current discourse for the first time. This is the essential difference with the type of GA-clefts performing focus-presentational function as shown in (19) where the predicate nouns are the presupposed information. Let us consider one more example of a typical prominence-presentational type:

8 (21) and (22) may be modified by replacing GA with WA forming WA-clefts. Please see Sunakawa (2005) pp.114-118 and pp.129-131 for an explanation of the differences between GA and WA-clefts.

(22) *Ningyo o meguru shinwa ya denshō wa, sekaijū itaru tokoro ni*
mermaids about mythology and folklore TOP world all over LOC
nokosarete iru ga, naka demo mottomo yūmeina no ga,
have been left but particularly most famous NOM SUBJ
Girisha-shinwa no Sērēn darō.
Greek mythology of Siren copula

(*Estaminet* Dec 1991)

'Mythology and folklore involving mermaids are found all over the world but the particularly famous one may be the Siren in Greek mythology.'

Some have argued that these types of GA-clefts should be considered as topic-less sentences (Shinya 1994, Noda 1996). However, these GA-clefts are not uttered without any presuppositions and therefore must be based on the context of the discourse and previous utterances. For example, looking at the sky and saying: *Ame ga futte kita yo* (It has started raining!) may be quite plausible but: *futte kita no ga ame da yo* (What has started falling is rain!) is not. It is because, unlike *genshōbun* (phenomenon descriptive sentences) that describe an incident directly as presented, these are explanatory type sentences that describe the situation arising from the presupposition based on the preceding contexts and utterances.

The presuppositions of sentences of the prominence-presentational GA-clefts are never presented explicitly as topics but implicitly contained within that which may be called *jōkyō-indai* (situationally inferred topic). Situationally inferred topics are by definition topics that are hidden but if we were to describe them in words, they may take the form as shown in the square brackets below:

(23) *Watashitachi wa satsuei o akiramete, sono ba o tachisarō to shita.*
we TOP shooting OBJ gave up that place OBJ be about to leave
[Soshite nani ga okotta ka to ieba]
and what SUBJ happened Q if say
Sono toki, mori no oku kara arawareta no ga, nanto
that time forest of depth from appeared NOM SUBJ my goodness
ippiki no ōkina kuma datta.
one big bear copula-PAST
'We gave up shooting and were about to leave the place. [And if we were to talk about what happened next] Just then, what appeared from the depths of the forest was, would you believe, a big bear.'

(24) *Ningyo o meguru shinwa ya denshō wa, sekaijū itaru tokoro ni*
mermaids about mythology and folklore TOP world all over LOC
nokosarete iru ga,
have been left but
[Sorera no shinwa ya denshō ni tsuite ieba]
these mythologies and folklores about if say
naka demo mottomo yūmeina no ga, Girisha-shinwa no Sērēn darō.
particularly most famous NOM SUBJ Greek mythology of Siren copula
(*Estaminet* Dec 1991)

'Mythology and folklore involving mermaids are found all over the world but [if we were to talk about such mythologies and folklore] the particularly famous one may be the Siren in Greek mythology.'

As it is necessary for the listener to hear the situationally inferred topic in order to understand sentences like these, more inference by the listener is required. Consequently, the listener has to devote more energy on processing the sentence and as a result, the listener is forced into making a conscious effort to focus on the meaning of the sentence. This is how a meaning that is more than a simple proposition, i.e., the specific referent is marked and emphasized, is deciphered by the listener.[9]

It must be noted that in these types of cleft sentences, certain parts of the sentences are often marked and emphasized by adverbs or conjunctions, as in the case of examples such as *Mottomo yūmeina no ga* (the most famous one is) and *Mazu kangaerareru no ga* (For a start it may be considered) where the underlined part indicates their markedness (Amano, 1996). When certain aspects of subjects are marked this way, the listener pays more attention to the referent indicated by the predicate and as a result, the referent becomes the topic, and this topic retains its prominence in the subsequent discourse.

Hetzron (1971) argues that moving a specific element from its usual position to another results in selective 'presentative function.' The GA-clefts under investigation here increase the level of prominence of the referent of the predicate noun by not only moving the specific element to a sentence final position but also by introducing *jōkyō-indai* (situationally inferred topic) that requires extra effort on the part of the listener in deciphering the sentence, or by emphasizing certain parts of the sentences by adverbs or conjunctions, the referent of the predicate noun is made more prominent.

From the above, it is possible to postulate that prominence-presentational GA-clefts perform the role of taking the information given in the preceding discourse and then presenting the referent that becomes the topic in the subsequent discourse in such a way that leaves a distinct impression upon the listener.

9 Please see Sunakawa (2005:118) for further discussion.

5 Topic development in discourse

In this section, focus-presentational functions of WA-clefts and the topic development of WA-clefts and GA-clefts are examined in order to explain the grammatical behaviours of both types of cleft sentences.

5.1 The focus-presentational function of WA-clefts

As discussed earlier, the main function of the structure of WA-clefts is focus presentation. The parameters of focus-presentational function, i.e., the function that determines the value of variable X and gives 'X is Y', are not restricted to those that seek specific referents by asking 'What is X?' or "Which one is X?' but include types that seek wide ranging answers such as causes, reasons and succession of events by posing questions such as 'Why X?' and 'X happens after what has happened?' and so on. WA-clefts can, therefore, seek not only the specific referent but also various different types of information, and answers to such questions are given in the focused predicate of the sentence. Consequently, not only nouns but various other expressions including subordinate clauses can form predicates of WA-clefts.

5.2 Topic development of WA-clefts

First let us examine cases of predicates being nouns. The following example has a noun as its predicate.

(25) *Watashi to Amerika o musubitsuketa no wa, chichi de aru.*
I and America OBJ linked NOM TOP father copula
Kare wa Waseda no Seikei o sotsugyō-go,…
he TOP Waseda of Politics and Economics OBJ graduated-after
(…description of father continues)

(*Bungei Shunjū* Jan 1993)

'Who linked me up with America was my father. Having graduated from *Waseda* majoring in Politics and Economics, he…'

In this sentence, the predicate noun is a specific person, *chichi* (father). It is nominative in relation to the verb *musubitsuketa* (linked) within the subject clause, and plays the grammatical function of subject to the verb. With regard to semantic roles and grammatical functions of nouns, Givón (1995: 46) presents a topic hierarchy in the discourse.

(26) Topic hierarchy in case roles:
  a. Semantic roles:
     Agent > Dative > Object > Locative > Instructive > Others
  b. Grammatical roles:
     Subject > Direct Object > Indirect Object

Following this hierarchy, *chichi* (father) in (25) is an agent, and as it is a subject of the verb in the clause, it is placed high in the topic hierarchy, and it is therefore expected that it will continue to be referenced in the subsequent discourse. Indeed, in this example *chichi* (father) is referred to as the topic in the subsequent discourse.

However, in a similar situation, there are some instances where the referent of the predicate noun is not carried on in subsequent discourse:

(27) *Tsui ni, to iuka, hatashite to iuka, waga mura no sūpā no tentō ni*
finally or at last shall we say our village of supermarket of shop LOC
doresu o kita kyūri ga tōjō shita. Saibai shite iru no wa,
dress OBJ wear cucumbers SUBJ appeared growing NOM TOP
tonari no machi no seinen de aru. / Doresu to itte mo, kifujin no
neighbouring town of young man copula dress say though lady of
yakaifuku to itta hade na mono de wa nai. Ga, sore demo,
evening dresses say ostentatious thing copula-NEG but it though
nanttatte kyūri de aru
whatever cucumbers copula

(Soichi Yamashita *Mura ni fuku kaze, Shincho Bunko*, 1989)

'Finally, or shall we say, at last, dressed cucumbers appeared at the counter of our village supermarket. The person growing them is a young man from the neighbouring town. / Though I call it a dress, it is not one of those ostentatious evening dresses a lady may wear. But, still, whatever one might call it, it is a cucumber.'

In respect to *tonari no machi no seinen* (a young man from the neighbouring town) being a person and the agent of the verb *saibai shite iru* (be growing) in the subject clause, this example is similar to (25). However the referent is not conveyed further and finishes within the current discourse. As seen in this example, in the case of WA-clefts, even those referents that are higher in the topic hierarchy are not always conveyed further as the topic of the subsequent discourse.

Furthermore, in the case of WA-clefts, it is not so rare that a referent lower in the topic hierarchy appears in the predicate, as the following example shows:

(28) *Sono hito ga norikonde kita no wa, tashika* <u>*Shiogama*</u>
that person SUBJ came on board NOM TOP surely *Shiogama*
datta to omou.
copula-PAST think

(*TrainVert* June 2003)

'Where the person came on board was <u>*Shiogama*</u>, I'm almost sure.'

The semantic role of *Shiogama* in relation to the verb *norikonde kita* (came on board) in the subject clause is a 'place' and the grammatical role is neither subject nor object. Because of this, it is positioned low in the topic hierarchy. In this example, what is carried on to the subsequent discourse is not *Shiogama*, but the person who came on board.

As shown above, predicate nouns of WA-clefts are not always carried on to the subsequent discourse, and it is not rare that the referent indicated by the noun is dropped immediately after it is introduced in the discourse. Also to be noted is that nouns lower in the topic hierarchy can appear in predicates.

Next, let us consider cases of predicates being subordinate clauses. Subordinate clauses represent events and attributions as well as the relationships between these subordinate clauses and the events and attributions of main clauses. When a certain concept is discussed as the topic of the discourse, it is more likely that the concrete referent will be conveyed further in the subsequent discourse rather than the abstract concept such as events, attributions and relationships (Chafe 1994: 67). Because of this, it is postulated that what a subordinate clause refers to may never become the topic of the subsequent discourse.

As predicted, among 21 examples where subordinate clauses form predicates, there was no sentence where the propositional content expressed in the subordinate clauses was conveyed further in the subsequent discourse, and in all examples what was found was the content left within the current discourse. An example of this is given below:

(29) <u>*Kaiga ni tsuyoku hikareru yō ni natta no wa, Hotta-tōdori no otomo de*</u>
paintings to strongly attracted became NOM TOP President Hotta's attendant
<u>*yoku tenrankai o mi ni itta sei deshō.*</u> *Chichi ga yūzen shokunin,*
often exhibitions OBJ see went reason copula father SUBJ *Yuzen* artisan
sobo ga makieshi data to iu kankyō mo atta ka
grandmother SUBJ *Makie* artist copula-PAST say surroundings too existed Q
to omoimasu. Saikin demo, yoku e no tenrankai ni wa ikimasu.
think recently even often paintings of exhibitions to TOP go
Senetsu desu ga, yōga ni kanshite wa hanbun puro o
presumptuous though western painting concerning TOP semi professional OBJ

jinin shite imasu.
acknowledge myself

(*Be-Common* Dec 1991)

'The reason why I became very interested in paintings may be because I used to accompany President Hotta to exhibitions very often. My father was a *Yuzen* artisan and my grandmother was a *Makie* artist and such a family background may have influenced me, too. Even now, I often go to exhibitions. Though I say so myself, I consider myself to be a semi-pro as far as western painting is concerned.'

From the information in the subordinate clause *Hotta tōdori no otomo de yoku tenrankai o mi ni itta sei* (Due to having accompanied President Hotta to exhibitions very often), only *tenrankai* (exhibition) is carried on in the subsequent discourse but the rest of the information is left within the current discourse. In addition to the fact that the meaning such as events, attributions, relationships presented by the subordinate clause is abstract, the high concentration of the information afforded by such subordinate clauses in comparison with nouns may also be the reason why the former cannot become the topic in subsequent discourse.

From these observations, it becomes clear that functions of WA-clefts are not primarily the introduction of the topic yet they perform other functions.[10]

5.3 Topic development of GA-clefts

As discussed in Section 4, there are two types of functions, namely the focus-presentational and prominence-presentational functions in GA-clefts. Out of 95 GA-cleft examples collected by the author, there were only five focus-presentational type sentences and the remaining 90 were prominent-presentational types. In case of all these 90 examples, the referent of predicate nouns was conveyed further in subsequent discourse. For example, in (30) below, *Kokuren Nihon seifu daihyōbu* (UN Japanese Government representatives) and in (31), *Zaōdō no gongyō* (divine services at the *Zaō* Temple) are conveyed further to the subsequent discourse.

(30) *Kono Kokuren ni taishi Nihon seifu o daihyō suru no ga*
this UN facing Japanese government OBJ represent NOM SUBJ
Kokuren-Nihon-seifu-daihyōbu de aru ga, watashi wa soko de
UN Japanese Government Representatives copula but I TOP there LOC
1988-nen made no 3-nen kan, zaimu-tantō-ittō-shokikan
1988 until 3 years period the first secretary responsible for financial affairs

10 Please see Sunakawa (2005: 112-131) for discussion on other functions of WA-clefts.

to shite kinmu shita.
as worked

(Japan Essayist Club *Haha no Shashin, Bungei Shunjū* 1994)

'Facing the United Nations, what represents the Japanese government is the UN Japanese Government Representatives, and this is where for three years until 1988, as the first secretary responsible for financial affairs, I worked.'

(31) *Soshite Yoshino e haitte ippaku shite, osusume na no ga*
and Yoshino to entered stayed overnight recommend NOM SUBJ
sōchō 6-ji goro ni hajimaru Zaōdō no gongyō.
early morning 6 o'clock around start *Zaō* Temple of service
Horagai to taiko to okyō to sore wa subarashii desu.
conch horns and drums and sutra chanting that TOP wonderful copula

(*Katei Gahō* Jul 1991)

'And getting to Yoshino and spending a night, what is recommended is the divine service at the *Zaō* Temple that starts in the early morning around 6 o'clock. Conch horns, drums, and sutra chanting; it's wonderful.'

The above observation indicates that prominence-presentational GA-clefts are used to introduce new referents, the referents that are introduced to the discourse for the first time and will be conveyed further to the subsequent discourse. Therefore, the referents of predicate nouns of prominence-presentational GA-clefts are 'New Topics' of the discourse.

On the other hand, the referents of predicate nouns of focus-presentational GA-clefts are the ones that follow from the previous discourse. Therefore, they are 'Old Topics' of the discourse.

Though there are such differences, in the case of both types, the referent of the predicate noun represents the topic of the discourse.

The grammatical characteristic of 'only nouns but not subordinate clauses being used for the predicate of GA-clefts' is attributed to the fact that predicates are the position that represents the topic of discourse that has been derived from the previous discourse, or is to be conveyed to the subsequent discourse. As discussed earlier, what is conveyed further as the topic of discourse is neither an event, an attribute or a relationship but the referent. Because of this, subordinate clauses are not used in predicates of GA-clefts. Instead, nouns are used in predicates.

Furthermore, this also explains the other characteristic of GA-clefts, where the tendency of the predicate nouns not accompanying case particles is particularly prominent in these constructions. In case of WA-clefts, predicate nouns do not always become the topic of discourse. On the other hand, predicate nouns in GA-clefts are always the

topic of discourse. This difference leads to the discrepancy in the way each type associates itself to the position within the topic hierarchy. In other words, the predicate nouns in GA-clefts that represent the discourse topic tend to associate themselves with higher ranking cases such as nominative and accusative in comparison with predicate nouns of WA-clefts that do not necessarily represent the discourse topic. Grammatical cases such as nominative and accusative are, unlike semantic cases such as instrumental and ablative, either cannot or do not normally take case particles when they are incorporated in the predicate of the cleft sentences. As a result, there are very few examples of GA-cleft predicate nouns taking case particles.

6 Skewed patterns of cleft sentences

As discussed earlier, WA-clefts are manifested in two types: "…WA + Noun + *da*" and "…WA + Subordinate clause + *da*" but in case of GA-clefts, in the examples found nouns were always contained but never subordinate clauses. Both types tend not to take case particles with their predicate nouns but the tendency is more apparent in the case of GA-clefts. In other words, GA-cleft distribution is skewed and most examples are found in the form of "…GA + Noun + *da*" and not in the form of "…GA + Noun + Case particle + da", and never in the form of "…GA + Subordinate clause + *da*." It is claimed in this paper that this phenomenon is not based on the restrictions at the sentence level but merely the patterns favoured by speakers in discourse depending on the differences of the functions of the two cleft sentence types. Based on the arguments thus far, the claims made in this paper and justifications for them are summarized below.

For WA-clefts having a focus-presentational function, it is possible for them to present a variety of information in the predicate, which is the focus. Because of this, not only nouns but subordinate clauses can take the predicate position.

GA-clefts have both focus-presentational and prominence-presentational functions. In case of GA-clefts with focus-presentational function, predicate nouns represent 'Old Topics' that have been carried forward from the previous discourse. On the other hand, in the case of prominence-presentational type of GA-clefts, predicate nouns represent 'New Topics' that are carried forward to the subsequent discourse. In summary, GA-clefts, whether they are focus-presentational or prominence-presentational, the referent of the predicate is always the discourse topic. Referents of nouns that represent relatively simple and concrete concepts such as inanimate objects or people are easier to convey as the topic of discourse, but abstract concepts such as events, attributes and relationships are not. Because of this, the predicates of GA-clefts are always nouns, and subordinate clauses are never used.

As for predicate nouns that present a topic, there is a tendency to use for nominative or accusative cases that are high in the topic hierarchy. Because nominative or accusative cases are clearly decipherable in terms of semantic roles, GA-clefts seldom accompany case particles in their predicate nouns.

7 Conclusions

Sentence patterns that are used frequently in discourse are gradually fixed and automated and enter higher-level grammar with distinctive regulatory power. Ohori (2004) points out that it is possible for grammaticalization to be found not only at the morphological or lexical level but also at the sentence level. The clefts examined in this paper are sentences that gradually form specific patterns based on functions performed in the discourse. These sentences may be considered as examples of day to day usage in discourse which encourage the development of certain patterns, making them fixed and automated, representing an example of sentence level grammaticalization.

Literature

Amano, M. (1996) Kōkō-shōten-bun no Meishi-jutsugo-bun: WA to GA no Kōsatsu no Kiten (Noun Predicate Sentences of Focus-final Type: A Starting Point of the Research on WA and GA). *Wakō Daigaku Jinbungakubu Kiyō* 30: 1-10.

Chafe, W. L. (1994) *Discourse, Consciousness, and Time.* Chicago: University of Chicago Press.

Givón, T. (1995) *Functionalizm and Grammar*. Amsterdam/Philadelphia: John Benjamins.

Hetzron, R. (1971) Presentative Function and Presentative Movement. *Studies in African Linguistics, Supplement*. 2: 79-105.

Kumamoto, C. (1989) Nichi-Eigo no Bunretsubun ni tsuite (On Cleft Sentences of Japanese and English). *Saga Daigaku Eibungaku Kenkyū* 17: 11-34.

Noda, H. (1996) *WA to GA (WA and GA)*. Tokyo: Kurosio Publishers.

Ohori, T. (2004) Bunpōka no Hirogari to Mondaiten (Expansion and Problems of Grammaticalization). *Gengo* 33(4): 26-35.

Prince, E. F. (1986) On the Syntactic Marking of Presupposed Open Propositions. *Chicago Linguistic Society.* 22: 233-259.

Shinya, T. (1994) Imikōzō kara Mita Heijobun Bunrui no Kokoromi (An Essay on the Classification of Declarative Sentences from the Perspective of Semantic Structure). *Tokyo Gaikokugo Daigaku Nihongo Gakka Nenpō.* 15: 1-14.

Sunakawa, Y. (1995) Nihongo ni okeru Bunretsubun no Kinō to Gojun no Genri (Japanese Cleft Sentences Concerning their Functions and Word Order). In: Nitta, Y. (ed.) *Fukubun no Kenkyū, jō* (*Research on Complex Sentences, part 1)*: 353-388, Tokyo: Kurosio Publishers.
Sunakawa, Y. (2005) *Bunpō to Danwa no Setten (Linkage between Grammar and Discourse)*. Tokyo: Kurosio Publishers.

要旨（Abstract in Japanese）

「日本語の分裂文の文法と談話における機能について」

砂川有里子（筑波大学）

分裂文にはハ分裂文「〜のは〜だ」とガ分裂文「〜のが〜だ」の2種がある。これらは述語に従属節を用いることが出来るかどうか、あるいは、述語名詞が格助詞を伴うことが出来るかどうかという点で異なった振る舞いを見せる。

本稿は、分裂文に見られる以上の相違に着目し、文レベルの制約だと思われていた現象が、談話において好まれて用いられる「型」の現れにすぎないものであることを述べる。

本稿の主張は、以下の2点である。

① 分裂文は焦点提示機能と特立提示機能という2種の談話機能を持つ。
② ハ分裂文とガ分裂文の文法的な振る舞いの異なりは両分裂文の談話機能によって説明できる。

II

Corpus-based research on discourse variety and lexis

5 Modal expressions and verbal interaction type: Suppositional adverbs as discriminators of Japanese corpora according to oral and written discourse varieties[1]

Andrej BEKEŠ
University of Ljubljana

Abstract
Modal expressions can be conceived as speaker's/writer's signals for a particular linguistic exchange or as a trace of such a linguistic exchange. Thus, modal expressions are the prime candidate as discriminators of the verbal interaction type, and consequently, of discourse type. In Japanese, the typical means to express modality are sentence-final modality expressions and modal adverbs. Because modal adverbs are easier to identify than sentence-final modal expressions, this chapter will examine the possibility of using a subset, i.e. suppositional adverbs, as discriminators of the Japanese oral and written discourse type. Several public and closed corpora belonging to different genres have been analysed here. Cluster analysis showed that distribution of suppositional adverbs in the examined corpora varied according to the discourse type. The outlier was a corpus of Elementary school *kokugo* textbooks, due to its heterogeneity of genres. The differences within similar corpora (Cluster of Informal conversations, Formal interviews, and CSJ lectures) mainly follow the axis of formality. On the other hand, in the more loosely correlated cluster of Diet speeches and Science textbooks there was a clear distinction between the stress on rhetoric in the former vs. stress on the precision of argument in the latter.

Keywords: verbal interaction type, discourse type, genre, suppositional adverbs, cluster analysis

0 Introduction

Text and discourse. Texts, either written, transcribed or recorded conversation or monologue, Aristotle's *ergon*, are not something static but traces of verbal interaction, of discourse: Aristotle's *energeia*, evolving in a particular social context with particular goals (cf. Coseriu 1973). Halliday (1978, 1991) describes the context of a situation, where communication takes place within the triplet of a *field,* i.e. 'subject matter and social action', *tenor,*

1 An earlier version of this chapter was presented at the International Conference: "Context-based Spoken Japanese Language" at the University Bordeaux Montaigne, April 4th-5th, 2014.

i.e. 'participants' relationship', and *mode,* i.e. 'medium and participation'. Modality in its broadest sense is concerned with the type of linguistic exchange in a particular situation. Thus, modal expressions can be conceived as a speaker's/writer's signals for the nature of a particular linguistic exchange or as a trace of such a linguistic exchange. This view is in line with Bakhtin's (1981) observation regarding the dialogic nature of text, which is, needless to say, more obvious in spoken interaction than in written.

Modality. The study of modality and related phenomena in Japanese has predominantly focused on the linguistic means which express modal meanings and the meaning of linguistic means, employed to express modality, the scope of investigation typically being a single utterance. With Jespersen as a starting point, Minami (1974, 1993) used Jespersen's terms *modus* and *dictum*. Teramura (1982), partly to avoid the misinterpretations based on traditional approaches, used *koto* (what is actually happening) and *mūdo* (from English *mood*). More recent approaches (cf. Nitta 1989, Masuoka 1991, Narrog 2009) prefer to view utterance as consisting of two parts: proposition (*meidai*) and modality (*modariti*), where modality reflects speaker's 'subjectivity'.

Modality, especially in planned linguistic exchanges such as written communication, tends to be formally expressed in a single utterance. Nonetheless, since modality expressions are not there just to fulfil some syntactic role, it is necessary to go beyond a single utterance in order to adequately understand modality related phenomena, as has already been pointed out by Teramura (1984:278-290).

In Japanese and in many other SOV languages, modality is typically expressed by the modal elements attached to the predicate (cf. Minami 1974). In addition, this can also be expressed by modal adverbs. As Minami (ibid., cf. also Narrog 2009), has shown, this is intimately connected with the layered nature of Japanese sentences. Co-occurrence of both modal adverb and sentence-final modality form, such as in (1) below, results in bracket structures (cf. Bekeš 2008c).

(1)	***tabun***	*kiku*	*-n*	***-darō…***	(from NUCC)
	probably	hear	*focus particle*	**probably**	(...she is probably going to hear...)

As has already been mentioned, taking verbal interaction / linguistic exchange as a starting point, modality is a signal/trace of the kind of verbal interaction taking place in a particular context of communication. Thus, we can expect that different verbal interactions in different contexts may have a different distribution of modal expressions, depending on the kind of linguistic exchange taking place. Moreover, the reverse is also possible, i.e. that similar types of distribution of modal expressions suggest similar types of verbal exchange.

Goal of this study. The goal of this study is to explore the possibility of using distribution of modal expressions as an analytic tool or indicator for discriminating

different types of verbal interactions represented by a particular text or groups of texts. Depending on the outcome, its usefulness as an indicator of genre will be assessed. To achieve this goal the focus will be on modal adverbs, in particular, on their subset of suppositional adverbs. Because of their formal simplicity, ease of identification and, as it will be shown, intricate relationship with different types of verbal interaction, suppositional adverbs seem to be most expedient for the purpose.

1 Discourse - text - genre - register - corpus

In empirical language research, two basic orientations can be distinguished. One type of research is predominantly concerned with clarifying "what is taking place in a particular linguistic exchange", i.e. the *ichigo-ichie* (once in a lifetime) type of approach. This is typical of conversation analysis and discourse analysis approaches, where the focus on the micro analysis of individual linguistic exchanges is predominant.

On the other hand, the so-called 'corpus approach' is more concerned with what is now called 'large data', focusing on the regularities in a mass of discourses produced within a definite period. This approach is typical of corpus lexicography, corpus based and corpus driven language studies and related research. As mentioned earlier, this present study uses a corpus approach. [2]

1.1 Text type, genre and register

It is expected that different communicative situations engender different goals and thus different types of texts. In situations with generically similar goals, similar solutions regarding the application of linguistic means for the achievement of a given goal emerge, resulting in notions of 'text type', 'genre' and 'register' (cf. Halliday 1991, Trosborg 1997). Definitions of what 'text type', 'genre' and 'register' actually are, and how they are related to each other, differ among authors and schools. This study follows Halliday (1991) and Hasan (2009). Figure 1 below shows the complex relationship between the context (including context of situation) and various aspects of language, i.e. system and instance. On the basis of this, register can be conceived as a toolbox of linguistic means, belonging to a language system, for achieving specific goals in a given (type of) context or situation. Correspondingly, text type is associated with situation type, different situations requiring different text types.

2 It is interesting to compare these research orientations with what historian Fernand Braudel (1958) called the opposition between "histoire événementielle", a non-structural history of events, and the history of "longue durée", i.e. history covering a longer period of time - which has the potential to highlight emergent structures in historical processes. As in linguistics, both views of historical processes are mutually interdependent.

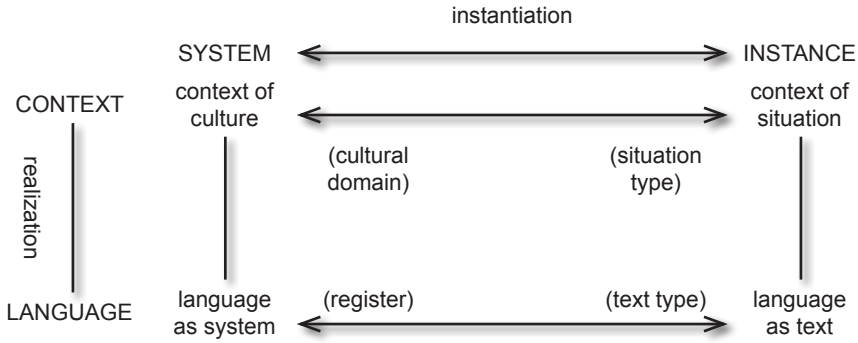


Figure 1. *Language and context: system and instance* (Hasan 2009 after Halliday 1991)

A more detailed relation between language/text, register, genre, and the context of situation is shown in Figure 2 below. Here, the term genre is used instead of the term text type as seen in Figure 1, as it is a more precise term referring to the generic purposes of communication. A type of situation defining generic purpose is also considered what motivates the choice of linguistic means for a particular occasion, i.e. the register. Register is thus closely associated with the features of the context of situation, i.e. with field, tenor and mode.

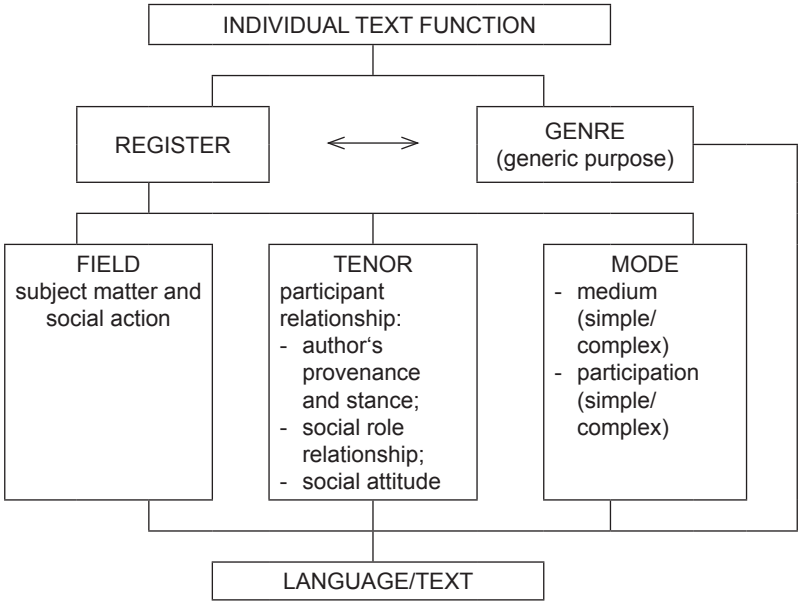


Figure 2. *Relation between register, genre and the context of situation* (cf. House 1997, Halliday 1991)

1.2 Modal adverbs in Japanese

As mentioned earlier, modal adverbs are one of the common means used to express modality in Japanese. Alongside sentence-final modal elements, attached to the predicate stem, they can be conceived as an important signal/trace of the kind of linguistic interaction taking place. Kudō's (2000) work, widely covering different types of adverbs, is a seminal work on this topic. Here, suppositional modal adverbs are of particular importance, existing in a kind of semantically based pseudo agreement with sentence-final modal expressions, as illustrated by example (1) above.

2 Suppositional adverbs as indicators of the type of linguistic exchange

2.0 Data

In this section, I will attempt to further investigate the potential of Japanese suppositional adverbs as indicators of the type of linguistic exchange.

The data used for this research consist of several corpora, some of which are openly accessible and some, due to copyright issues, are used only for research and are not openly accessible. These corpora were chosen because they consist of texts, representing widely different types of linguistic exchange, i.e.:

(i) Informal conversations: Ohso, Mieko (2003) *Meidai kaiwa kōpasu* (Nagoya University Conversation Corpus, NUCC). About 100 informal conversations between familiar participants. Text file size 3.5 MB.
(ii) Formal interviews: Oikawa Terufumi (1998) *Jinbunkagaku to konpyūtā DATABASE vol. 1*, Sōgō kenkyū daigaku. Transcribed interviews with 50 Japanese native speakers. Unfamiliar participants, systematic status differences. Text file size 0.82 MB.
(iii) Academic conference lectures: NINJAL CSJ corpus core data (Himawari 1.02 CSJ core data, 2005). Core part of the NINJAL Contemporary spoken Japanese Corpus.
(iv) Japanese *Kokkai* (the Diet) speeches: NINJAL Chūnagon (1990-1999), Transcribed Diet speeches are part of this corpus.
(v) Science textbooks: Nishina Kikuko (unpublished material prepared at the Tokyo Institute of Technology) 16 University Science textbooks for basic subjects - corpus of Science texts.
(vi) Elementary school *kokugo* (national language) textbooks: Nishina Kikuko's (unpublished material prepared at the Tokyo Institute of Technology) 60 elementary school *kokugo* textbooks. Corpus of elementary school texts.

Characteristics of the contextual features of texts from the corpora above are displayed in Table 1 below.

'Undemanding' refers to the linguistic interaction regarding common topics, where specialised knowledge of a particular field is not required. On the other hand, 'demanding' refers to linguistic interaction requiring specialised knowledge of a particular field.

Table 1. *Contextual features of texts in the analysed corpora*

Contextual features		**Informal conversations**	**Formal interviews**	**CSJ lectures**	***Kokkai* speeches**	**Science textbooks**	**Elem. s. texts**
FIELD	*subject matter*	private topics, undemanding	private topics, undemanding	public topics, science, demanding	public topics, state policy issues, demanding	public topics, science, demanding	fiction, medium demanding
	social action	private conversation	public interview, onesided conversation	public, one-sided imparting of knowledge, expository	public, antagonistic debate, argumentative	public, one-sided imparting of knowledge, expository	public, literary text,- classroom activity,
TENOR	*social role relationship*	family, friends, colleagues, little or no seniority, high familiarity	interviewer vs(+seniority) interviewee (-seniority), high seniority, low familiarity	lecturer vs. audience (students), high seniority, low familiarity	peer vs peer (Diet members), low familiarity	public, writer-reader, low familiarity	public, writer-reader
	social attitude	informal, familiar	formal	formal	formal	formal	formal
MODE	*medium*	spoken	spoken	spoken	spoken	written	written
	participation	direct	direct	direct	indirect	indirect	indirect

2.1 Suppositional adverbs

The following is a list of the most common suppositional adverbs, based on Kudō (2000).

Table 2. *Suppositional adverbs* (cf. Kudō 2000)

Modality type	No.	Adverb
NECESSITY **(NEC)**	*1*	*kitto* (surely)
	2	*kanarazu* (certainly)
	3	*zettai ni* (absolutely)
EXPECTATION **(EXP)**	*4*	*osoraku* (probably)
	5	*tabun* (likely)
	6	*sazo* (surely)
	7	*ōkata* (probably)
	8	*taitei* (usually)
	9	*taigai* (mostly)
CONJECTURE **(CON)**	*10*	*dōyara* (somehow)
	11	*yohodo/yoppodo* (quite)
POSSIBILITY **(POSS)**	*12*	*moshika-suruto/-shitara/-sureba* (maybe)
	13	*hyottosuruto /hyottoshitara* (possibly)
	14	*kotoniyoreba /kotoniyoruto* (possibly)
	15	*angai* (fairly)
	16	*kanarazushimo …(nai)* ([not] necessarily)

According to the modality type they encode, these adverbs are divided into four discrete groups; from the greatest to the least probability: "necessity" (abbreviated as NEC in subsequent tables and graphs), "expectation" (EXP), "conjecture" (CON), and "possibility" (POSS).

2.2 Distribution of suppositional adverbs in different types of corpora

Following Kudō (2000), the distribution of suppositional adverbs, either alone or in association with sentence final modal expressions, has been further examined in various corpora by Srdanović et al. (2008a,b), Srdanović et al. (2009), etc. The purpose was to verify Kudō's observations, evaluate the importance of suppositional adverb - sentence-final modal expression associations for inclusion in Japanese language curricula, and genre characteristics. The results of present study are presented in Table 3, where some differences between the examined corpora are obvious already on the basis of raw frequencies (**frq** in the table) and relative frequencies (expressed as percentages) of suppositional adverbs in chosen corpora selected; for example adverbs such *tabun* (likely) being frequent in informal conversations, yet less likely to appear in other types of linguistic exchange.

Table 3. *Distribution of suppositional adverbs*

Modality type	No.	Adverb	Informal convers.			Formal interviews			CSJ lectures			*Kokkai* (Diet) speeches			Science textbooks			Elem. school texts		
			frq	%	dens.	frq	%	dens.	frq	%	dens.	frq	%	dens.	frq	%	dens.	frq	%	dens.
NEC	1	*kitto*	172	18	**184**	11	9	**52**	0	0	0	9	2	**73**	0	0	**0**	129	44	**168**
	2	*kanarazu*	42	4	**45**	19	16	**90**	0	0	0	84	15	**684**	56	46	**86**	42	14	**55**
	3	*zettai ni*	35	4	37	1	1	5	0	0	0	37	6	301	5	4	8	12	4	16
EXP	4	*osoraku*	9	1	**10**	8	7	**38**	1	33	41	194	33	**1579**	2	2	3	8	3	10
	5	*tabun*	580	61	**621**	50	40	**236**	2	67	82	72	12	**586**	1	1	2	10	3	13
	6	*sazo*	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	15	5	20
	7	*ōkata*	0	0	**0**	1	1	**5**	0	0	0	7	1	**57**	0	0	**0**	2	1	3
	8	*taitei*	5	1	5	10	8	47	0	0	0	4	1	33	0	0	0	28	9	36
	9	*taigai*	11	1	12	1	1	5	0	0	0	2	0	16	0	0	0	0	0	0
CON	10	*dōyara*	11	1	**12**	0	0	**0**	0	0	0	5	1	**41**	0	0	0	11	4	14
	11	*yohodo/ yoppodo*	34	4	**36**	2	2	**9**	0	0	0	0	0	**0**	0	0	0	17	6	22
POSS	12	*moshika-suruto/ -shitara/ -sureba*	33	3	35	12	10	57	0	0	0	10	2	81	0	0	0	12	4	16
	13	*hyottosuruto /hyottoshi-tara*	4	0	4	1	1	5	0	0	0	3	1	24	0	0	0	3	1	4
	14	*kotoniyor-eba /kotoni-yoruto*	0	0	0	0	0	0	0	0	0	0	0	0	1	1	2	0	0	0
	15	*angai*	14	1	**15**	1	1	**5**	0	0	0	4	1	**33**	1	1	**2**	5	2	7
	16	*kanarazu shimo … (nai)*	8	1	**9**	4	3	**19**	0	0	0	148	25	**1205**	54	45	**83**	1	0	1

The table above and the results of research are further elaborated by adding the density of expressions in each corpus, expressed as the observed frequency per 1.000.000 morphemes (**dens.** in the table). Both, relative frequency and density, make a comparison between different corpora or texts more direct and meaningful than raw frequencies since larger corpora include more tokens than smaller corpora. Corpora can be similar in their distribution of relative frequencies of suppositional adverbs, though their density may differ, or they can be similar in the distribution of the density of suppositional adverbs.

For example, the density of *tabun* per million morphemes is high, not only in informal conversations, but, which must be noted, also in *Kokkai* (the Diet) speeches where raw frequencies were considerably lower due to the smaller size of the corpus.

In order to gain a clearer picture of the general properties of the modal adverb distribution of corpora, the agglomeration of data across the modality type is shown in Table 4.

Table 4a. *Agglomerated distribution of suppositional/evidential adverbs across the modality type (density, i.e. frq/1.000.000 morphemes)*

Modality type	**Informal convers.**	**Formal interviews**	**CSJ lectures**	***Kokkai* (Diet) speeches**	**Science textbooks**	**Elem. school *kokugo* texts**
NEC	266	147	0	1058	94	239
EXP	649	331	123	2271	5	82
CON	48	9	0	41	0	36
POSS	63	86	0	1343	87	28

In Table 4a, different shades of grey reflect the relative prevalence or relative absence of a certain modality type in different corpora, based on density. In informal conversations, suppositional adverbs are employed quite frequently, with prevailing types of suppositional modality being NECESSITY and EXPECTATION. On the other hand, in *Kokkai* (the Diet) speeches, suppositional adverbs are used markedly more frequently. In relation to NECESSITY, the use of EXPECTATION is comparable to that used in Informal conversations, yet in addition to this the very frequent use of POSSIBILITY has also been observed. CSJ academic lectures, as they are a monologues in public/formal settings, display, as expected, a lower frequency of the use of suppositional adverbs, with only one prevailing type of suppositional modality, i.e. EXPECTATION. Science textbooks and Elementary school *kokugo* (national language) textbooks also display complementary distributions. As expected, more formal science textbooks display a markedly lower frequency of modal adverbs than Elementary school *kokugo* textbooks, actually, the second lowest frequency among all corpora. Similar conclusions, though not identical, can be reached also on the basis relative frequencies as shown in Table 4b.

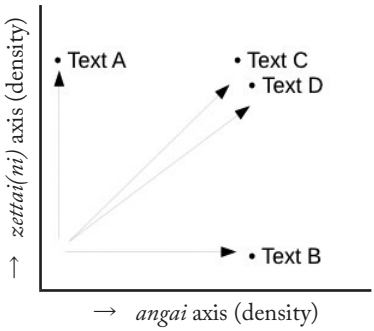
Table 4b. *Agglomerated distribution of suppositional/evidential adverbs across the modality type (relative frequencies)*

Modality type	**Informal convers.**	**Formal interviews**	**CSJ lectures**	***Kokkai* (Diet) speeches**	**Science textbooks**	**Elem. school *kokugo* texts**
NEC	26	26	0	23	50	62
EXP	64	57	100	47	3	21
CON	5	2	0	1	0	10
POSS	5	15	0	29	47	7

These observations have led to the conclusion, that when language is considered, it is difficult to think about language in general, as the use of linguistic means greatly varies across different types of linguistic exchange. The picture that emerges - though based here only on the limited case of modal adverbs - is that, within a language, there might be different sets of "sub-languages", used for different purposes in different linguistic exchanges. That is, a specific set of the means of expression is mobilised in order to achieve the goals for each specific linguistic exchange; thus, similar exchanges imply similar solutions, hence the similarities of suppositional adverb distributions within each homogeneous corpus. This brings us back to the notion of register in Section 1 (cf. Figures 1 and 2).

2.3 Similarities of text types: clustering of different types of text data based on WORD-SPACE

The notion of WORD-SPACE (cf. Widdows 2004) will be illustrated here by example of a 2-dimensional modality space, spanned between two suppositional adverbs. Let us make the assumption that there are three hypothetical texts: text A asserting a fact of which the speaker/writer is absolutely sure, with only the adverb *zettai(ni)*) used (absolutely); text B, in which the speaker/writer expresses facts as possibilities, with only the adverb *angai* (fairly) used; and text C, which is of a mixed nature, in which the speaker/writer both asserts clear facts - using *zettai(ni)* while stating other facts as possibilities, using *angai* in equal proportions. There is also text D, in which *angai* and *zettai(ni)* are used in proportions as in text C. If the frequencies of adverbs are provided as relative frequencies or densities, this situation can be visualised as texts represented by points in a WORD-SPACE, an abstract space of texts or sets of texts, spanned over word vectors; in our case the words are *zettai(ni)* and *angai*. The position of a text in WORD-SPACE depends on the relative frequency or density of each word vector, as shown in Figure 3 below. Text A includes only *zettai(ni)* and is represented by a point in the upper left corner, text B, including only the word vector *angai* is represented by a point in the lower right corner and texts C and D, including both *zettai(ni) and angai* are points in similar proportions located in the upper right corner, between texts A and B.



Text A: only *zettai(ni)*

Text B: only *angai*

Text C: *zettai(ni)* and *angai* in similar proportion

Text D: similar to C

Figure 3. *Text/corpus data in a toy modality adverb word space*

This idea can also be applied to a more complex abstract space, including more spanning word vectors (in our case 16 suppositional adverbs, as shown in Table 1). Texts with similar proportions of defining words (expressed as word vectors), will be represented as points in close proximity to each other and texts with different proportions as points at a greater distance from each other. The similarity between texts and of course organised groups of texts, such as corpora, can be expressed as the distance between the points representing them or as the angle between text (or corpus) vectors (i.e., lines connecting text points with their origin). On the basis of this, cluster analysis can be systematically applied to group texts or sets of texts.

"Distance" and "angle" as measures of similarity, which highlight the different aspects of similarity and the resulting groupings, are not necessarily the same, as can be seen in Figures 4 and 5.

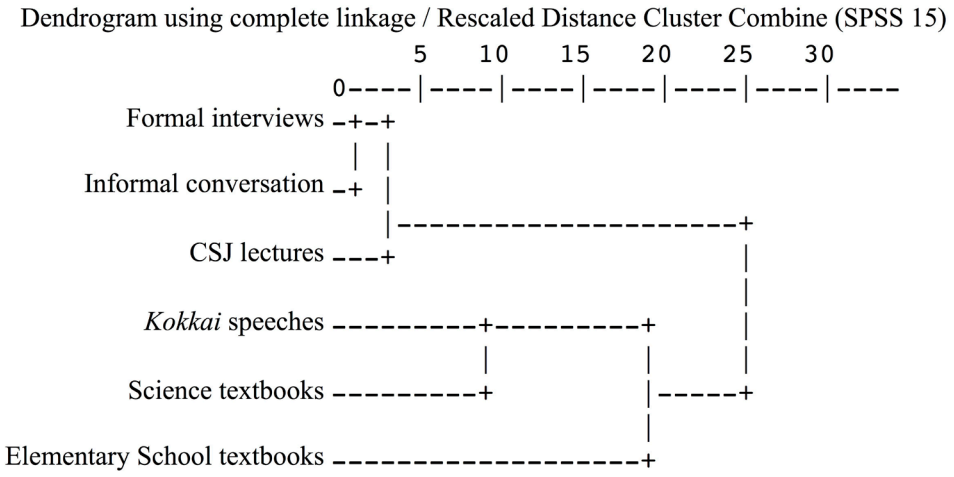


Figure 4. *Comparison of corpora 1 – Euclidean "distance"*

In Figure 4, the *Kokkai* (the Diet) speeches are shown as a corpus classified in a group of its own, in contrast to the group consisting of all other corpora. With distance as a measure, extremely high normalised frequencies of adverbs of all modality types come out much more strongly than the similarity of their proportions. This results in a dendrogram, where the *Kokkai* (the Diet) speeches are an outlier, at a great distance from the origin and represented with the least similarity to other corpora.

On the other hand, Informal conversations are also a kind of outlier in this cluster analysis, since they also display very high normalised frequency of *tabun*, with the second highest *kitto* being more than three times less frequent.

This similarity of texts and, consequentially, the corpora based on a comparison of the relative frequencies of suppositional adverbs provides us with rather different result, as shown in Figure 5 below.

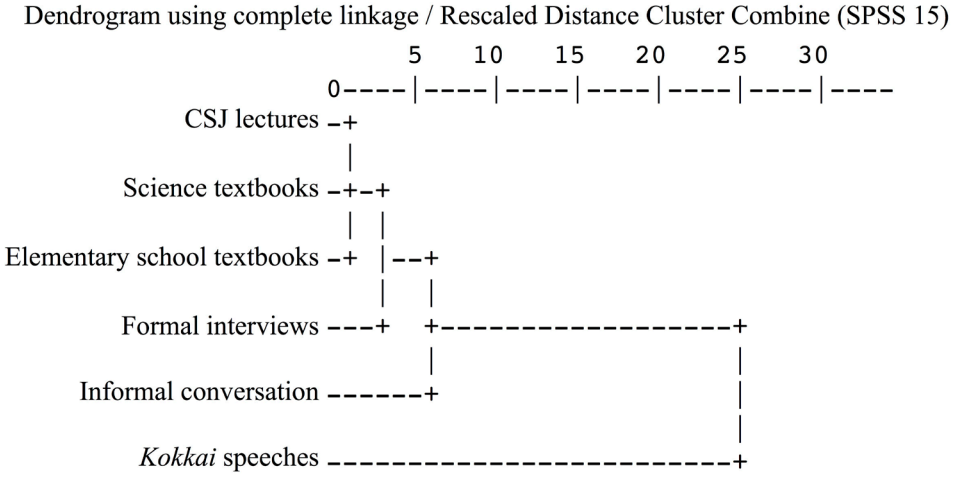


Figure 5. *Comparison of corpora 2 – Similarities in relative frequencies – cosine distance*

Here the spoken corpora, i.e. Formal Interviews, Informal conversations and CSJ lectures, form quite a distinct group. As can be seen in Table 4b, a similarity based on the proportional predominance of adverbs belonging to the "expected" (*tabun* in particular), are predominant in all three groups over all other modalities. In all the other three corpora *osoraku* is predominant in the category "expected". Thus, *Kokkai* (the Diet) speeches and Science textbooks, while differing in proportion of the "expected", share similar proportions in the category of "necessity" and "possibility". In the second cluster, the outlier consists of a corpus of Elementary school textbooks, where only the proportion of "necessity" is clearly greater than all other modality categories.

Identical results can be obtained without "adverb based word space" interpretation, simply by comparing the correlations (i.e., statistical similarities) between the

distribution of adverbs in corpora, as shown in Table 5 below. The degree of correlation between different corpora is shown with different shades of grey. As in the cluster analysis in Figure 4, Informal conversations and Formal interviews are most closely correlated (0.91), followed by CSJ lectures (0.84 with Informal conversations and 0.83 with Formal interviews). The corpora in this cluster are thus internally closely correlated with each other.

Table 5. *Correlations of adverb distribution between corpora*

	Informal convers.	**Formal interviews**	**CSJ lectures**	***Kokkai* speeches**	**Science textbooks**	**Elem. s. texts**
Informal convers.	---					
Formal interviews	0.91	---				
CSJ lectures	0.84	0.83	---			
***Kokkai* speeches**	0.13	0.29	0.48	---		
Science textbooks	-0.09	0.13	-0.12	0.55	---	
Elem. s. texts	0.20	0.17	-0.11	-0.10	0.03	---

The other, more loosely correlated cluster is that of the *Kokkai* speeches and Science textbooks (0.55), corresponding to the cluster in the dendrogram in Figure 4. This cluster is loosely correlated to the CSJ lectures in first cluster (0.48). The outlier here is an elementary school *kokugo* texts corpus, which is understandable, since it is an agglomeration of various literary and expository texts and thus has a much lesser internal consistency than other corpora.

Based on the observations of results of both types of cluster analysis, as represented in Figures 4 and 5, Table 5, *Kokkai* (the Diet) speeches display the most complex use, employing all categories of suppositional modality adverbs and, thus, suppositional modality. Indeed, as can be seen in Table 3, *Kokkai* (the Diet) speeches have the highest density (freq./1.000.000 morphemes) of 3 out of 4 suppositional modality categories. It is interesting to note here that, while formal varieties of suppositional adverbs are employed more frequently, informal counterparts also display relatively high densities when compared to other corpora. Indeed, a comparison with Informal conversations shows that *Kokkai* (the Diet) speeches - though public and formal, still display a high frequency of *tabun*, an adverb that is associated with informal contexts of use.

The correlation between the uses of certain forms along the formal-informal axis is obvious, when Informal conversations and Formal interviews are compared.

Table 6a. *Formal – informal axis* – osoraku *and* tabun

Adverb	**Informal conversations**	**Formal interviews**
osoraku (probably)	10	38
tabun (likely)	621	236

Chi-square test: $p < 0.0005$

Table 6b. *Formal – informal axis* – kitto *and* kanarazu

Adverb	**Informal conversations**	**Formal interviews**
kitto (surely)	184	52
kanarazu (certainly)	45	90

Chi-square test: $p < 0.005$

As Tables 6a and 6b show, there is a clear correlation between *osoraku* as a formal choice and *tabun* as an informal one. The same is true also for the pair *kitto* and *kanarazu*: *kitto* prevails in informal conversations and *kanarazu* in the formal ones.

Table 6c. *Proportions of formal vs. informal use (*osoraku *vs.* tabun*)*

	***Kokkai* (Diet) speeches**	**Style**
osoraku (probably)	1579	formal
tabun (likely)	586	informal

Variations between the formal and informal, observed in *Kokkai* (the Diet) speeches (Table 6c above) may be attributed to differences between individual speakers. Yet other corpora with a high degree of formality, such as Science textbooks, do not show this variation to the same extent. What may be the case here is that, though mediated as recorded texts, *Kokkai* (the Diet) speeches still reflect the dynamics of spoken communication. Speakers may be resorting to strategies, such as "characterisation" (cf. Coseriu 1973, *linguaggi d'imitazione*, i.e. imitation languages, Sadanobu 2011, *kyarakuta*), switching between formal and informal registers, in order to achieve certain goals such as attracting greater sympathy with the audience and so on.

This possibility of variation also explains the similarities and dissimilarities between *Kokkai* (the Diet) speeches and Science textbooks. What is similar in the linguistic exchanges of both of these corpora is logical argumentation. On the other hand, where scientific discourse requires a precision of argument, discourse in the Diet preferentially seeks rhetoric effects. Skilful hopping between registers may be a way to achieve this goal.

All corpora, except Elementary school *kokugo* textbooks, seem to be quite homogeneous and are close to what we intuitively understand as genre, i.e. generic purpose (see Figure 2 above). On the other hand, Elementary school *kokugo* textbooks consist of a variety of different texts, fragments of literary works, explanations and comments, and provide a picture of a much dispersed use of various suppositional modalities. Because of this lack of homogeneity, what can be stated here is that this corpus does not represent a genre in the narrow sense of the word, such as academic prose etc.

3 Conclusions

The distribution of suppositional adverbs in the corpora examined here reveals a clearly profiled, distinctive use of suppositional adverbs, according to the type of text. The corpora were internally rather homogeneous, their text type being close to what is also intuitively understood as a *genre.* The outlier *was* the corpus of Elementary school *kokugo* textbooks, because of its heterogeneous nature as a collection of various fragments of literary texts and expository prose, not correlating well with any other corpus.

Differences within similar corpora (Cluster of Informal conversations, Formal interviews, and CSJ lectures are mainly along the axis of formality, i.e. the employment of informal vs. formal items (*tabun* vs. *osoraku*, *kitto* vs. *kanarazu* etc.). On the other hand, in the more loosely correlated cluster of *Kokkai* speeches and Science textbooks there was clear distinction between a stress on rhetoric (use of informal *tabun* when expedient), in the former vs. a stress on precision of argument in the latter.

The results above quite clearly show that elements such as modal adverbs, and the suppositional adverbs among them, intrinsically connected to the type of verbal interaction, are indeed efficient discriminators of various text types. Yet further refinement is a possibility, by using other stable and transparent linguistic elements, also intrinsically connected to various types of verbal interaction. What comes to mind are:

(i) linguistic elements directly related to types of verbal interaction such as other types of modal adverbs, utterance-final particles, etc.
(ii) linguistic elements directly related to types of discourse structuring such as connectives (*setsuzoku hyōgen*), thematising elements (*teidai hyōgen*), etc…

On the other hand, elements whose function is predominantly referential, even though they may be relevant in different text types or genres, such as the nouns *hito, renchū* and *yatsu,* all referring to a person or human being, are not suitable, as their use depends not only on the type of verbal interaction involved but also on a type of extra-linguistic reality. Since there may be similar text types relating to different kinds of extra-linguistic reality, it can be expected that referential elements are less suitable as efficient discriminators of text types.

Literature

Bakhtin, M. M. (1986) The Problem of Speech Genres, in *Speech Genres and Other Late Essays*. Translated by Vern W. McGee. Austin: University of Texas Press.

Bakhtin, M. M. (1981) Discourse in the novel. In Michael Holquist (ed.) *The dialogic imagination* (translation by Caryl Emerson and Michael Holquist, of *Voprosy literatury i estetiki*). Austin: University of Texas Press, Slavic series; No. 1, pp. 259-422 (translation of "Slovo v romane" 1934–1935, also available in original at http://www.gumer.info/bibliotek_Buks/Literat/bahtin/slov_rom.php - last accessed August 2, 2015).

Bekeš, A. (2008a) Suppositional adverbs as indicators of discourse genre. In: Petrovčič, Mateja (ed.) *Current issues in East Asian languages*, (*Asian and African studies*, ISSN 1408-5429, vol. 12, issue 3). Ljubljana: Department of Asian and African Studies, Faculty of Arts, pp. 5-16.

Bekeš, A. (2008b) *Text and Boundary: A Sideways Glance at Textual Phenomena in Japanese*. Ljubljana: Ljubljana University Press.

Braudel, F. (1958) Histoire et Sciences sociales: La longue durée. *Annales. Économies, Sociétés, Civilisations*. 13-4, pp. 725-753.

Coseriu, E. (1973) *Lezioni di linguistica generale*. Editore Boringhieri.

De Beaugrande, R. A., Dressler, W. (1981), *Introduction to Text Linguistics*. London: Longman.

Halliday, M. A. K. (1978) *Language as Social Semiotic: The Social Interpretation of Language and Meaning*. London: Edward Arnold.

Halliday, M. A. K. (1985) *An introduction to functional grammar*. London: Edward Arnold.

Halliday, M. A. K. (1990) Some grammatical problems in scientific English. *Annual Review of Applied Linguistics* 6:13-37.

Halliday, M. A. K. (1991) Language and context: system and instance (in 2002-2007 *Collected Works of M.A.K. Halliday* vol.9, London: Bloomsbury).

Halliday, M. A. K. (1994) *An Introduction to Functional Grammar*. Arnold.

Halliday, M. A. K. & Ruqaiya Hasan (1976) *Cohesion in Englis*h. London: Longman.

Hasan, R. (2009) The place of context in a systemic functional model. In M.A.K. Halliday and JonathanJ. Webster (eds.) *Continuum Companion to Systemic Functional Linguistics*, pp. 166-189. New York: Continuum.

House, J. (1997). *A Model for Translation Quality Assessment*. Second edition. Tübingen: Günter Narr.

Kudō, H. (2000) Fukushi to bun no chinjutsu no taipu (Adverbs and the type of sentence-final modality) In Nitta Yoshio and Masuoka Takashi (eds.) *Nihongo no bunpō 3* – modariti (Japanese grammar 3: modality) pp. 161-234. Tokyo: Iwanami shoten.

Narrog, H. (2009) *Modality in Japanese: The layered structure of the clause and hierarchies of functional categories*. Amsterdam: John Benjamins.

Nitta, Y. & Takashi M. (1989) *Nihongo no modariti* (Japanese modality) Tokyo: Kuroshio Shuppan.

Sadanobu, T. (2011) Kyarakuta wa bunpō o doko made kaeru ka? (How much can a 'character' change the grammar) In Kinsui Satoshi (ed.) *Yakuwari-go kenkyū no hatten* (development of the role language) Tōkyō: Kuroshio shuppan, pp. 17 - 26.

Sano, M. (2008) Text-classification system for large scale balanced corpus: a systemic functional approach. *Proceedings of the 2008 General Meeting of the MEXT grant-in-aid for Scientific Research Priority Area Program "Japanese Corpus" (in Japanese)*.

Srdanović Erjavec, I., Bekeš, A., Nishina, K. (2007) Cluster analysis of suppositional adverbs and clause-final modality. *Asian and African Studies*, University of Ljubljana Faculty of arts, pp. 21-31.

Srdanović Erjavec, I., Bekeš, A., Nishina, K. (2008)a Distant collocations between suppositional adverbs and clause-final modality forms in Japanese language corpora. In: *Large-scale knowledge resources: construction and application* ; Third international conference on large-scale knowledge resources, LKR 2008, Tokyo, Japan, March 3-5, 2008; proceedings, (Lecture Notes in Computer Science, 4938), (Lecture Notes in Artificial Intelligence). Berlin: Springer, pp. 252-266.

Srdanović Erjavec, I., Bekeš, A., Nishina, K. (2008)b Adverbs and clause-final modality collocations in various corpora. In: *Tokutei ryōiki kenkyū 'Nihongo kōpasu': Heisei 19 nendo kōkai wākushoppu (kenkyūseika hokokukai) Yokōshū : 15.3.-16. 3. Tokyo: Monbukagakusho kagakukenkyūhi tokuteiryōiki kenkyū 'Nihongo kōpasu' Sōkatsu ban*, pp. 223-230.

Srdanović I., Hodošček B., Bekeš A., Nishina K. (2009) Uebukōpasu to kensaku shisutemu o riyō shita suiryō fukushi to modariti keishiki no enkaku kyōki chūshutsu to nihongo kyōiku e no ōyō (Extracting distant collocations of adverbs and modality forms using a web corpus and a query system). *Shizen gengo shori* (Journal of Natural Language Processing) Vol. 16 No. 4.

Srdanović I., Bekeš A., Nishina K. (2009) Kōpasu ni motozuita goi shirabasu sakusei ni mukete -- suiryō-teki fukushi to bunmatsu modariti no kyōki o chūshin ni shite (Towards the corpus based vocabulary syllabus with special reference to suppositional modality adverbs and sentence-final modality), *Nihongo kyōiku* (Japanese language education), 142 : 69-79.

Teramura, H. (1982) *Nihongo no shintakusu to imi I*. (Japanese syntax and meaning I) Tokyo: Kurosio.

Teramura, H. (1984) *Nihongo no shintakusu to imi II* (Japanese syntax and meaning 2). Tokyo: Kuroshio Shuppan.

Trosborg, A. (1997) Text Typology: Register, Genre and Text Type. In Anna, Trosborg (ed.) *Text typology and translation.* Amsterdam: John Benjamins Publishing Company, pp.3-22.

Widdows, D. (2004) *Geometry and meaning*. Stanford: CSLI.

Corpora

- NUCC Nagoya University Conversation Corpus
  Ohso, Mieko (2003) *Meidai kaiwa kōpasu* (Nagoya University Conversation Corpus).
  About 100 informal conversations between familiar participants. Text file size 3.5 MB.
- Formal interviews
  Oikawa Terufumi (1998) *Jinbunkagaku to konpyūtā* DATABASE vol. 1, Sōgō kenkyū daigaku. Transcribed interviews by 50 Japanese native speakers. Unfamiliar participants, systematic status differences. Text file size 0.82 MB.
- CSJ conference lectures
  NINJAL CSJ corpus core data (Himawari 1.02 CSJ core data, 2005)
- *Kokkai* (Diet) speeches
  NINJAL Chūnagon (1990-1999)
- Science textbooks
  Nishina Kikuko (unpublished) 16 University Science textbooks for basic subjects - corpus of Science texts
- Elementary school *kokugo* textbooks
  Nishina Kikuko (unpublished) 60 elementary school *kokugo* textbooks. Corpus of elementary school Japanese language texts.

概要 (Abstract in Japanese)

「モダリティ表現と言語による相互作用―推量副詞による日本語の会話・文章コーパスのタイプを識別する一可能性について―」

アンドレイ・ベケシュ（リュブリャナ大学）

談話におけるモダリティ表現は、話し手・書き手から聞き手・読み手への、進行中の言葉によるやり取りがどのようなものなのかを示す合図である。そのため、談話分析においてモダリティ表現は、言葉によるやり取りのタイプ、言い換えれば、談話およびテキストのタイプ、を識別する重要な手掛かりと言える。

日本語の場合、モダリティを表す手段として、文末モダリティ表現と、いわゆる陳述副詞がある。陳述副詞は、文末モダリティ表現に比べ同定が容易である。本稿では、日本語の談話およびテクストのタイプを識別するために、陳述副詞のサブタイプである推量的副詞を用いる可能性を検討する。分析の対象として、それぞれが異なったジャンルに属する複数のコーパスを用いる。

分析の結果、推量的副詞の分布がコーパスの種類によって異なることが明らかになった。特に異なったのは小学校の国語教科書コーパスの推量的副詞の分布である。

一方、推量的副詞の分布がある程度類似している親しい仲間たちの会話コーパス、フォーマルなインタビューのコーパスおよびCSJ講義コーパスにおける相違点は、主として、フォーマル・インフォーマルの軸に沿って選択されたレジスタの違いとして現れている。

さらに、国会演説コーパスと自然科学の教科書コーパスのような、若干しか類似していないコーパスの相違点は、主として、レトリックにおける違いおよび論法の精度における違いに現れるようである。

6 Adjectives on –i in Japanese language corpora: Distribution, patterns and lexical constraints[1]

Irena SRDANOVIĆ
Juraj Dobrila University of Pula

Abstract
This paper explores Japanese i-adjectives using empirical methods of corpus linguistics and employing state-of-art language resources and a lexical profiling tool. Firstly, this research presents resources that have been used in the analysis and explains their relevance and characteristics. These resources are used to examine the distribution of i-adjectives in the large-scale corpora of contemporary written Japanese, which clarifies which i-adjectives predominate in the overall usage of i-adjectives and how some i-adjectives and adjectival suffixes are more productive than others. Next, this research analyses the distribution of the patterns of the three major roles of adjectives and shows the different tendencies in the usage of their roles and patterns among adjectives. The research focuses on i-adjectives in their attributive role preceding a modified noun and reveals their complexity of patterns and the need to further subcategorize the types of attributive role adjectives have. Furthermore, this study examines lexical constraints in the attributive role of i-adjectives, while discovering some adjectives with no or a rare attributive role.

Keywords: i-adjectives, corpora, distribution, patterns, constraints

1 Introduction

The computer-assisted systematic research of carefully collected large-scale authentic data confirms that lexical items retrieved for utterance constrain the syntactic structure that can be employed for their construction (cf. Schönefeld 1999: 138-9, Stefanowitsch and Gries 2003: 209-10). This empirically based confirmation is one of the main achievements of corpus linguistic research in the analysis of human spoken and written discourse and reminds us of the necessity of employing this methodology to further analyse particular languages and human language in general. Accordingly, the aim of this research is to explore the Japanese language patterns of i-adjectives, with special focus on their role as modifiers of nouns, using the

1 The previous version of the study "Distribution, semantic and syntactic profile of Japanese i-adjectives" was presented at the conference "XXVII es Journées de Linguistique d'Asie Orientale" in Paris, 2014.

empirical methods of corpus linguistics and employing the latest language resources and lexical profiling tools.

The first part of this paper examines the distribution of i-adjectives in present-day large scale written corpora, differentiating between very frequent and vary rare adjectives and their role in Japanese language productivity. This research touches upon the most productive adjectival suffixes in Japanese which form compound and derived adjectives.

The second part of the paper analyses the distribution of the patterns of the three major roles of adjectives (predicative, attributive and adverbial) as recognized in previous studies (cf. Suzuki 1972, Nishio 1972, Hashimoto and Aoyama 1992) and explores how tendencies of the usage of their roles differ among adjectives. The usage of large-scale corpora, BCCWJ and JpTenTen, and the user-friendly tool Sketch Engine (Kilgarriff et al. 2004, Srdanović et al. 2008, 2013) enable a more thorough exploration of the patterns in their use and some possible constraints. This study uses several adjective examples in order to explore in detail the attributive role of i-adjectives preceding a noun.

This paper is structured as follows: Section 2 explains the resources used in analysis, Section 3 presents the results of the analysis of the distribution of adjectives and their productivity, and Section 4 describes the analysis and results of different patterns and their constraints.

2 BCCWJ, JpTenTen and Sketch Engine: resources used in analysis

This section introduces two large-scale Japanese language corpora, BCCWJ and JpTenTen, and the state of the art corpus query system Sketch Engine. These corpora and the tool are used in this analysis.

Like the British National Corpus (BNC) and many other national corpora, the Balanced Corpus of Contemporary Written Japanese (BCCWJ) was compiled with the aim to be as balanced as possible and with a size of around 100 million words. As described by Maekawa et al. (2013) the originality of this corpus is in its sampling policy. The first two sub-corpora: Publication sub-corpus (PSC) and Library sub-corpus (LSC) used the technique of stratified random sampling. PSC represents the actual state of the publication in Japan and reflects the aspects of the "production" of written texts from books, magazines, and newspapers published between 2001 and 2005. LSC reflects the "circulation" aspect of written Japanese, covering a wide range of the books published between 1986 and 2005 and registered in more than 13 public libraries in the Tokyo metropolis. The third sub-corpus, the Special-purpose sub-corpus (SSC), included sampled data of some specific types of texts, such as public data distributed by the Japanese government, best-selling books, blog data etc.

Thus far there have been various approaches in Japanese word identification, since the Japanese language has no explicit word boundaries and it poses great challenges when identifying words for language processing and corpus analysis. The BCCWJ corpus project decided to employ a two-fold morphological annotation approach with short-unit word (SUW) and long-unit word (LUW) identification (Ogura et al. 2011) using the electronic dictionary UniDic and the morphological annotation tool MeCab. This new approach also covers various orthographic variations under the same units.

The following two examples show how Japanese words are identified using short- and long-unit word annotation. While some words such as particles and single-morpheme units are marked as one unit in both annotation systems (e.g. 私 *watashi* 'I', は *wa* '[theme particle]', で *de* '[place particle]'), some words are divided into more units in SUW annotation (e.g. 日本/語 *Nihon/go* 'Japanese language [Japanese/language], 勉強/し/て/いる *benkyō/shi/te/iru* 'study [study_N/suru_AuxV_base/suru_AuxV_renyō/iru_AuxV]]" and combined into one or several units in LUW annotation (e.g. 日本語 *Nihongo* 'Japanese language', 勉強/し/ている *benkyō/shi/teiru* 'study [study_N/suru_AuxV_base/teiru_AuxV]). The adjectives on *–i* are typically divided into more units in SUW annotation in case of compound and derived adjectives, such as 興味/深い *kyōmi/bukai* 'interesting [interest_N/deep_Ai]', and combined into one unit in LUW annotation 興味深い *kyōmibukai* 'interesting'. The ordinary simple adjectives are identified as one unit in both SUW and LUW annotation (e.g. 新しい *atarashii* 'new').

Example 1-1: Two-fold annotation system (short- and long-unit words)

SUW:	/私/は/プーラ/大学/で/日本/語/を/勉強/し/て/いる/。
	Watashi/wa/Puura/daigaku/de/Nihon/go/wo/benkyō/shi/te/iru/.
	'I study Japanese at Pula University.'
LUW:	/私/は/プーラ大学/で/日本語/を/勉強し/ている/。
	Watashi/wa/Puuradaigaku/de/Nihongo/wo/benkyō/shi/teiru/.
	'I study Japanese at Pula University.'

Example 1-2: Two-fold annotation system (short- and long-unit words) applied on i-adjectives

SUW:	新しい	興味/深い
	atarashii	*kyōmi/bukai*
LUW:	新しい	興味深い
	atarashii	*kyōmibukai*
	'new'	'interesting'

The next generation of compiled corpora, following the large-scale balanced national corpora such as BNC and BCCWJ, aimed at obtaining corpora much larger in

size by using various methodologies in order to collect data from the web and had also achieved a relatively good balance of the data (c.f. Baroni and Bernardini 2004; Baroni and Ueyama 2006; Baroni and Kilgarriff 2006; Sharoff 2006). The first web corpora were close to the national corpora size (c.f. Srdanović et al. (2008) as an example of such Japanese language web data) but a decade later super large-scale corpora were created. For example, the Corpus factory project (Kilgarriff et al. 2010) provided a number of such corpora, including the 10-billion-word Japanese language corpus JpTenTen (Pomikalek and Suchomel 2012; Srdanović et al. 2013), which we will use in the analysis. This corpus uses the same tools for word identification as BCCWJ: MeCab and UniDic with a two-fold annotation system (SUW and LUW).

Finally, we will use the corpus query and lexical profiling system Sketch Engine (Kilgarriff et al. 2004) to search through the two large-scale corpora. The advantages of this tool are in various search possibilities including the advanced functionality Word Sketches that provide a detailed summary of keyword patterns and collocations. In addition, when required, we use the corpus query systems Chunagon to search for data in BCCWJ sub-corpora.

3 Distribution of i-adjectives

3.1 SUW and LUW i-adjectives in BCCWJ

Joyce and Hodošček (2012) report an "approximately thirteen-fold increase in the number of LUW lemma types over the number of SUW lemma types" in BCCWJ. The study on SUW and LUW lemma types for i-adjectives in BCCWJ data (Srdanović 2013a) revealed that there are 761 adjectives annotated as SUW lemma types and 12585 adjectives annotated as LUW lemma types (Table 1). The analysis include general i-adjectives (annotated as Ai.g, jap. 形容詞-一般 *keiyōshi ippan*, such as *yoi*, *tanoshii*, *fukai*), bound adjectives (annotated as Ai.bnd, jap. 形容詞-非自立可能 *keiyōshi-hijiritsukanō,* such as *nai*, *yoi*) and adjectival suffixes (annotated as Suff.ai, jap., 接尾辞-形容詞的 *setsubiji-keiyōshiteki*, such as -rashii, -*ppoi*).[2] The results show that there is an approximately 15-fold increase in the number of LUW over the SUW lemma types for i-adjectives in this large-scale corpus. The large gap reminds us that much linguistic production is actually achieved through combinations of a limited number of elements, which is widely known as the phenomenon of language economy.

2 While bound adjectives do not appear as a category in LUW, there are still some adjectival suffixes not attached to any LUW item in the BCCWJ data.

Table 1. I-adjectives annotated as SUW and LUW lemma units in the corpus BCCWJ (based on UniDic + MeCab annotation)

Annotation schema	**Examples of annotated i-adjectives (general and bound adjectives)**	**No. of i-adjective lemma units in JpTenTen**
SUW	無い *nai* 'nonexistent, [indicates negation]', 良い *yoi* 'good', 美味しい *oishii* 'delicious', 楽しい *tanoshii* 'funny', らしい *rashii* 'such as', 易い *yasui* 'simple'	761
LUW	無い *nai* 'nonexistent, [indicates negation]', 良い *yoi* 'good', 美味しい *oishii* 'delicious', 楽しい *tanoshii* 'enjoyable', 人間らしい *ningenrashii* 'human', 色っぽい *iroppoi* 'amorous', 間違い無い *machigainai* 'certain', 分かり易い *wakariyasui* 'easy to understand'	12585

3.2 I-adjectives with high frequency and high coverage

As previously discussed in Srdanović (2013a), the study of i-adjectives annotated as LUW lemma units in BCCWJ reveals that highly frequent 25 (24+1) i-adjectives take up 62% (20%+42%) of the overall occurrences of i-adjectives, when calculating their tokens (Table 2). The first 127 i-adjectives on the frequency list take up almost 90% (62%+27%) of the overall occurrences of this part of speech category. We can see that the number of highly frequent adjectives is relatively small and that these adjectives constitute the highest coverage of adjective usage in BCCWJ. On the other hand, there is great number of very rare adjectives that appear only once or rarely. The rare adjectives that appear from 1 to 10 times cover around 87% of overall adjective usage, when calculating types. Figure 1 shows the usage of adjectives in relation to their frequency span in the corpus and the number of i-adjective lemmas that appear in the corpus in the specific frequency span.

It must be noted that highly frequent items that constitute the large amount of coverage in a language are more likely to be encountered by foreign language learners and therefore should be given priority in vocabulary learning. As Nation (2001) already proposed for English vocabulary learning, a basic 2000-3000 words cover around 70-80% of language usage and therefore need to be placed on the priority list for learners. This approach can be applied to other languages and is in line with the results of the distribution of i-adjectives, where the first group of adjectives takes up 62% of overall usage. The next group of 102 i-adjectives is also very widely used (27%) and, therefore, could be covered gradually after the first group of i-adjectives. All in all, both groups of i-adjectives cover 89% of i-adjective usage.

Table 2. The number of LUW i-adjective lemma tokens and types per frequency span in BCCWJ

Freq span of Ai	LUW Adj (token)	%
100000～	279682	20,02
10000～100000	591458	42,33
1000～10000	378934	27,12
100～1000	94210	6,74
10～100	33739	2,41
1～10	19122	1,37
Total	1397145	100

Freq span of Ai	LUW Adj (type)	%
100000～	1	0,01
10000～100000	24	0,19
1000～10000	102	0,81
100～1000	341	2,71
10～100	1064	8,45
1～10	11053	87,83
Total	12585	100

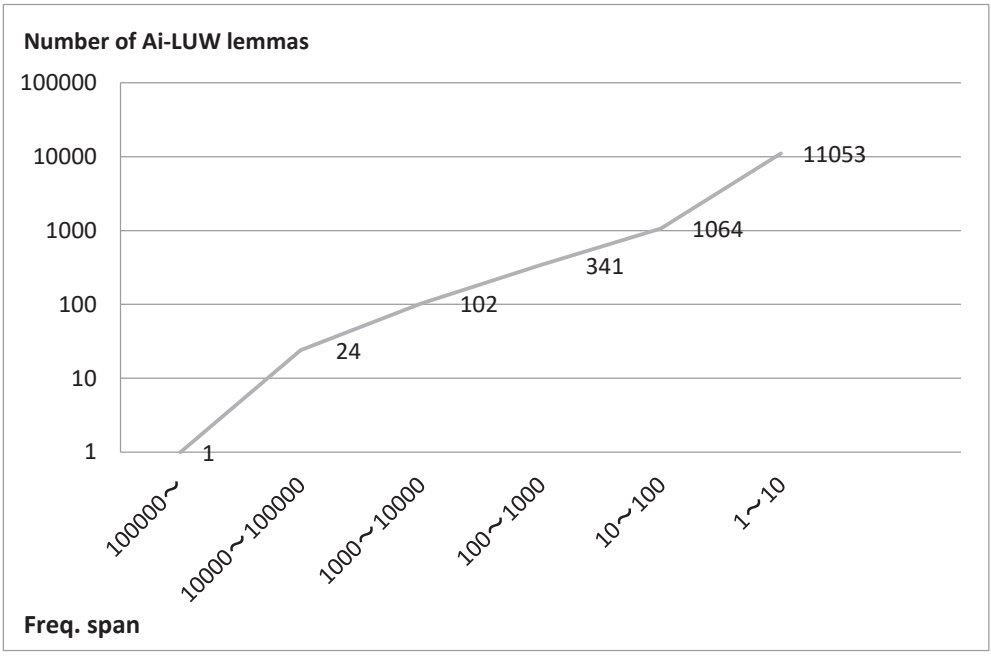


Figure 1. I-adjective LUW lemma appearance in BCCWJ based on frequency span (type-token ratio)

What follows is a list of the 25 most frequent i-adjectives (annotated as LUW lemma) that appear in BCCWJ.[3]

無い *nai* 'non-existent, [indicates negation]', 良い *yoi* 'good, [indicates permission]', 良い *yoi* 'good', 多い *ōi* 'many', 高い *takai* 'high', 大きい *ōkii* 'big', 悪い *warui* 'bad', 強い *tsuyoi* 'strong', 新しい *atarashii* 'new', 早い *hayai* 'fast', 長い *nagai* 'long', 少ない *sukunai* 'few', 旨い *umai* 'skillful', 若い *wakai* 'young', 凄い *sugoi* 'amazing', 小さい *chīsai* 'small', 難しい *muzukashii* 'difficult', 楽しい *tanoshii* 'joyful', 深い *fukai* 'deep', 美味しい *oishii* 'tasteful', 近い *chikai* 'close', 低い *hikui* 'low', 嬉しい *ureshii* 'happy', 面白い *omoshiroi* 'interesting', 広い *hiroi* 'wide'

3.3 Low frequency adjectives, suffixes and their productivity

Interestingly, there is a very high number of low frequency adjectives – more than 11000 adjectives that appear only once or up to ten times. Carefully observing their production in the corpus examples, we notice a large number of compound and derived adjectives that are formed using various suffixes, which indicates the capacity of i-adjectives for productivity and creativity in Japanese. Table 3 shows the list of all suffixes that are used to produce more than a hundred various i-adjectives in the corpus. The second line shows the number of different adjectives created using a specific suffix, where the frequency of each adjective is only one. The third line shows the total number of different adjectives that are created using a specific suffix. Finally, the fourth line shows the total number of all the appearances of adjectives with a specific suffix. The most productive suffixes are *-rashii*, *-yasui*, *-ppoi*, *-gatai/nikui.*[4] However, different kinds of productivity can be noticed: *-ppoi/-poi* seems to have the most tendencies for variety and creativity as the number of adjective that appear only once is quite high, in comparison to the number of different adjectives and all their appearances. *-rashii* and *–tsurai* have similar tendencies, so we might expect more derived adjectives in the case of these suffixes. On the other hand, *-nai, -yoi/ii, -fukai/bukai* have a limited number of different adjectives that are repeated quite often.

3 This list also includes the most frequent function words *nai* (無い) and *yoi/ii* (良い). Note that the word *tai* (たい) is annotated as an auxiliary verb. Also, note that there are some slight differences among the list of 25 of the most frequent i-adjectives in SUW and LUW annotation data – 欲しい *hoshii* 'to want [general i-adjective]' and 易い *yasui* 'easy; likely to …[general i-adjective and adjectival suffix]' appear higher in the frequency list in data annotated as SUW.

4 Refer to Srdanović (2013) for more details on the compounding of adjectives with their suffixes, most frequent compounds, i-adjectives with the suffix *-kusai*, and for a discussion on the implications of the results on Japanese language learners.

Table 3. Suffixes and their productivity in producing i-adjectives

Suffix	No of Ai (freq=1)	No of Ai (diff)	No of Ai (total)
らしい *rashii* 'seemingly/like'	2376	2995	15734
易い *yasui* 'easy'	1213	2316	22738
っぽい *ppoi* 'like'	1031	1326	5694
難い *nikui/gatai* 'hard'	939	1718	15158
無い *nai* 'non-existent [negation]'	458	632	18540
辛い *tsurai* 'painful/tough'	203	343	1650
良い *yoi* 'good'	194	245	5200
臭い *kusai* 'smelly/-ish'	181	292	3796
ぽい *poi* 'like'	158	185	238
深い *fukai/bukai* 'deep'	124	210	4705

When closely observing the results from the viewpoint of Japanese language education, it can be noticed that it would be very effective to provide learners with detailed information on the process of suffix productivity in order to provide support during the process of learning how to produce new, correct and meaningful adjective combinations, especially in cases of highly productive suffixes. On the other hand, restricted and lexicalized adjectives can be introduced and learned one by one or in meaningful groups in order to make learning maximally efficient.

4 I-adjective patterns and their constraints

This section explores the distribution of i-adjective patterns in Japanese language corpora and their constraints in the attributive role.

4.1 Distribution of patterns for i-adjectives: case of *takai*

This section takes the Japanese i-adjective *takai* 'high, tall, expensive' as an example and explores its most frequent patterns that appear in the corpus JpTenTen, using the Sketch Engine Word sketch tool. Figure 2 shows that among the recognized patterns the following appear most frequently:

1) *Takai* as a noun modifier, in its attributive role (*rentai-kei*) preceding a noun (高い +N, *takai* + N 'high + N', 18%), such as 高い山 *takai yama* 'a high mountain';
2) *Takai* as an adjectival predicate (*shūshi-kei*) (Nが高い[concl], *N ga takai* 'N is high', 18%), such as 効果が高い *kōka ga takai* 'the effect is high'; and

3) *Takai* in combination with some suffixes, such as 高さ *takasa* 'height', 高過ぎ *takasugi* 'too high' (18%).

In the annotated corpus data, another type of attributive role (*rentai-kei*) of the adjective *takai* appears as an adjectival predicate in adnominal clauses and is quite frequent:

4) Adjectival predicate in adnominal clause (Nの/が高いN, *N no/ga takai N* 'N with high N', 16%).

This attributive role of *takai* is both a noun modifier preceding a noun but also an adjectival predicate of a preceding relative clause, forming the construction *N ga/no Ai N*. For example, 完成度の高い作品 *kanseido no takai sakuhin* 'work with a high degree of perfection/completion', or 背が高い方 *se ga takai kata 'a tall person [lit. a person with a high back*]'. In this construction, the so called '*ga/no* conversion' (cf. Harada 1971) occurs with the case particles *ga* and *no*. Since this kind of usage is quite prominent in the case of the adjective *takai*, yet also appearing in the usage of some other adjectives (Srdanović 2013b), the author suggests that there is a need to differentiate it from a pure attributive role in a noun phrase, at least in corpus annotation data. This would be possible either by using a narrower tagset for grammatical roles (forming a new inflection tag 'attributive-predicative role') or by improving parsing (syntactic analysis) for these kinds of grammatical constructions. Also, this pattern with *ga/no* conversion needs to be differentiated from the pattern with the possessive *no* particle, where two attributives (nominal and adjectival) appear modifying the same noun (*N1noAiN2*, see the case of adjective *aoi* in Section 4.2).

Besides these usages, there are also:

5) Adnominal clause with adjectival predicate and omitted case particle or a compound of a noun and *takai* (10%), such as テンション高い *tenshon-takai* 'high tension'.
6) Adverbial form preceding a verb (*renyō-kei*) (高く+V, *takaku* + *V* 'V + high_adv)' (9%), such as 高くなる *takaku naru* 'to become expensive'、高く売る *takaku uru* 'to sell at a high price'. For adverbial forms, although considered one of main three forms in adjective usage, the results showed that, in the case of *takai*, it is not as prominent when compared to other forms.
7) Conjunctive form of *takai* used to link adjectives or clauses (*renyō-kei*) (Nが高く(て)[cont], *N ga takaku(te)[cont]* 'N is high and …' (8%).

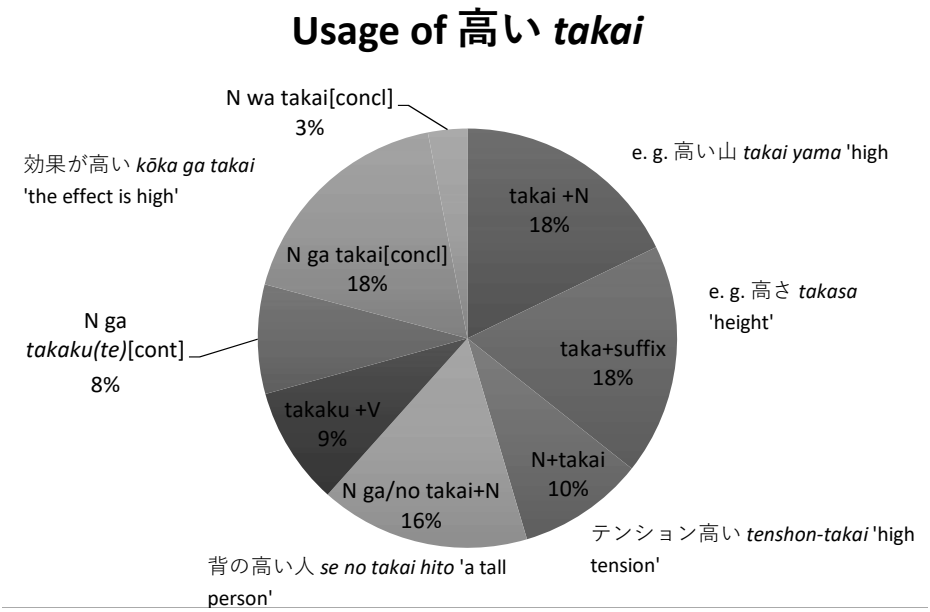


Figure 2. Usage of *takai* in the JpTenTen corpus

4.2 Adjectives in attributive form preceding a modified noun: a previous study revisited

In a previous study (Srdanović 2013b), the author explored the syntactic structure of the attributive role of an adjective preceding a noun (Ai_rentai + N) for a number of adjectives of high, medium and low frequency with an aim to gain better withdrawal results for Japanese adjectives in their attributive form from the SkE collocational functionality Word Sketches. A hundred random examples of *Ai_rentai +N* have been taken from the corpus for each adjective and analysed.

The research findings showed that different adjectives have different tendencies in forming the patterns. For example, the adjectives 寒い *samui* 'cold', 親しい *shitashii* 'close', 甘い *amai* 'sweet' have the tendency to appear mostly in simple noun phrases of an adjective modifying a noun (Ai + N) (90%, 93%, 86% respectively), such as 寒い季節 *samui kisetsu* 'cold+season = cold season'. The adjectives 多い *ooi* 'a lot of' and 高い *takai* 'high' show a very high tendency when forming adnominal clauses with an adjectival predicate. More specifically, there is a large number of *ooi* + *N* or *takai* + *N* cases (91% and 54% respectively) where the adjective has the role of a modifier and at the same time the role of an attributive predicate (for example, 雨の多い国 *ame no ōi kuni* 'rain+-CONV_no+lots of+country = a country with lots of rain, or 質の高いサービス *shitsu no takai sābisu* 'quality+CONV_no+high+service = a high-quality service'). Such cases

clearly differentiate from other observed patterns, such as simple *Ai + N* pattern, where an adjective has the pure role of a noun modifier, and therefore they deserve their own sub-classification. This is especially relevant in the domain of corpus annotation, but might also deserve to be reconsidered within the domain of Japanese language grammar.

Furthermore, there are patterns with two attributives (nominal and adjectival) modifying the same noun (*N1noAiN2*), that are more characteristic for adjectives such as青い *aoi* 'blue' or 甘辛い *amakarai* 'salty-sweet'. In this pattern, the particle *no* connects *N1* to *N2* and the i-adjective modifies *N2*, so that the order of modification is *N1no (Ai+N2)* (17% and 12% respectively). For example, 日本の青い空 *Nihon no aoi sora* 'Japan + POSS_no +blue+sky = Japanese blue sky'. This kind of pattern needs to be differentiated from adnominal clauses with an adjectival predicate, where the particle *no* appears as, the so called '*ga/no* conversion' (cf. Harada 1971), which we will explore further in Section 4.3.

There are also a few cases of adnominal clauses with adjectival predicate and omitted case particle or compounds of a noun and an adjective, also in the case of adjectives such as 多い *ōi* 'a lot of' and 高い *takai* 'high' (2% and 5%), e. g. 香り高いコーヒー *kaoritakai kōhī* 'aroma+high+coffee = aromatic coffee'.

The above results of the corpus-based study, including the results from Figure 2, show that there is a need to reconsider corpora annotation but also the traditional way of observing the syntax of adjectives and there usage in the Japanese language. The three major roles of adjectives: predicative, attributive and adverbial (cf. Suzuki 1972, Nishio 1972, Hashimoto and Aoyama 1992) need to be reconsidered for their subgroups and further explored based on the large-scale corpora available for contemporary Japanese to widen the understanding of their behaviour and the patterns of adjectives.

4.3 Exploring the pattern 'N no/ga takai N'

This section takes the patterns Nが高いN *N ga takai N* 'N with high N' and Nの高いN" (*N no takai N* 'N with high N') as an example and shows the results of the quantitative corpus-based analysis of these patterns. Table 4a and 4b display the reoccurrences of various combinations within the pattern. For example, "~度が高いこと/もの/作品" (… *do ga takai koto/mono/sakuhin* 'thing/work with high level of ...') in Table 4a, 質の高いサービス (*shitsu no takai sābisu* 'service of a high quality') in Table 4b, or 背の/が高い人・方 (*se no/ga takai hito/kata* 'a tall person [lit. a person with a high back]' both Tables 4a and 4b).

The *ga/no* conversion phenomenon, where the particle *ga* stands for 'Nominative Case' and *no* for 'Genitive Case', has been explored and explained in various studies (cf. Harada 1971, Tsujimura 1996: 264-266). These studies describe the possibilities of the conversion in relative clauses in Japanese and state that there is no apparent

Table 4a: Corpus-based analysis of the pattern Nが高いN *N ga takai N* 'N with high N')

~が高い~ ga takai	こと koto 'thing, [nominalizer]'	ため tame 'for, because'	もの mono 'thing'	人 hito 'person'	方 kata 'person'	気 ki 'mood, spirit'	場合 baai 'in case of'	ところ tokoro 'place, when'	そう sō 'it seems, it is said'	わけ wake 'reason, it is not so...'	時 toki 'times, when'	せい sei 'back'	事 koto 'thing'	分 bun 'part, point'	作品 sakuhin 'work'	理由 riyū 'reason'	傾向 keikō 'tendency'
度 do 'degree'				227	197	155	167	119	53	149	48	61	114	193	205		53
率 ritsu 'rate'			215	107	97	374	144	112	198	112	45	64	158	49		141	199
～度 do 'degree'	971	614	664														
～率 ritsu 'rate '	1409	578															
背 se 'back'	251	82	49	565	311					79		53		56			
レベル reberu 'level'	343	110	144	240	108	63	78	93		74			49				
人気 ninki 'popularity'	250	178	178						179						56	127	
効果 kōka 'effect'	567	172	269						110				52				
値段 nedan 'price'	316	93	219		66	81		52					55	88			
能力 nōryoku 'ability'	306	153	60	260	113					62							
リスク risuku 'risk'	494	180	104	102			72										
評価 hyōka 'evaluation'	189		152	69											124	93	
敷居 shikii 'treshhold'	61		175			145											
確率 kakuritsu 'probability'	256	98	75			119			99	71			50				
標高 hyōkō 'elevation'	61	258						170				102					
気温 kion 'temperature'	56	104									97	121					
方 hou, kata 'more, person'	159	50	60			139	96			56							46
湿度 shitsudo 'humidity'	63	117					58				113	86					
頻度 hindo 'frequency'	205	105	174	65	55		76										
価格 kakaku 'price'	225	107	100														

Table 4b: Corpus-based analysis of the pattern Nの高いN *N no takai* N 'N with high N')

~の高い~ no takai	もの mono 'thing'	人 hito 'person'	作品 sakuhin 'work'	方 kata 'person'	サービス sābisu 'service'	商品 shōhin 'product'	ところ tokoro 'place'	製品 seihin 'product'	情報 jōhō 'information'	選手 senshu 'player'	男 otoko 'man'	女性 josei 'woman'	医療 iryō 'healthcare'	場所 basho 'place'	靴 kutsu 'shoes'
度 do 'degree'		303		220	180	287	131	246	134	87		63	57	118	
質 shitsu 'quality'	727		447		1474	200		170	410	70			933		
背 se 'back'	283	1411		628						141	992	392			
レベル reberu 'level'	587	574	320	163	66		246			250		82			
人気 ninki 'popularity'	266		250		48	519									
～度 do 'degree'	1984		1303												
価値 kachi 'value'	858		73		365	357		323	167						
効果 kōka ''effect	489				47	154		85							
クオリティ kuoriti 'quality'	389		368		92	135		89							
率 ritsu 'rate'	268	53		57		180	102	66						54	
精度 seido 'accuracy'	254							62	185						
能力 nōryoku 'ability'	107	416		243						399					
標高 hyōkō 'elevation'							526							356	
意識 ishiki 'awareness'		521		302						52		161			
湿度 shitsudo 'humidity'							117							220	
リスク risuku 'risk'	150	182		64		57									
頻度 hindo 'frequency'	433														
評価 hyōka 'evaluation'	181	86	263			49				68					
身分 mibun 'status'	70	468		50								85			
ヒール hīru 'heel'															614
品質 hinshitsu 'quality'	182				64	111		118							
カロリー karorī 'calories'	359														

difference in meaning. Looking into the differences in occurrences of the particles *ga* and *no* in the specified patterns throughout the corpus, revealed that there are differences in the distribution between the use of *ga* and *no* in this kind of pattern and that some combinations of words in the pattern tend to appear more or only with one of the particles. The results also revealed that *ga* tends to be more in use in combination with functional words and in an abstract sense (e.g. こと *koto* 'thing, [nominalizer]', ため *tame* 'for, because'). On the other hand, *no* is used twice to three times more in this kind of pattern than *ga*, and tends to appear more in concrete cases (質の高いサービス *shitsu no takai sābisu* 'high level service', ～度の高い作品 *…do no takai sakuhin* 'work with a high degree of…'[5], ヒールの高い靴 *hīru no takai kutsu* 'shoes with high heels' etc.).

The details of the most frequent patterns are also presented in Table 5, where unmarked patterns appear frequently only with one of the case particles, *no* or *ga*, while marked patterns appear in both cases.[6] For example, ～度 -- 高いもの *…do -- takai mono* 'a thing with a high degree of ...' (*ga*: 664 / *no*: 1984) and 背 -- 高い人 *se -- takai hito* 'a tall person' (*ga*: 565 / *no*: 1411) appear with both case particles quite frequently, although we can notice almost three times more usage with *no* in these particular examples as well. There are plenty of examples where combinations have high tendencies and appear more often or only with one case particle. The particle *no* particularly tends to appear more than the particle *ga* within patterns where N1 is 質 *shitsu* 'quality' and N2 are words such as サービス *sābisu* 'service', 医療 *iryō* 'healthcare', 教育 *kyōiku* 'education', 情報 *jōhō* 'information'. On the other hand, the particle *ga* rather appears in the patterns with funcational words ため *tame* 'for, because', こと *koto* 'thing, matter [nominalizer]', such as ～度が高いため *… do ga takai tame* 'for there is a high degree of …', 率が高いため *ritsu ga takai tame* 'for there is a high rate', 割合が高いこと *wariai ga takai koto* 'that the proportion is high'. The reason for this is related to the fact that the conversion from *ga* to *no* following N1 enables another *ga* to appear as a particle following N2 and marking it as a subject of a sentence. Further elaboration on syntactic, semantic or pragmatic reasons for this is required.

5 度 *do* 'degree' appears as a suffix which forms words with various nouns. Within the target pattern these nouns are, for example, 完成度 *kanseido* 'degree of perfection; level of completion', or 難易度 *nan'ido* 'degree of difficulty'.

6 For each pattern, *N1ga takai N2* and *N1 no takai N2,* the most frequent thousand occurances are extracted, thus the patterns of frequency less than 16 for the pattern *N1ga takai N2* and less than 31 for the pattern *N1 no takai N2* are not taken into account.

Table 5. Tendencies in the pattern: particles *ga* and *no*

N1 ga takai N2	Frequency	N1 no takai N2	Frequency
率 -- 高い こと *ritsu -- takai koto* 'that there is a high rate of'	1409	度 -- 高い もの *do -- takai mono* 'a thing with a high degree of'	1984
度 -- 高い こと *do -- takai koto* 'that there is a high degree of'	971	質 -- 高い サービス *shitsu -- takai sābisu* 'a high-quality service'	1474
度 -- 高い もの *do -- takai mono* 'a thing with a high degree of'	664	背 -- 高い 人 *se -- takai hito* 'a tall person'	1411
度 -- 高い ため *do -- takai tame* 'for there is a high degree of'	614	度 -- 高い 作品 *do -- takai sakuhin* 'work with a high degree of'	1303
率 -- 高い ため *ritsu -- takai tame* 'for there is a high rate of'	578	背 -- 高い 男 *se -- takai otoko* 'a tall man'	992
効果 -- 高い こと *kōka -- takai koto* 'that there is a high effect'	567	質 -- 高い 医療 *shitsu -- takai iryō* 'a high-quality healthcare'	933
背 -- 高い 人 *se -- takai hito* 'a tall person'	565	価値 -- 高い もの *kachi -- takai mono* 'a high value thing'	858
リスク -- 高い こと *risuku -- takai koto* 'that there is a high risk of'	494	質 -- 高い もの *shitsu -- takai mono* 'a high quality thing'	727
率 -- 高い 気(...) *ritsu -- takai ki(...)* '(it seems) that there is a high rate of'	374	質 -- 高い 教育 *shitsu -- takai kyōiku* 'high-quality education'	653
レベル -- 高い こと *reberu -- takai koto* 'that there is a high level of'	343	背 -- 高い 方 *se -- takai kata* 'a tall person'	628
割合 -- 高い こと *wariai -- takai koto* 'that the proportion is high'	328	ヒール -- 高い 靴 *hīru -- takai kutsu* 'high heel shoes'	614
値段 -- 高い こと *nedan -- takai koto* 'that the price is high'	316	レベル -- 高い もの *reberu -- takai mono* 'a thing with a high level of '	587
背 -- 高い 方 *se -- takai kata* 'a tall person'	311	レベル -- 高い 人 *reberu -- takai hito* 'a person with a high level of'	574

4.4 Lexical constraints of attributive roles

In order to explore the collocational relations of i-adjectives and nouns, the 500 most frequent adjectives and their most frequent collocates have been selected from JpTenTen and BCCWJ (Srdanović 2014). For highly frequent adjectives up to 100 collocates are taken, for the remainder, up to 50 collocates. To avoid unclear data, collocates below frequency 5 for JpTenTen and 2 for BCCWJ have been excluded (since BCCWJ is smaller in size, set frequency is also lower). As can be noted in Table 6, while 500 of the most frequent adjectives could be retrieved from both corpora without too many problems, the number of discovered collocations significantly lowers in the case of a smaller corpus, which confirms what has been already stated about corpora usage limitations in relation to its size and different language phenomena. This must be considered in descriptive linguistic studies and lexicographic work in order to uncover and describe complete collocational information.

Furthermore, the analysis of the retrieved data on i-adjectives and noun relations revealed that a number of target adjectives discovered in both corpora have no

attributive role or a quite rare one when compared to the other usages of a particular adjective. Table 6 reveals a number of adjectives with no or a very rare attributive role. Here we can expectedly notice the opposite trend, i.e., that the smaller corpus lists a larger number of adjectives with no or a rare attributive role (83), while the larger corpus recognized more cases of attributive role usage in some of the adjectives and, thus, lists a much smaller number of adjectives with no or a rare attributive role (23). For example, 止む無い *yamunai* 'unavoidable', listed in BCCWJ as one with no or a rare attributive role (only 3 cases), appears in JpTenTen with an attributive role in more than 200 cases (for example, 止む無い事情 *yamunai jijō* 'unavoidable circumstances', 止む無いこと *yamunai koto* 'unavoidable thing' etc.).

A closer look into the results of some of the i-adjectives implies the need to use larger language data when attempting to discover particular lexical constraints. Thus, larger corpora is more reliable for this purpose.

Table 6. Lexical constraints of attributive roles in i-adjectives observed in two corpora

Corpus	Number of adjectives	Number of collocations	No or very rare attributive role
JpTenTen	500	23220	23
BCCWJ	500	9218	83

Table 7 shows a more detailed corpus analysis in some of the retrieved adjectives with no or a very rare attributive role. For some adjectives, no attributive role is discovered in the corpus and they do not directly proceed and modify a noun, such as the adjectives *tegarui* 'easy' and *tokorosemai* 'crowded'. These adjectives appear, rather, in other forms and patterns: *tegarui* in its nominalized form *tegarusa* 'easiness' or in compound 手軽すぎる *tegarusugiru* 'too easy', and *tokorosemai* in its predicative form preceding a clause, such as *tokorosemashi to narande iru* 'to be lined up crowdedly'.

However, some adjectives do appear in attributive roles but these were not retrieved by the search engine for various reasons (e.g. different orthographic form), for example 訝しい *ibukashii* 'suspicious' with rare attributive role いぶかしい顔 *ibukashii kao* 'suspicious face'.

A future task of this study would be to explore some other search methods in order to retrieve more definite data on lexical constraints from the available corpora and to differentiate between non-attributive roles and very rare attributive role.

Table 7. I-adjective with no or very rare attributive role: revisited

Adjectives (Ai)	Freq	Most frequent patterns	Modified noun
tegarui 手軽い 'easy'	14542	手軽さ *tegarusa* 'easiness', お手軽さ *otegarusa* 'easiness', (お) 手軽すぎる *(o)tegarusugiru* 'too easy'	/
ibukashii 訝しい 'suspicious'	12271	いぶかしげ な顔 を する *ibukashigena kao wo suru* 'to have a suspicious face', いぶかしく 思う *ibukashiku omou* 'to think suspiciously', 訝しがる *ibukashigaru* 'to be suspicious', いぶかしげ に *ibukashige ni* 'suspiciously'	* with different orthographic form いぶかしい顔 *ibukashii kao* 'suspicious face' (13) ・表情 *hyōjō* 'expression' (11)…
tokorosemai 所狭い 'crowded'	11569	所狭しと 並ん で いる *tokorosemashi to narande iru* 'to be lined up crowdedly', 所狭し と 並べ られ て	/
nikui/gatai 難い 'hard'	9273	し難い *shinikui* 'hard to do', わかり難い *wakarinikui* 'difficult to understand', サビ にくい *sabinikui* 'resistant to rust', 代え 難い すばらしい もの *kaegatai subarashii mono* 'amazing things hard to replace' (?)	* as a compound adjective
omowashii 思わしい 'suitable, well, convenient'	7588	体調 が 思わしく ない *taichō ga omowashiku nai* 'not to feel well / to be in poor health', 結果 が 思わしく なかっ た *kekka ga omowashiku nakatta* 'the results were not favorable' *often used in negation	思わしい 結果 *omowashii kekka* 'faborable results' (41)・効果 *kōka* 'effects' (13) が 出ない *ga denai* 'cannot get'…

5 Conclusion

This research showed the importance of using large-scale language resources and state-of-the-art tools in empirical studies in order to challenge the currently available traditional approaches to language study and existing findings. It explores Japanese i-adjectives from various perspectives, focusing on the distribution of i-adjectives in present-day corpora, their patterns and constraints. The research on the distribution of i-adjectives revealed the i-adjectives that predominate over the usage of other i-adjectives and provided some new insights into the productivity of adjectival suffixes in Japanese. Next, this research explored the distribution of patterns and their adjectival

roles in the case of the adjective *takai* 'high' and revisited the previous research on the usage patterns of a number of i-adjectives in their attributive role. The research results indicated a need to reconsider the three major roles of adjectives: predicative, attributive and adverbial for their subcategorization in the domain of corpus annotation but also within the domain of the grammar of the Japanese language.

Furthermore, the patterns Nが 高い N *N ga takai N* 'N with high N' and Nの高い N" (*N no takai N* 'N with high N') are examined and other than providing the most frequent patterns, the study had some interesting finding on the *ga/no* conversion in the specified pattern: e. g. *ga* tends to be used for more abstract phenomena while the more frequent *no* for more concrete things. Finally, this study examined the lexical constraints in the attributive forms of i-adjectives and discovered some adjectives with no or a rare attributive role.

Literature

Baroni, M. and Ueyama, M. (2006) Building general- and special-purpose corpora by Web crawling. *Proceedings of the 13th NIJL International Symposium*, 31-40.

Baroni, M. and Bernardini, S. (2004) Bootcat: Bootstrapping corpora and terms from the web. *Proceedings of LREC*, 1313-1316.

Baroni, M. and Kilgarriff, A. (2006) Large linguistically-processed web corpora for multiple languages. *Proceedings of EACL*, 87-90.

Joyce, T., Hodošček, B. and Nishina, K. (2012) Orthographic representation and variation within the Japanese writing system: Some corpus-based observations, *Written Language and Literacy*, 15 (2) (Special issue: Units of language - units of writing, edited by T. Joyce and D. Roberts), John Benjamins Publishing Company, 254-278.

Kilgarriff, A., Reddy, S., Pomikálek, J. and PVS, A. (2010) A corpus factory for many languages. *Proceedings of LREC*, 904-910.

Kilgarriff, A., Rychly P., Smrž, P., Tugwell, D. (2004) The Sketch Engine. *Proceedings Euralex*, 105-116.

Harada, S. (1971) Ga-no conversion and idiolectal variations in Japanese. *Gengo Kenkyū (Language Research)* 60, 25-38.

Hashimoto, M. and Aoyama, F. (1992) Three usages of adjectives. *Kēryo Kokugogaku (Mathematical Linguistics)*, 18 (5), 201-214.

Nation, P. (2001) *Learning vocabulary in another language*. Cambridge University Press

Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., Koiso, H., Yamaguchi, M., Tanaka, M. and Den, Y. (2013) Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation*. Springer Netherlands

Nishio, T. (1972) *A descriptive study of the meaning and uses of Japanese adjectives. NLRI 44*, Shuei Shuppan

Ogura, H., Koiso, H., Fujiike, Y., Miyauchi, S., Konishi, H., Hara, Y. (2011) *Gendai nihongo kakikotoba kinkō kōpasu keitai-ron jōhō kitei-shū dai-yon-ban*. Technical Report LR-CCG-20-05-01, The National Institute of Japanese Language and Linguistics.

Pomikálek, J. and Suchomel, V. (2012) Efficient web crawling for large text corpora, *Proceedings of the Seventh Web as Corpus Workshop (WAC7)*.

Schönefeld, D. (1999) Corpus Linguistics and Cognitivism. *International Journal of Corpus Linguistics*, 4 (1), 137-171.

Sharoff, S. (2006) Creating general-purpose corpora using automated search engine queries. *WaCky! Working papers on the Web as Corpus*. Gedit.

Srdanović, I., Erjavec, T. and Kilgarriff, A. (2008) A web corpus and word-sketches for Japanese. *Shizen gengo shori (Journal of Natural Language Processing)* 15 (2), 137-159.

Srdanović, I. (2013a). Japanese i-adjectives as short and long unit words: implications for language learning. *PACLING 2013: Conference proceedings*, September 24, 2013 Tokyo: Pacific Association for Computational Linguistics, 8pp.

Srdanović, I. (2013b) Collocation and Syntax: Adjective and Noun Collocations, *Proceeding of the 4th Japanese corpus linguistics workshop*, Department of Corpus Studies/Center for Corpus Development, NINJAL, 267-284.

Srdanović, I., Suchomel, V., Ogiso, T., Kilgarriff, A. (2013) Japanese Language Lexical and Grammatical Profiling Using the Web Corpus JpTenTen, *Proceeding of the 3rd Japanese corpus linguistics workshop*, Department of Corpus Studies/Center for Corpus Development, NINJAL, 229-238.

Srdanović, I. (2014). Corpus based collocation research targeted at Japanese language learners. *Acta linguistica asiatica.* 4 (2), 25-35.

Stefanowitsch, A. and Gries, S. T. (2003) Collostructions: investigating the interaction of words and constructions. *International Journal of Corpus Linguistics*, 8 (2), 209-243.

Suzuki, S. (1972) *Japanese Grammar and Morphology*. Mugi shobo.

Tsujimura, N. (1996) *An Introduction to Japanese Linguistics*. Blackwell Publishers.

Internet resources

Sketch Engine: http://www.sketchengine.co.uk/ (17.8.2019.)

Chunagon: https://chunagon.ninjal.ac.jp/bccwj-nt/search (10.10.2019.)

要旨 (Abstract in Japanese)

「日本語のコーパスにおけるイ形容詞–分布、パターン、語彙制約—」

イレーナ・スルダノヴィッチ
(ユライドブリラ大学プーラ)

本稿では、コーパス言語学の実証的手法および最先端の言語資源と語彙プロファイリングツールを用いて日本語の形容詞の使用について検討する。まず、分析に利用したリソースを紹介し、その重要性と特徴について述べる。次に、現代日本語大規模コーパスにおける形容詞の分布を分析することにより、使用頻度の高い形容詞、さらには複合形容詞を形成する生産性の高い形容詞および接尾辞の用法の多様性を明らかにする。続いて、形容詞が持つ3つの主要な役割のパターンの分布を分析し、形容詞によって役割とパターンに異なる使用傾向が見られることを示す。特に、形容詞の連体修飾用法に焦点を当て、用法のパターンの複雑さを明らかにし、形容詞の連体修飾用法のタイプをより詳細に分類する必要があることを指摘する。さらに、形容詞の連体修飾用法における語彙的制約を調査した結果、連体修飾用法を全く持たない、あるいはほぼ持たない形容詞を見出すことができた。

III

Research of corpora and Japanese language learning

7 Readability measurement of Japanese texts based on levelled corpora

LEE Jae-ho
Waseda University
HASEBE Yoichiro
Doshisha University

Abstract
A method is presented here to measure the readability of Japanese texts using levelled corpora. Two sets of levelled corpora were constructed for this purpose: one was used as model data to devise a readability measurement formula, and the other as test data to check the validity and reliability of the formula. Six-level model corpora were constructed at first using texts extracted from Japanese textbooks and Japanese Diet meeting transcripts. We examined these corpora both manually and statistically. Then a multiple regression analysis of the results of these examinations was carried out. Among the five models produced, the best model was selected and used to construct a readability formula. The formula was tested using the other set of levelled corpora based on 25 years of reading passages from the Japanese-Language Proficiency Test (JLPT), and its reliability was confirmed. A web-based system was also developed using the formula to aid teachers of Japanese in preparing reading materials that match student levels. The system also has much reading-related functionality, making it helpful to teachers and learners, as well as allowing wide access of the present research to a broad range of people involved in teaching, learning, and studying Japanese.

Keywords: readability, levelled corpora, regression analysis, web-based system

1 Background and purpose

Text readability studies aim to devise methods to measure reading difficulty in natural language texts. Research in this field has developed systematic procedures that rank the level of a given text based on various indices such as the mean number of words per sentence. There is a long tradition of such attempts for texts written in English, and a number of methods and formulae have been proposed (e.g., Flesch 1948; Smith and Kincaid 1970). In recent years, readability studies have also been actively pursued to measure texts in Japanese (e.g., Sakamoto 1964; Tateishi et al. 1988; Shibasaki and Hara 2010; Sakai 2011; Sato 2011). Moreover, several web-based systems targeted at Japanese native speakers have been developed utilizing various methods and formulae[1].

1 Shibasaki and Hara (2010) have made their online system available at the following website: http://readability.nagaokaut.ac.jp/readability, and Sato (2011) at: http://kotoba.nuee.nagoya-u.ac.jp/sc/readability/index.html.

No matter what the target language, virtually all studies in text readability measurements have been completed with the following two points in mind: 1) What are the essential factors that determine the level of the text? 2) How is it possible to formalize the relationship among various factors and produce a readability formula? As to question 1, the factors need to be broadly divided into two types. On the one hand, there are macro factors such as topics and coherence, and on the other, there are micro factors such as levels of vocabulary items, degrees of complexity of grammatical structures, and length of words and sentences. Focused primarily on the factors of the latter type, Shibasaki and Hara (2010) produced a readability formula for Japanese texts by using a linear regression analysis which included indices such as the proportion of *hiragana* characters in the text, the mean number of predicates per sentence, the mean number of characters per sentence, and the mean number of *bunsetsu* boundaries[2] per sentence. As to question 2, much previous research thus far has adopted statistical methods, such as principal component analysis and regression analysis, applying them to Japanese text data that were formatted in specific ways.

The research presented in this paper aims at advancing text readability studies for the Japanese language and devises a practical and useful system that contributes to Japanese language teaching, learning, and research. More specifically, utilizing levelled corpora, mainly consisting of texts from Japanese textbooks[3], we produced the following formula to measure the readability level of a given text in a six-level scale: X = {mean length of sentence ∗ -0.056} + {proportion of *kango* ∗ -0.126} + {proportion of *wago* ∗ -0.042} + {proportion of number of verbs among all words ∗ -0.145} + {proportion of the number of auxiliary verbs ∗ -0.044} + 11.724 (R^2=0.896). The formula was tested against another set of levelled texts in Japanese to prove its reliability[4]. Lastly, the method was implemented in a computer system that calculates and produces the estimated level of a text via a web-based online interface.

It should be noted that the project presented in this paper is original in several ways. Firstly, the readability formula we constructed is intended especially for learners of Japanese as a foreign language, whereas many existing formulas such as those by Shibasaki and Hara (2010) and Sato (2011) are intended for native readers of Japanese. Secondly, our online implementation offers new functionalities that are not available in existing systems for reading support. These points are explicated in the following sections.

---

2 A *bunsetsu* is a unit of text in Japanese that is comprised of a content word plus the optional function word(s) that immediately follow it (Zhang and Ozeki 1998).

3 In the present paper, "Japanese textbooks" refer to "textbooks used for teaching Japanese to non-native learners".

4 A *kango* is a Japanese word of Chinese-origin and thus is typically written in *kanji* characters, whereas a *wago* is a Japanese word that is neither loaned nor derived from words in a foreign language. A *wago* is typically written in *hiragana* or *kanji* characters in contemporary Japanese.

2. Data and methods

2.1 Overview

Two different sets of data were prepared for our research: model data and test data. The former consists of two types of text: one comprised of text from 83 Japanese textbooks, ranging from introductory to advanced, and the other comprised of text from National Diet meeting transcripts, selected according to the criteria explained in 2.2. From this basic data, we created corpora of six different levels. The readability measurement formula was produced by analyzing these levelled corpora. The latter dataset, that for testing the formula, consists of texts derived from 25 years of the Japanese-Language Proficiency Test (JLPT).

The levelled corpora for analysis were created from the original data in the following way. First, all texts were split into separate files of roughly the same size (around 1,000 characters). Second, each file was manually examined and then analyzed computationally and this enabled us to obtain corpora of six different levels. Then, each component text file in each of these levelled corpora was analyzed further using natural-language processing (NLP) tools, and various text features such as the frequency of words of different categories and different parts-of-speech were obtained. Using such numerical data as input values, a multiple linear regression analysis was conducted and, as a result, our readability measurement formula was finally obtained. The formula was then tested against a second dataset derived from JLPT, and its effectiveness was verified. The whole process is schematically summarized in Figure 1.

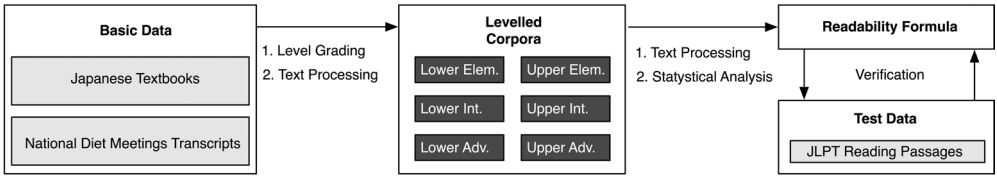


Figure 1. Data and procedures

2.2 Creating levelled corpora

The six-level scale we utilized throughout our research corresponds to lower-elementary, upper-elementary, lower-intermediate, upper-intermediate, lower-advanced, and upper-advanced levels. The model corpora of the first five levels were created using texts in Japanese textbooks, and that of the most advanced level was created from the text of National Diet meeting transcripts, which were included in the *Balanced Corpus*

of Contemporary Written Japanese (BCCWJ)[5]. The total number of words in the levelled corpora is 595,360. Table 1 shows how these texts were divided[6].

Table 1. Basic statistics of the levelled corpora

	Lower-elem. (133)	Upper-elem. (117)	Lower-int. (148)	Upper-int. (286)	Lower-adv. (117)	Upper-adv. (194)
Word types	3,178	2,858	5,156	10,291	6,833	4,712
Word tokens	72,691	68,746	87,433	174,953	69,268	122,269

*Numbers inside parentheses represent the number of text passages included

The actual procedure for grouping the original data into these levels was comprised of three steps. First, the first author of the present paper checked the general design (such as purpose, contents, and featured study items) of each of the textbooks in the original dataset, and categorized them into five levels from lower-elementary to lower-advanced. Second, we asked three practicing teachers of Japanese to manually examine the text passages thus categorized and choose only those that they thought truly matched the given level. Finally, the results were further verified using the statistical method of discriminant analysis.

2.2.1 Choice of data and data size

There are two supplementary comments on the basic statistics of the levelled corpora presented in Table 1. The first concerns the choice of the original data, and the second concerns data size.

The decision to use Japanese textbooks to construct a corpus for each of the five levels from lower-elementary to lower-advanced was motivated by the following: in text readability studies, it is required that a clear indication of the level of the model data be already given so that a formula can be drawn by analyzing it. Thus, it has traditionally been the case that readability research uses language textbooks. The reasoning behind this is obvious: textbooks are written according to the assumed levels of the readers who use them. The vocabulary, idioms, structures, and types of logic used in textbooks of

5 http://pj.ninjal.ac.jp/corpus_center/bccwj/en/

6 One may expect vocabulary variation (number of word types) to increase as the level of difficulty increases. In Table 1, however, there are less word types in the upper-advanced corpus (Diet transcripts) than in the lower-advanced (textbooks), even though it is larger than the lower-advanced corpus. This is probably due to the fact that Diet transcripts repeatedly deal with a rather limited set of topics. Another possibility is that sentences in the Diet transcripts tend to be composed by combining two or more clauses with conjunctions, and as a result they contain a relatively larger number of functional words than sentences of other types of text.

different grades are, in general, fairly controlled. We find this characteristic of language textbooks ideal for our purpose. In fact, however, there are some researchers who see language in textbooks as unnatural, or at least somehow different from language observed elsewhere. This is actually a matter of degree and the same can be said about written language of any kind. We concluded that the benefit of using textbooks exceeded any possible drawbacks.

We used National Diet meeting transcripts for our highest-level corpus based on the following four motivations. First, these are transcripts of genuine utterances, and are not artificially created data. This results in a variety of styles in the data, which is often considered characteristic of a highly advanced set of linguistic data. Second, this approach provided a sufficient amount of text. As shown in Table 1, the number of words for this level is comparable to those of other levels, even if not necessarily exceeding them. Third, the sentences used are relatively long, which is broadly considered a condition for an advanced text. Finally, the fourth reason was that the data contained utterances dealing with abstract concepts and ideas. For these reasons, and also taking into consideration the facts observed in texts of different registers by Lee (2011), we made a decision to exclusively use National Diet meeting transcripts to compile the corpus associated with our upper-advanced level. Lee (2011) carried out a close examination of the text in National Diet meeting transcripts and showed that it should be placed well beyond the level of texts used for JLPT L1 tests (highest-level).

Another point that requires comment is that the data size of the five corpora from lower-elementary to lower-advanced is not balanced, as is apparent in Table 1. This is due to the fact, firstly, that there is a relatively larger number of available titles of textbooks at the intermediate level. Secondly, elementary level textbooks contain shorter sentences and they accordingly have fewer words. A third reason is that there are only a limited number of available titles for advanced learners. Thus, the data size of the corpora at different levels is different. Still, each corpus has a fairly large amount of text, and the effect of size difference among corpora was considered very small, if any.

2.2.2. The rationale behind these six levels

So far we have not presented a sufficient explanation as to why all texts were split into files of approximately 1,000 characters and why we adopted the six level-scale in the first place.

The reason behind creating text files of roughly the same size had much to do with the fact that in text readability studies, various indices regarding "length" observed in a text have much to do with the level of the text. Such indices include the mean length of sentences, the mean length of words, and the total number of words in a text. It is essential for text readability research to make certain that such indices are retrieved as

accurately and efficiently as possible.[7] Thus, standardizing text size is a prerequisite for obtaining characteristics regarding readability. The choice of 1,000 for the number of words contained in one text file is, however, rather arbitrary. It is not necessarily based on a specific scientific fact. Rather, it is motivated by the fact that in many Japanese language courses for non-native speakers a text of about 1,000 characters is typically preferred because this fits in well with the duration of a lesson.

We categorized the model text files into six levels. It would have been possible for us to choose two or three category levels instead, as many textbooks are simply levelled as "elementary," "intermediate," and "advanced," but did not for several reasons. Firstly, a textbook teaching Japanese often contains materials of different levels; the level of the very first chapter in the book can be largely different from that of the last chapter in the same book. Actually, this is quite a natural phenomenon, as textbooks are designed so that their users' ability gradually increases as the pages proceed. This fact urged us to break up one single textbook into parts and put them into different categories according to the specific content level. Secondly, the common practice of dividing textbooks into "elementary," "intermediate," and "advanced" is not necessarily rigidly standardized among publishers and authors. Some textbooks adopt a level system based on the frequency of the use of complex grammatical constructions, but others adopt a level system based exclusively on the vocabulary items used. For these reasons, we devised a six-level scale, which does not exclusively depend upon either grammatical or lexical characteristics, nor did we risk placing texts into two or three haphazard levels.

As a result of splitting the dataset into files of about 1,000 characters, we obtained 995 text files in Japanese. Levelling them was not, however, necessarily completed by reading text files and manually sifting them one-by-one. This would be not only time and resource consuming, but also highly ineffectual. Thus, we devised a method of text categorization that made use of both human graders and computational tools. First, the textbook dataset was roughly sorted into five levels from lower-elementary to lower-advanced by one of the authors of this paper (mainly according to the general facts already known about the titles). Then we asked three teachers of Japanese who have more than 10 years of teaching experience to examine all the files in each of the levelled file pools that were created in the pre-categorization process. They were then asked to pick out exactly 30 files that they thought contained texts quite representative for each of the five levels. Then we selected 20 files that were chosen by multiple graders for each level, creating a subset of the original dataset that was comprised of five groups of 20 files, each of which is thought to be more or less prototypical of each level. Furthermore, we

7 As is pointed out by some of the pioneers in this field such as Flesch (1948), Sakamoto (1964), and Smith and Kincaid (1970), the larger the length of a sequence, whether it is of a sentence or a word, the heavier the burden on working memory. The readability of a text is roughly in negative correlation with the mean size of various textual elements.

carried out a discriminant analysis (described in more detail in 3.1), using these core data as a model, against the text files that had been "filtered out" in the previous process, and finally obtained the levelled corpora of text, each of which contained not only 20 files, but also files that supposedly have similar textual characteristics to those of the core data. Table 2 shows descriptions of the assumed abilities of readers of each level given to the graders before they examined the texts.

Table 2. Descriptions of readers' reading abilities for six levels

Level	**Description**
Upper-advanced	The reader is able to fully understand highly technical writing. S/he has no difficulty dealing with virtually any kind of text in Japanese.
Lower-advanced	The reader is able to mostly understand technical writing. S/he can deal with complex structures often observed in literary works.
Upper-intermediate	The reader is able to grasp the overall structure of technical writing. S/he can deal with Japanese texts found in most day-to-day situations without much difficulty.
Lower-intermediate	The reader is able to read relatively simple writing and can deal with texts comprising multiple sentences.
Upper-elementary	The reader can understand basic vocabulary items and grammatical patterns. S/he can deal with complex sentences of basic types such as ones involving *-te* form.
Lower-elementary	The reader can understand the most fundamental Japanese expressions used in simple sentences. S/he has difficulty in dealing with complex sentences or sentences containing adnominal modifiers.

The following passages are samples of the core data collected as a result of the process mentioned above for the five category levels from lower-elementary to lower-advanced, and a sample of text of the upper-advanced level, which is from National Diet meeting transcripts.

1) Lower-elementary

音楽がすきですから，よくＣＤを聞きます．日本がすきですから，日本語を勉強します．安かったですから，買いました．ディズニーランドは楽しかったです．教室は静かでした．わたしはラーメンがすきです．わたしはたばこがきらいです．ワンさんは日本語が上手です．わたしは料理が下手です．

I like music, so I often listen to CDs. / I like Japan, so I study Japanese. / I bought it because it was inexpensive. / I had fun at Disneyland. / It was quiet in the classroom. / I like ramen. / I hate cigarette smoke. / Mr. Wang is good at speaking Japanese. / I am not good at cooking.

2) Upper-elementary

わたしは夏休みに国へ帰らないつもりです．わたしは30歳まで結婚しないつもりです．わたしは大学へ行かないつもりです．わたしは学校では母国語を使わないつもりです．わたしは車に乗らないつもりです．今年の夏も国へ帰りますか．はい，そのつもりです．いいえ，帰らないつもりです．

I am not going to my home country during the summer holiday. / I will not marry until I am thirty.

I am not going to a university. / I will not use my native language at school. / I will not drive a car. / Are you going to your country in the summer again this year? Yes, I am. / No, I'm not.

3) Lower-intermediate

毎週1回は祖母の家に子どもたちが孫たちをつれて集まります．とてもにぎやかです．祖母の80さいの誕生日には，マニラで一番大きなホテルを借りて，大家族の全員と親しい友人が，全部で500人以上集まりました．ごちそうを食べたり，ダンスをしたり，歌をうたったりして，とてもにぎやかでした．祖母もワルツやチャチャチャをおどりました．それから子どもと孫の全員が花をプレゼントしました．

My grandmother has all her children and grandchildren gather at her house once a week. Her house is filled with laughter and lively conversations. On her 80th birthday, we held a big party at the biggest hotel in Manila with more than 500 participants, including all of her family members and friends. At the boisterous party, we enjoyed wonderful food, sang songs together, and danced to the music. Grandma herself danced a waltz and the cha-cha. At the close of the party, all her children and grandchildren presented her with flowers.

4) Upper-intermediate

今でいうリフォーム，リサイクルをごく当たり前のこととしてやっていました．日本は，1950年代後半から経済の成長がいちじるしく，供給がどんどん増加し，国民一人あたりの所得も上がってきました．この時代を境にして，需要と供給のバランスが逆転しました．現在の日本は完全に供給が過剰，需要が不足している時代です．ものをつくる企業はこういうときにどうするでしょうか．

Back then, many Japanese people were already doing what we now call “reforms” and “recycles.” However, the economy grew so rapidly in the 1950s that there was a tremendous increase of supply, and the average income earned per person rose markedly. It was the time when the supply-demand unbalance gradually set in. Now, Japan is in a state of excess supply with limited demand. We should think about what manufacturers can do under such circumstances.

5) Lower-advanced

「実現への戦略―それは産業界を納得させる手順」―土地税制改革のため、大蔵省が省内論議をまとめた内部資料に、こんなくだりがある。「土地などの資産所得が勤労所得よりはるかに大きくなり，勤労意欲を低下させて日本の経済・社会基盤を揺るがしている」「いや，資産効果で消費は拡大，重厚長大産業は土地を活用して企業基盤を強くしている面もある」…連日の，こんな議論を経て土地保有税創設に動き出した大蔵省が，土地税制改革のポイントは対財界戦略にあるとみていることを示す文言だった．

“The Strategy for Realization: A Procedure to Persuade Business Sectors”—this is a phrase found in one of the Finance Ministry’s internal documents that contain summaries of discussions on the land-tax system. “For many people, the income from their real estate assets has become far larger than their earnings from work, and the decrease in their incentives to work is starting to largely influence the economic and social foundations of Japan.” “It does not necessarily explain the whole picture—the asset effect has increased consumption, and the heavy industries are strengthening the corporate infrastructure by making good use of the land they own.” The Finance Ministry, after repeated discussions like this, is currently doing preparatory work for creating a land-holding tax. They consider it extremely important to be highly strategic in negotiating with the major players in the business community, as is apparently shown in the above-mentioned passage.

6) Upper-advanced

「あの際の米軍による行動が，イラクに関連する一連の国連安保理決議の履行を確保するため，それに必要な措置ということであれば，我が国としてはこれを理解し，支持する，こういうことを申したわけでございまして，我が国としてはいわば無条件で米国のやることはすべて支持しますよということは申し上げておりません．御承知のとおり，国連には一連の決議がございまして，イラク軍が北部イラク地域から撤退するようにということをずっと国連として求めておったわけでございます．そういったことが確保されるために必要な措置ということで米軍が行動するのであれば，それは理解し，支持する，こういうことを明らかにしたということでございます．」

"What I intended to state was only that our country should understand and support the conduct of the American Armed Forces on that occasion because it was a necessary procedure to make sure the series of UN Security Council Resolutions on Iraq were implemented; I was not saying that our country would endorse everything that the US does without any conditions. As you know very well, the UN has made a series of resolutions, and they clearly requested that Iraqi forces withdraw from the Northern Iraq regions. If the US Armed Forces were conducting a necessary action to ensure the implementation of the UN resolutions, then the Japanese government should understand and support it. That is what I intended to make clear then."

2.3 Selection of formula variables

In order to construct a formula to calculate the readability of Japanese texts, firstly it was necessary to analyze our model data with NLP tools. Thus, we analyzed our dataset using the Japanese morphological analyzer MeCab 0.996 with UniDic 2.2.0.[8] Obtained from this process were types of data such as: 1) mean length of sentence, 2) proportion of nouns, 3) proportion of auxiliary verbs, 4) proportion of verbs, 5) proportion of subsidiary verbs, 6) proportion of adjectives, 7) proportion of *wago* words, and 8) proportion of *kango* words. We selected these elements based on work by Shibasaki and Hara (2010), as candidates for variables to be used in our formula.

In our selection of elements for use as variables, there were limitations that needed to be considered. Firstly, since the resulting formula would be computationally implemented in a web-based readability measurement system, only values that could be immediately calculated were available to us. In reality, there could be numerous variables that affect the readability of texts. Theoretically, it is conceivable that there are not only

8 MeCab (http://taku910.github.io/mecab/) can be used with one of several available dictionary packages, of which UniDic is one option (http://osdn.jp/projects/unidic/). UniDic is superior to other dictionaries in that the format of its entry items is systematically standardized based on short-unit words (SUW) and it offers richer lexical information including that of word types regarding etymological origins (*wago,* words of Japanese-origin; *kango,* words of Chinese-origin; or *gairaigo*, words of Western-origin). See Den (2009) for further details about UniDic.

purely numerical ones such as the frequency of certain type of words, but also those that represent more abstract aspects of texts such as the overall cohesion, the stylistic tone of the text, or even the size of font type and the color of a printed text. However, we had to exclude from our formula those types of information that are difficult to obtain computationally, even though some might be effective in determining the real readability of a text.

Secondly, although using an NLP dependency analysis tool could be helpful for producing an accurate formula, it was not a realistic option. In fact, Shibasaki and Hara (2010) used the results of dependency parsing in their model. Tools for dependency parsing are currently available, including ones that were adopted by Shibasaki and Hara (2010)[9]. However, they suffer from a problem of insufficient accuracy (more than 10 percent of text is analyzed incorrectly). Thus, we decided not to use this type of technology in constructing our formula and the web-based system we built based on the formula.

Thirdly, we chose to use only variables that are proportional, instead of those that are numerically absolute. The output of a formula that adopts the latter types of variables would be significantly influenced by the size of the input text. This makes it difficult to compare readability scores for texts of different sizes. By using only proportional frequencies, we can measure the readability of texts of any size and we can make sure that the resulting scores are comparable to each other.

The formula was constructed with linear regression analysis. Linear regression analysis is a statistical method that has also been used in past readability studies (e.g., Tateishi et al. 1988; Shibasaki and Hara 2010). It is helpful when explaining the correlation among two or more variables based on a linear model. We conducted multiple linear regression analysis using IBM SPSS (ver. 22).

2.4 About test data

In addition to the levelled corpora based on the core dataset described above, we also built a test corpus comprised of text files other than those contained in the latter to confirm the validity and the reliability of the formula.

There is an important fact to note regarding the test data. The levels estimated for input texts using our formula do not necessarily have pre-existing external criteria. In fact, this is the case with virtually every attempt in text readability measurement. Suppose, for instance, one desires to measure the readability of a Japanese newspaper article by applying a readability formula to the text and obtains an estimated upper-advanced level. How do we verify that the result is correct, or reject it as incorrect? As such, readability levels are to some extent inevitably subjective. Thus, the verification of the readability formula is not necessarily an easy task.

9 Shibasaki and Hara (2010) used CaboCha, a Japanese dependency structure analyzer (http://taku910.github.io/cabocha/).

To minimize such concerns and also to verify that the application of our formula was as reliable and usable as possible, we constructed test corpora using texts from reading passages in JLPT from 1984 to 2008. The division of the data is presented in Table 3.

Table 3. Test corpora

Level	Number of words	Mean number of words per sentence
L1 (78)	50,511	28.3
L2 (66)	42,586	24.5
L3 (17)	10,541	16.4
L4 (11)	6,242	10.9

* Numbers inside parentheses represent the number of text passages included

As in the case of the model data, the test data consisted of text files, each of which contained around 1,000 characters. The L1 level (highest-level) corpus had 50,511 words in total and was comprised of 78 files. The corpora of other levels were constructed in the same fashion. Also, as in the case of the model data, the higher the level, the greater the number of words in the corpus. This is mainly because the JLPT tests for more advanced levels have longer sentences than those for lower levels. This is apparent from the mean number of words per sentence in each of the test corpora: 28.3 for L1, 24.5 for L2, 16.4 for L3, and 10.9 for L4.

The test was carried out by examining the degree of match between the test corpora and the estimated levels obtained by applying the data in the test corpora to our formula.

3 Results and discussion

This section describes the procedures and results of the analysis in further detail. 3.1 presents a closer look at the way the levelled data of the model corpora were constructed, explaining how the division of the corpora was drawn from the discriminant analysis. In 3.2, the results of the multiple linear regression analysis carried out to construct the formula are expounded. And in 3.3, the results of the verification of the formula using the test data are presented.

3.1 Results of the discriminant analysis: Constructing the levelled corpora

As briefly described previously, we manually classified the original data and then extracted 20 text files containing data that assumedly matched each of the six levels from

lower-elementary to upper-advanced. The resulting "core" data of 120 files were utilized to classify the other 875 files, that is, the rest of the original dataset of 995 text files, using discriminant analysis. As a result, for the lower-elementary level, 78 text files were re-selected out of 113 files that had been rejected from the core data by graders by way of manual examination. Similarly, 37 files out of 97 files for upper-elementary, 58 files out of 128 files for lower-intermediate, 102 files out of 266 files for upper-intermediate, 60 files out of 97 files for lower-advanced, and 152 files out of 174 files for upper-advanced that had been once rejected by graders were re-selected for the respective levels as presented in Table 4.

Table 4. Discriminant analysis results

		Levels predicted by discriminant analysis						
		Upper-adv.	**Lower-adv.**	**Upper-int.**	**Lower-int.**	**Upper-elem.**	**Lower-elem.**	**Total**
Original levels	**Upper-adv.**	152	14	8	0	0	0	174
	Lower-adv.	6	60	24	7	0	0	97
	Upper-int.	8	70	102	61	22	3	266
	Lower-int.	0	4	39	58	21	6	128
	Upper-elem.	0	1	14	28	37	17	97
	Lower-elem.	0	0	0	7	28	78	113
	Total	166	149	187	161	108	104	875

Finally, among the 995 text files contained in the original dataset, 607 were used to construct the levelled corpora and the other 388 files were filtered out, as the latter files were not grouped to levels either in the selection process by human graders or the discriminant analysis.

3.2 The readability formula

The readability formula was selected from five models generated as a result of multiple linear regression analysis. Figures involved in the analysis are shown in Table 5.

Table 5. Multiple linear regression analysis results

Models		Coefficient	R^2
Model 1	(Constant)	5.938	0.787
	Mean length of sentence	-0.099	
Model 2	(Constant)	6.691	0.839
	Mean length of sentence	-0.082	
	Proportion of *kango*	-0.073	
Model 3	(Constant)	13.195	0.878
	Mean length of sentence	-0.063	
	Proportion of *kango*	-0.153	
	Proportion of *wago*	-0.086	
Model 4	(Constant)	12.128	0.893
	Mean length of sentence	-0.057	
	Proportion of *kango*	-0.142	
	Proportion of *wago*	-0.061	
	Proportion of verbs	-0.159	
Model 5	(Constant)	11.724	0.896
	Mean length of sentence	-0.056	
	Proportion of *kango*	-0.126	
	Proportion of *wago*	-0.042	
	Proportion of verbs	-0.145	
	Proportion of auxiliary verbs	-0.044	

Among the five models constructed by using the multiple linear analysis in Table 5, Model 1 is the simplest. It is composed only of a constant and the mean length of sentences. Its R^2, an index that shows prediction accuracy, is 0.787. Model 2 includes the proportion of *kango*, words of Chinese origin, in addition to a constant and the mean length of sentences, with its R^2 being 0.839. Having examined Models 3 to 5 in the same token, the R^2, the coefficient of determination, of each of the 5 models is plotted as in Figure 2.

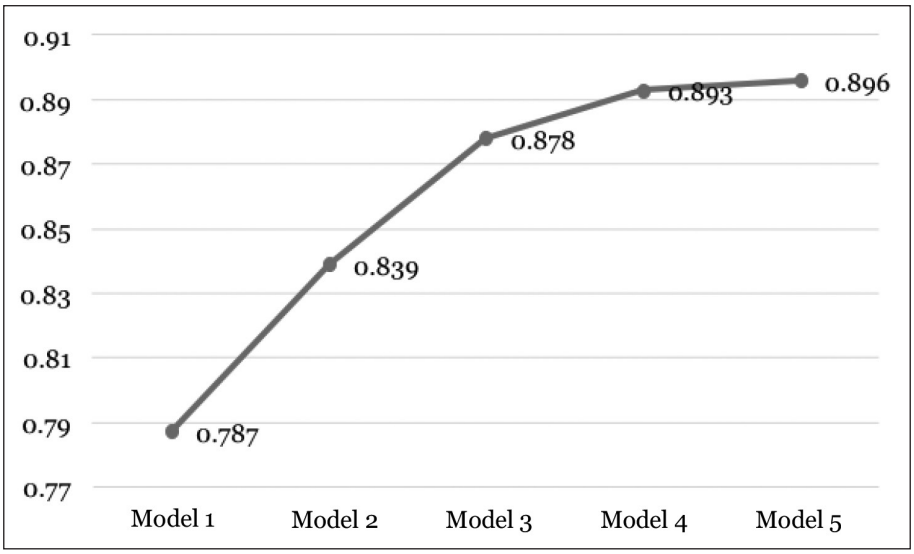


Figure 2. Transition of the coefficient of determination

Among the five models, Model 5 was finally selected as it showed the highest prediction accuracy. Based on this model, the following readability formula was obtained.

Readability Formula for Japanese Language Education (R^2 = 0.896)

X = {mean length of sentence * -0.056} + {proportion of *kango* words * -0.126} + {proportion of *wago* words * -0.042} + {proportion of number of verbs among all words * -0.145} + {proportion of number of auxiliary verbs * -0.044 } + 11.724

The formula shows that three indices are especially effective when measuring the readability level of a Japanese text (for language education). First among them is the mean length of sentence, as would be naturally expected. It is considered that this indirectly reflects the degree of structural complexity of a sentence in the text passage. Secondly, the proportions of *kango* and *wago* are effective. This is considered to be due to the fact that many words of technical and/or abstract concepts tend to be realized as *kango*, whereas most *wago* are considered more basic and fundamental. And thirdly, the proportion of verbs and the proportion of auxiliary verbs are also effective. It is assumed that these two indices reflect, again, the degree of structural complexity of the text. For a more concrete example, our formula is applied to a sample text of the lower-elementary level presented in 2.2 as follows:

{8.56*-0.056}+{9.09*-0.126}+{63.64*-0.042}+{2.60*-0.145}+{22.08*-0.044}+11.724=6.08

The resulting score, 6.08, can be interpreted using a correspondence table as in Table 6. It is within the range of 5.5 to 6.4, thus the text is interpreted as lower-elementary.

Table 6. Levels and readability scores

Level	Readability score range
Upper-advanced	0.5 - 1.4
Lower-advanced	1.5 - 2.4
Upper-intermediate	2.5 - 3.4
Lower-intermediate	3.5 - 4.4
Upper-elementary	4.5 - 5.4
Lower-elementary	5.5 - 6.4

There is a caveat. The resulting readability score could be smaller than 0.5, the lower limit on the table, or larger than 6.4, the higher limit. When such a case arises, then the text can be considered to have some characteristics that our formula cannot properly deal with. For example, an extremely short text that includes many *kango* in long sentences could produce a score less than 0.5. On the contrary, a text passage having many *wago* in extremely short sentences could produce a score over 6.4. In any case, such instances are rightfully considered exceptional when dealing with texts for Japanese reading education.

3.3. Verification results using test data

In this section, the results of verification using the test data introduced in 2.4 are presented. The logic behind the procedure is this: if readability scores produced by applying the formula to texts from JLPT tests, which have already been levelled, predict the text levels sufficiently correctly, then the formula is considered highly valid. The resulting figures of this experiment are summarized in Table 7.

Table 7 presents a cross tabulation of JLPT levels of the test data, on the one hand, and the estimated readability levels calculated using the formula, on the other. Several things can be noted here: 1) the reading passages in JLPT L1 are mostly estimated to be of upper-intermediate or lower-advanced, 2) the reading passages in JLPT L2 are mostly estimated to be lower-intermediate or upper-intermediate, 3) the reading passages in JLPT L3 are exclusively estimated to be upper-elementary or lower-intermediate, and 4) the reading passages in JLPT L4 are exclusively estimated to be lower-elementary or upper-elementary.

Now let us examine the results of the same experiment in the form of numeral scores, instead of discrete levels. Figure 3 represents the distribution of the scores in the form of a

Table 7. Cross tabulation of JLPT levels and levels estimated using the formula

			Estimated readability level					Total
			Lower-elem.	Upper-elem.	Lower-int.	Upper-int.	Lower-adv.	
JLPT Level	L1	Num. of passages	0	0	6	47	25	78
		%	0.0	0.0	7.7	60.3	32.1	100.0
	L2	Num. of passages	0	1	19	44	2	66
		%	0.0	1.5	28.8	66.7	3.0	100.0
	L3	Num. of passages	0	7	10	0	0	17
		%	0.0	41.2	58.8	0.0	0.0	100.0
	L4	Num. of passages	5	6	0	0	0	11
		%	45.5	54.5	0.0	0.0	0.0	100.0
Total		Num. of passages	5	14	35	91	27	172
		%	2.9	8.1	20.3	52.9	15.7	100.0

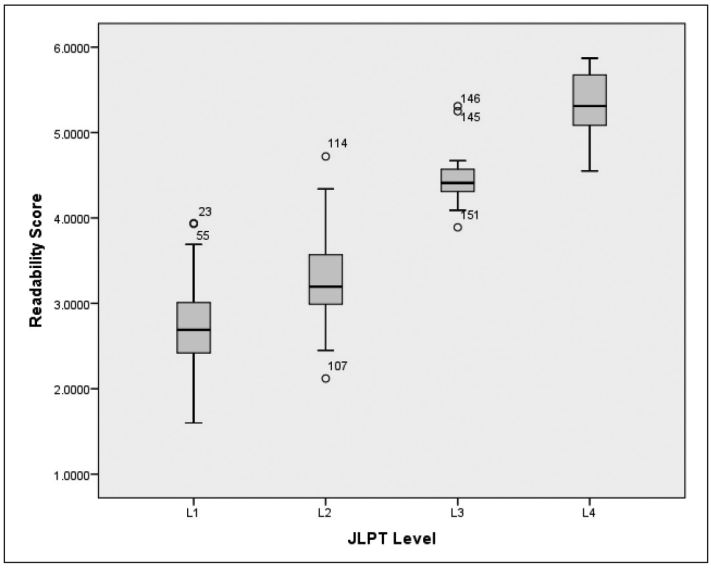


Figure 3. Estimated readability levels of text in JLPT L1 to L4

box plot. The figure shows that the larger the JLPT level-number, the higher the readability score estimated by our formula (Note that a larger JLPT level-number represents a less advanced test level, and a higher readability score means the text in question is relatively easy). One-way analysis of variance showed that the difference among the four groups in terms of their mean numbers is statistically significant ($F(3, 168) = 141.035$, $p < 0.001$).

Another important fact noted for Figure 3 is about the overall tendency of the results. According to the estimated levels worked out by the calculation using our formula, while L1 and L2 show a relatively small gap between them, the gap between L2 and L3 is larger. It is also larger than the gap between L3 and L4. In fact, this conspicuous gap between text passages of L2 and those of L3 has been known among people involved in the test and has been addressed in the new version of JLPT that is divided into 5 levels. The present experiment finally attests to this.

In concluding this section, the estimated scores of text (and accordingly the levels) obtained using our formula with the JLPT reading passages largely correspond to the original JLPT text levels. This confirms the high reliability of the formula gained as a result of the present research.

4 Web system implementation

4.1 Overview

As an attempt to utilize the output of our research presented thus far, we developed a web system that estimates readability scores and levels of texts in Japanese. The system is currently available at http://jreadability.net. We expect that primary users will be practicing teachers of Japanese who need to prepare reading materials for classes to match student levels. Our system also makes several features available that will be helpful not only to teachers, but also to learners.

There are existing systems available for automatic readability assessment such as those developed and introduced in Shibasaki and Hara (2010) and Sato (2011), but they are built on corpora of textbooks written in Japanese for native speakers of Japanese; their formulae consequently assess the readability of Japanese texts on a scale corresponding to Japanese school grades, and as such are not directly applicable to selecting texts for readers of Japanese as a foreign language. On the other hand, our formula is built on levelled corpora of textbooks for learners of Japanese as a foreign language. Therefore, it is expected to be easier to use for both teachers and learners of Japanese.

4.2 Basic system design

In order to calculate readability scores and levels from input texts using our formula, the system needs to first parse the text into words. The input text is split into sentences

by a full-stop symbol and then each of these sentences is further split into words. Since word boundaries in Japanese texts are not indicated by spaces, splitting sentences into words requires an NLP tool called a morphological analyzer. To create a system that is capable of accomplishing this in the same fashion as when we dealt with corpus data to extract lexical information in the process presented in 2.3, we adopted the same set of equipment, MeCab (0.996) and UniDic (2.2.0). With these tools working on the backend, the system extracts five numerical indices from the input text: 1) mean length of sentences, 2) proportion of *kango*, 3) proportion of *wago*, 4) proportion of verbs, and 5) proportion of auxiliary verbs. The system applies these values to our formula to obtain the readability score.

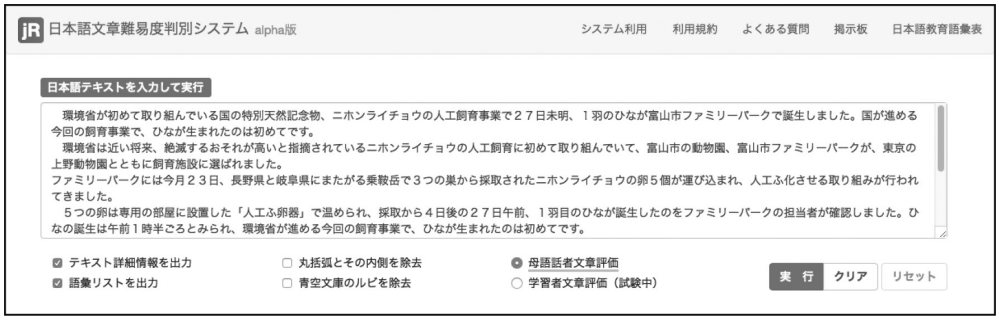


Figure 4. Input form of online readability measurement system

Figure 4 is a screenshot of the text input form of this online system. The user inputs the text and presses the 実行 *jikkō* ('run') button. The results are immediately presented as shown in Figure 5.

As mentioned above, although the calculation of the readability score needs values for only five types of variables, other types of data obtained as a result of the text analysis using MeCab and UniDic are also presented. Among those are the total token number of words and the total type number of words of the input text, as well as the frequency and distribution of vocabulary items of different levels, the frequency and distribution of vocabulary items of different parts-of-speech, and the frequency and distribution of vocabulary items of different types of origin (such as *wago, kango*, and *gairaigo*) as shown in Figure 5.

テキストの概要

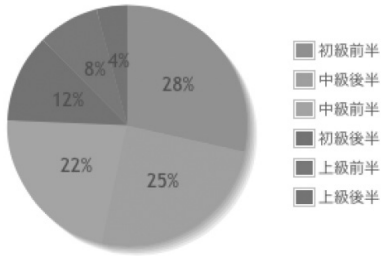
総形態素数（異なり）を表示するには「語彙リストを出力」をオンに

文章難易度	上級前半
リーダビリティ・スコア	2.15
総文数	12
総形態素数（延べ）	458
総形態素数（異なり）	184
総文字数（記号・空白を含む）	741
一文の平均語数	38.17

語彙レベル構成

語彙レベル情報を持っている形態素だけを集計

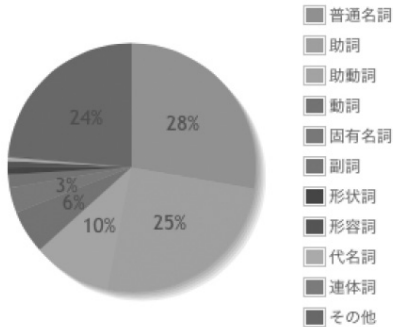
初級前半	55
中級後半	48
中級前半	43
初級後半	23
上級前半	16
上級後半	8



品詞構成

記号類は除外

普通名詞	127
助詞	116
助動詞	47
動詞	26
固有名詞	15
副詞	8
形状詞	4
形容詞	4
代名詞	2
連体詞	1
その他	108



語種構成

定型句は「ありがとう」などを指す

和語	276
漢語	108
外来語	10
混種語	4
定型句	0

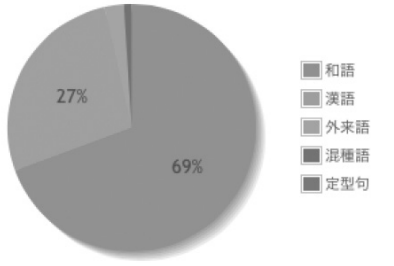


Figure 5. Sample results of readability measurement

4.3 Additional features

The statistics and graphs in Figure 5 are presented on the pane with a tab titled テキスト情報 *tekisuto jōhō* ('Text Information'). There are two other tabs next to it, one of them being テキスト詳細 *tekisuto shōsai* ('Text Details'), and the other 語彙リスト *goi risuto* ('Vocabulary List'). Selecting テキスト詳細 *tekisuto shōsai*, the user is presented with the input text, with its component sentences sequentially numbered and words highlighted with different colors according to the vocabulary level as shown in Figure 6.

Figure 6. Text details

The system has a levelled vocabulary list for learners of Japanese in its background that was produced by Sunakawa et al. (2012). The list consists of six sub-lists of different levels (lower-elementary, upper-elementary, lower-intermediate, upper-intermediate, lower-advanced, upper-advanced).

A similar feature is already available in the reading support system Reading Tutor (Kawamura 1999)[10]. However, while Reading Tutor categorizes vocabulary according to

10 http://language.tiu.ac.jp/index_e.html

the 4 levels of the old version of JLPT, our system uses a more fine-grained six-level vocabulary list, which is expected to be more easily applicable to actual learning environments. Moreover, the system also includes a built-in dictionary with definitions and example sentences. Inside our system, each of the words in the input text is checked to see if it is included in one of the sub-lists of the levelled vocabulary list. If this is the case, the word is highlighted with a color according to the level. When one of the highlighted words is clicked, a pop-up window will appear showing dictionary definitions and example sentences of the word, which were also provided as a product of Sunakawa et al. (2012).

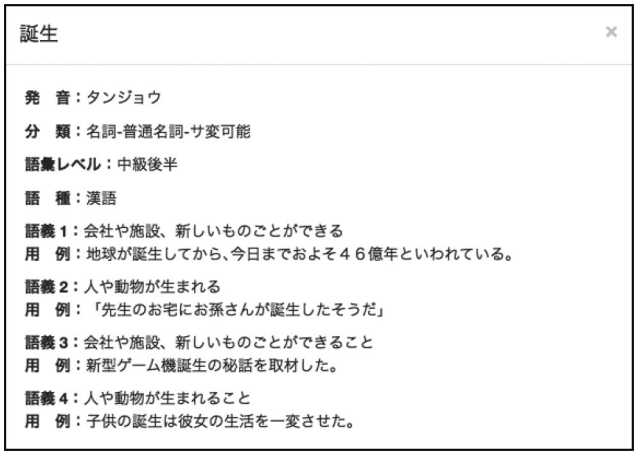


Figure 7. Pop-up window showing definitions and examples

The features presented above will be helpful for teachers of Japanese and also for learners. Other features for learners implemented in this system include the text read aloud with synthesized voice. Once a readability measurement process has been completed, a headphone-like little icon appears above the text input form if the web-browser being used is natively capable of the text read aloud. Clicking on this icon will play the input read aloud text in a synthesized voice.[11]

Our web system has some other features that may also be helpful to researchers of Japanese as a foreign language, or Japanese linguistics in a broad sense. Once a readability measurement process has been completed, a 語彙リスト *goi risuto* ('Vocabulary List') tab appears. Clicking on the tab, the user will be presented with a list of all the words in the input text as in Figure 8. The data are aggregated into their basic forms (e.g., 取り組む *torikumu* ('work on') is the basic form for variations such as 取り組み *torikumi* or

11 As of writing, not many web-browsers support the Web Speech API, which our system depends on for its read aloud functionality. Currently, we have only tested this functionality on Google Chrome, one of a few browsers that support the API.

取り組ん *torikun*) and include the following types of related information, by which the user can sort and rearrange the data on their web-browser. These data are downloadable in the comma-separated value (CSV) format.

Basic form	取り組む	(*torikumu*)
Pronunciation	トリクム	(*torikumu*)
Grammatical category	動詞-一般	(verb-general)
Surface form(s)	取り組ん	(*torikun-*)
Frequency (%)	2（0.44%）	
Vocabulary level	中級後半	(upper-intermediate)


本システムについて　テキスト情報　テキスト詳細　語彙リスト

語彙リスト

結果保存（CSV：Shift-JIS）　結果保存（CSV：UTF-8）

形態素数（延べ）：458　形態素数（異なり）：184

リンクをクリックすると辞書引きを行います。

出現順	基本形	読み	分類	出現順で並べ替え	読みで並べ替え		分類で並べ替え	頻度で並べ替え	語彙レベルで並べ替え
1	環境	カンキョウ	名詞-普通名詞-一般		3	0.66	環境 (3)		中級後半
2	省	ショウ	接尾辞-名詞的-一般		3	0.66	省 (3)		中級後半
3	が	ガ	助詞-格助詞		16	3.49	が (16)		
4	初めて	ハジメテ	副詞		4	0.87	初めて (4)		初級後半
5	取り組む	トリクム	動詞-一般		2	0.44	取り組ん (2)		中級後半
6	で	テ	助詞-接続助詞		2	0.44	で (2)		
7	いる	イル	動詞-非自立可能		3	0.66	いる (3)		初級前半
8	国	クニ	名詞-普通名詞-一般		2	0.44	国 (2)		初級前半
9	の	ノ	助詞-格助詞		27	5.9	の (27)		
10	特別	トクベツ	形状詞-一般		1	0.22	特別 (1)		中級前半
11	天然	テンネン	名詞-普通名詞-一般		1	0.22	天然 (1)		中級後半
12	記念	キネン	名詞-普通名詞-サ変可能		1	0.22	記念 (1)		中級後半
13	物	ブツ	接尾辞-名詞的-一般		1	0.22	物 (1)		中級後半
14	、		補助記号-読点		23	5.02	、 (23)		

Figure 8. Vocabulary list (partial)

4.4 System limitations

The online system allows a user to measure the readability of a Japanese text and also offers many functions useful to educators, learners, and researchers of the language. There are, however, some limitations that the user should note. Firstly, depending on the nature of the input text, the system may not perfectly parse the text and break it into individual words in the most appropriate way. The model data used to devise the readability formula are mostly from textbooks of Japanese. The NLP tools are able to analyze such text easily because it does not have many neologisms; it is mainly composed of words that are well established in the language. The online-system we developed,

however, is required to analyze whatever type of text that the user inputs. Accordingly, the text could be of various types such as a piece of text written especially for elementary learners using quite a limited number and variety of words, or a blog text containing many newly-coined words and/or highly technical terms, which would be difficult for the NLP tools to handle properly.

A second limitation concerns the morphological analysis completed using the NLP tools. Normally, texts in Japanese do not have intervening spaces to make the boundaries of words visible. Morphemes are combined with each other forming larger units, namely words. They are combined with each other with different strengths, making the distinction between morphemes and words less clear. Thus, there are a couple of different ways in which the size of a word-unit is determined for Japanese text. We adopted short-unit words (SUW) among other possible word-units such as long-unit words (LUW) mainly because of the specifications of the NLP tools we used. With SUW, a sequence such as 環境省 *kankyōshō* ('environment ministry') is analyzed as two individual words 環境 *kankyō* ('environment') and 省 *shō* ('ministry') sequentially arranged back-to-back. Some users might find it slightly unnatural since what is referred to by this sequence of two morphemes is just one single concept, or institution. They may prefer to have such a sequence treated as a single (compound) word, rather than as two individual elements.

The latter limitation is, however, mostly at a presentation level, and it does not significantly affect the readability measurement. It is possible that future enhancements and improvements of the NLP tools will enable us to repeat the same set of procedures as described in the present paper to devise a possibly better readability formula based on a different type of unit of words such as LUW. The current formula based on SUW nonetheless has been proven effective as presented in Section 3.

5 Conclusion

This paper presented a method for measuring the readability of Japanese texts using levelled corpora. First, we built a set of six-level corpora using text data extracted from textbooks of Japanese and National Diet meeting transcripts. We examined these corpora both manually and statistically. Then a multiple regression analysis of the results of these examinations was carried out. Among five models produced, we selected the best one and used it to construct our readability formula. The formula was tested using another set of levelled corpora built from 25 years of JLPT tests, and its reliability was confirmed. Our readability assessment formula is original in that it is built upon corpora of textbooks for learners of Japanese as a foreign language and thus it is considered more usable to assess the readability of texts used to teach or learn Japanese than other formulae developed on corpora of texts written for native readers of Japanese.

Moreover, we developed a web-based system using the formula to aid teachers of Japanese in preparing reading materials that match the level of their students. The system is also equipped with many reading-related functionalities that make it helpful not only to teachers, but also learners. Text highlighting according to the fine-grained six-level vocabulary list and pop-up dictionary with word definitions and example sentences are among the functionalities developed especially having learners' convenience in mind. Although a few limitations exist in this system, it is hoped that the system will enable a wide range of people involved in Japanese language instruction to benefit from the present research.

Literature

Den, Y. [伝康晴]. (2009) Tayō na mokuteki ni tekishita keitaiso kaiseki shisutemu yō denshika jisho [多様な目的に適した形態素解析システム用電子化辞書] ("A multi-purpose electronic dictionary for morphological analyzers"). *Journal of the Japanese Society for Artificial Intelligence* 24(5), 640-646.

Flesch, R. (1948) A new readability yardstick. *Journal of Applied Psychology* 32(3), 221-233.

Kawamura, Y. [川村よし子]. (1999) Goi chekkā o mochiita dokkai tekisuto no bunseki [語彙チェッカーを用いた読解テキストの分析] ("Analyzing text for reading using a vocabulary checker"). *Kōza Nihongo Kyōiku* [講座日本語教育] ("Lectures on Japanese Language Teaching") 34, 1-22.

Lee, J. [李在鎬]. (2011) Daikibo tesuto no dokkai mondai sakusei katei e no kōpasu riyō no kanōsei [大規模テストの読解問題作成過程へのコーパス利用の可能性] ("Using corpora to create materials for reading section of large-scale tests"). *Nihongo Kyōiku* [日本語教育] ("Journal of Japanese Language Teaching") 148, 84-98.

Sakai, Y. [酒井由紀子]. (2011) Kenkō igaku jōhō o tsutaeru nihongo tekisuto no rīdabiritī no kaizen to sono hyōka: Ippan muke shippei setsumei tekisuto no yomiyasusa to naiyō rikai no shiyasusa no kaizen jikken [健康医学情報を伝える日本語テキストのリーダビリティの改善とその評価：一般市民向け疾病説明テキストの読みやすさと内容理解のしやすさの改善実験] ("Improvement and evaluation of readability of Japanese health information texts: An experiment on the ease of reading and understanding written texts on disease"). *Library and Information Science* 65, 1-35.

Sakamoto, I. [阪本一郎]. (1964) Bun no nagasa no hijū no sokuteihō: Readability kenkyū no kokoromi [文の長さの比重の測定法：Readability 研究の試み] ("Assessing the weight of sentence length: An attempt to approach the readability"). *Dokusho Kagaku* [読書科学] ("Science of Reading") 8(1), 1-6.

Sato, S. [佐藤理史]. (2011) Kinkō kōpasu o kihan to suru tekisuto nan'ido sokutei [均衡コーパスを規範とするテキスト難易度測定] ("Measuring text readability based on balanced corpus"). *IPSJ Journal* 52(4), 1777-1789.

Shibasaki, H. [柴崎秀子] and Shin-ichiro Hara [原信一郎]. (2010) 12 gakunen o nan'i shakudo to suru nihongo rīdabiritī hanteishiki [12学年を難易尺度とする日本語リーダビリティー判定式] ("The readability formula to predict school grades 1-12 based on Japanese language school textbooks"). *Keiryō Kokugogaku* [計量国語学] ("Mathematical Linguistics") 27(6), 215-232.

Smith, E. A. and J. P. Kincaid. (1970) Derivation and validation of the automated readability index for use with technical materials. *Human Factors* 12(5), 457-464.

Sunakawa, Yuriko, J. Lee, and M. Takahara. (2012) The construction of a database to support the compilation of Japanese learners' dictionaries. *Acta Linguistica Asiatica* 2(2), 97-115.

Tateishi, Y. Y. Ono, and H. Yamada. (1988) A computer readability formula of Japanese texts for machine scoring. *Proceedings of the 12th Conference on Computational Linguistics*, 649-654.

Zhang, Y. and K. Ozeki (1998) Automatic *bunsetsu* segmentation of Japanese sentences using a classification tree. *Language, Information and Computation (Proceedings of PACLIC12)*, 230-235.

要旨（Abstract in Japanese）

「日本語のレベル別コーパスに基づいたリーダビリティ測定」

李在鎬（早稲田大学）、長谷部陽一郎（同志社大学）

本研究ではコーパスを用いて日本語テキストのリーダビリティを測定する方法の開発を行った。これにあたり2種のレベル別コーパスを構築した。リーダビリティ測定用の公式を構築するためのモデルとなるコーパスと、得られた公式の妥当性・信頼性を評価するためのコーパスである。作業は次のように行われた。まず、日本語の教科書と国会議事録から抽出したデータから、6レベルのモデルコーパスを作成した。次に、回帰分析を用いて得られたモデルの中から最も予測精度の高いものを選び、それを元にリーダビリティ公式を構築した。次にこの公式を、25年分の旧日本語能力試験（Japanese Language Proficiency Test; JLPT）の読解問題データから作成した評価用コーパスに適用した。その結果、この公式によって高い精度で日本語テキストのレベル判別が可能なことが明らかとなった。現在、この成果を元にして開発したオンラインのリーダビリティ判定システムを公開している。

8 Analysis of correctness in adverb use in the Japanese composition support system Nutmeg

Bor HODOŠČEK, NISHINA Kikuko, YAGI Yutaka, ABEKAWA Takeshi
Osaka University, Tokyo Institute of Technology, Picolab Co., Ltd., National Institute of Informatics

Abstract
Nutmeg (http://hinoki-project.org/nutmeg/) is a writing support system for Japanese language learners. It can identify probable mistakes in learner writing by classifying expressions based on their frequency distribution across several native Japanese corpora representing various registers. Namely, it divides the corpora into a positive group representing the target register and a negative group representing registers considered to contain inappropriate stylistic features. The purpose of this study is to examine adverb usage within the Japanese academic register and to evaluate the classification results of the system. The system classified 2,919 adverbs extracted from the electronic dictionary UniDic into 'acceptable', 'unacceptable' or 'unknown' classes. These results were compared to an independent classification by an L2 education expert and revealed differences, especially in the low recall performance of the system. Furthermore, adverbs that had a relatively high frequency in the positive corpus set were incorrectly classified as unacceptable. An investigation into these problems revealed that the classification of a lemma according to its different orthographic forms resulted in some of the differences between the human and system evaluations. Because the system classification works at the level of single morphemes, it could not arrive at the right conclusion in instances where the correct unit of classification was a morpheme compound. Other future tasks include classifying the multiple usage and meanings of a single lemma as separate items.

Keywords: academic writing, Japanese language learner, writing support system, register, large-scale Japanese corpora, Scientific and Technical Japanese Corpus, adverbs

1 Introduction

Foreign students enrolled in undergraduate programs in science, technology, engineering, and mathematics (STEM) fields in Japan are often required to write homework assignments, experimental results, graduation theses as well as research papers in academic Japanese. However, courses geared towards beginner and intermediate level learners of Japanese as a second language tend to emphasize the acquisition of spoken language. As a result, learners are inadequately prepared for academic writing and often struggle over how to correctly write academic texts. A common example is choosing the more

appropriate writing form between *de aru* or *da* forms (copular verbs corresponding to 'is/are'). The following short passage, extracted from the Natane learners corpus[1], is written by a native Chinese first-year science student and illustrates the use of spoken words (*sugoku, totemo*) where more appropriate semantically-compatible replacements exist (*kiwamete, hijōni*):

Ex. 1) 日本の婚姻制度は中国と大体同じである。ふるい時代にくらべて、すごく自由、平等になった。日本の法律によると、未成年も結婚できることを了解して、とてもびっくりした。 *Nihon no kon'inseido wa Chūgoku to daitai onaji de aru. Furui jidai ni kurabete, sugoku jiyū, byōdō ni natta. Nihon no hōritsu ni yoru to, miseinen mo kekkondekiru koto wo ryōkaishite, totemo bikkurishita.* 'The marriage system of Japan is almost the same as that of China. It has become much more free and fair when compared to previous eras. I was very surprised to learn that adolescents were allowed to marry under Japanese law.'

An earlier survey based on the Natane error annotations revealed that adverb related errors, similar to those of Ex. 1, were among the most frequent (Yagi et al. 2014a; Yagi et al. 2014b). Adverbs are also an advantageous research target because a relatively smaller set of them are used in academic writing compared to spoken language. Also, the variety of adverb usage greatly differs along register lines.

Furthermore, while sentence-final expressions and function words that connect phrases and sentences are perhaps more indicative of register differences (Srdanović, Hodošček, Bekeš, & Nishina 2009), making them a valid subject of such a study, they have the undesirable property of transcending morpheme and phrasal (*bunsetsu*) boundaries, the latter form of which are not supported in the error classification API used in this research. In most cases adverbs are formed from one morpheme and are thus a more immediately tractable target, with the exception of the adverbs examined in Section 4.2.

Nutmeg's main focus is to assist the process of writing academic Japanese. It analyzes the user's text input and points out any expressions that are inconsistent with the academic writing register, thereby forcing users to reflect on their word choice and in the process, hopefully improve their writing (Yagi, Hodošček, Abekawa, Nishina, & Murota 2014). The current focus is on the identification of errors and not the automatic suggestion of alternative expressions, the development of which are left to future research.

Our research uses language-processing techniques on large-scale corpora, and aims to provide automatic corrections that are appropriate for the register required by the learner. Ng et al. (2014) describe a recent task on grammatical error correction in which many teams made use of machine learning, statistical machine translation and

1 The Natane Learners Corpus contains over 200 essays collected or elicited from L2 Japanese learners and is available from https://hinoki-project.org/natane/. Example 1 is available from http://hinoki-project.org/natane/document/151_a/show.

rule-based approaches to various degrees of success. At the present time, our research only aims for the identification of errors and does not aim to give automatic corrections. Indeed, Hodošček (2011) previously showed that classifying words or expressions for suitability to a genre using only frequency data from large-scale corpora is a feasible and simple approach.

From the viewpoint of the overall effectiveness of writing assistance systems, Yagi et al. (2014a; 2014b) conducted experiments on how learners react to errors shown by the Nutmeg system. The results recommend showing learners only a few example sentences when they correct texts by themselves. Abekawa et al. (2015) analyze tendencies in learner errors related to adverbs from the viewpoint of the academic register by comparing learner errors and adverbs listed in the official vocabulary of the pre-2010 JLPT (Japanese Language Proficiency Test) in order to help develop methods of correction.

From a narrower perspective of error-types, the Chantokun system (Mizumoto 2012) identifies and corrects Japanese case particle usage using a classifier trained by feeding in Japanese learner texts and their associated corrected versions constructed by native speakers on the Lang-8 website[2]. The major difference between Mizumoto (2012) and our present research is that the former employs the use of error-corrected learner corpora for misuse detection, whereas the latter uses native corpora representing varied registers for misuse detection.

From the perspective of research in second language education, Watanabe (2010) analyzes differences in adverb usage within academic reports written by learners and native speakers. The research shows that learners tend to use inappropriate degree adverbs such as the colloquial 一番 *ichiban* 'the most'. Watanabe's research is similar to our present study as both focus on adverb usage, but differs in methodology. Watanabe's research is based on a manual analysis, whereas our research is based on a predictive analysis using corpus data. Moreover, Watanabe (2010) addresses the issue of the L2 Japanese academic writing curriculum as part of the research aim, whereas we developed this system as a tool for self-study.

2 Academic Japanese and official orthographic policy

Textual genres are often described in terms of differences: spoken vs. written, colloquial vs. formal, objective and logical reasoning in essays vs. subjective emotional descriptions. This carries over into the expressions used in those registers. Varieties of language, directly connected to the situation of their use, are referred to as register (Halliday &

2 Accessible from http://lang-8.com/.

Hasan 1976). The present work thus focuses on the appropriateness of learners' expressions to the academic register, with the goal of improving learner composition by conforming to the academic writing style.

Compared with several other languages with clearly encoded spelling rules, a compulsory orthographic policy for the Japanese language has not yet been established. The Ministry of Education, Culture, Sports and Technology (MEXT) has published a standardized manual called *Kōbunsho Sakusei Yōryō* 'Criteria for the writing of official documents' (Kōbunsho Manual) for various orthographies of official government documents (MEXT 2014). However, magazines, newspapers, and other media in Japan are not bound to its rules and tend to have their own more-or-less equivalent but differing internal style manuals. As a consequence of the lack of standardization across these groups and organizations, the orthography of the Japanese writing system remains complicated. In this research we assume that academic communities have their own writing rules that are relatively close to the standard orthographic rules set out by MEXT, but will examine the validity of this assumption in the succeeding sections.

3 Methods and materials

3.1 Corpus selection criteria

The selection of appropriate corpora is essential to realizing the goals of the system: namely, to provide feedback on learners' written errors within the genre of scientific and technical academic writing. The combination of the Balanced Corpus of Contemporary Written Japanese (BCCWJ) (Maekawa et al. 2013), which comprises a diverse range of registers including informal and spoken text, is combined with the Scientific and Technical Japanese Corpus[3] to satisfy these requirements.

Classifying a word as either appropriate or inappropriate for the academic register relies on quantifying its appropriateness with respect to a variety of corpora with well-defined situational characteristics. Therefore, taking those corpora from the BCCWJ and the STJC for which the situational characteristics most closely align with the academic register, one can attempt to infer a word's appropriateness. For words that do not appear or are rarely found within corpora belonging to the academic register, one approach is to just mark them as inappropriate. The approach outlined in this paper takes a different stance in which, in order to identify a word as inappropriate, it is not enough to simply find the word within the set of corpora closest to the academic register, but also necessary to have a separate set of corpora for which the situational

3 The Scientific and Technical Japanese Corpus (STJC) is an ongoing project seeking to form a representative sample of scientific and technical Japanese. It is formed from Japanese language journals and proceedings in such fields as Natural Language Processing, Civil Engineering, Electrical Engineering, and Medicine.

characteristics differ enough from the academic register so that a significant presence of a word within them is taken to be a strong indicator of inappropriate use within the academic register.

For the positive corpus set, the STJC along with the White papers and Law documents media from the BCCWJ were selected, while for the negative corpus set, Yahoo! Q&A, Yahoo! Blogs, and the Minutes of the Diet media, all from the BCCWJ, were selected. While the White paper and Law documents sub-corpora are not strictly academic in nature, many of their situational characteristics are shared with or similar to the STJC. Indeed, previous research has shown that their writing style (Hodošček 2011), and specifically the fact that they can be considered to have undergone editing for consistency with other publications in their fields and are meant for an expert audience, are similar enough to the STJC that we are able to justify their inclusion in the positive set. As the choice of corpora for the negative set was constrained to the corpora available within the BCCWJ, sub-corpora that consistently contain either transcribed speech (Minutes of the Diet) or contain informal writing (Yahoo! Blogs and Yahoo! Q&A) were selected. Finally, the remaining corpora are essential for deciding whether an adverb's relative frequency is exceptionally high or low in the positive and negative corpus sets when compared to 'average' Japanese prose.

3.2 Adverb selection criteria

The list of adverbs examined within this study was compiled from the full list of adverbs within the UniDic morphological dictionary. UniDic was jointly developed alongside the BCCWJ and employs a hierarchical structure that captures the orthographic variation inherent within the Japanese writing system. Morphemes are organized under their lemma, word, and orthographic form to encode the structure shown in Figure 1 (Ogiso et al. 2010), where the adverb *yahari* 'well' is divided into 6 or more (see Table 5 in Section 4.1.2) orthographic forms (やはり , ヤハリ , 矢張り , やっぱり etc.). Unlike a traditional dictionary, the lemma is organized at the level of meaning so that polysemious words having identical word and orthographic forms can be organized under two or more different lemmas.

The question of which orthographic form of a word to choose from when writing is dependent on the particular writing context into which the word is to be inserted. For the purposes of this study, we hypothesize that for the academic register, in which clarity of communication is at a premium, standardization within most academic domains will mean that there is in general a single preferred way of writing a word. We therefore choose to prioritize the analysis of words at their orthographic form level first, and then to select several examples to be additionally analyzed from the lemma level. There are merits and demerits to both approaches. As mentioned in Section 2 above, we need to

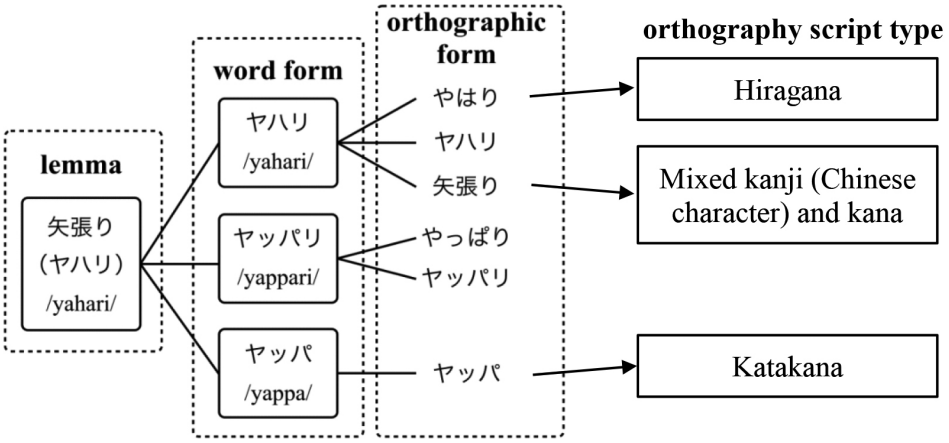


Figure 1. Word and orthographic forms of the adverb 矢張り *yahari* 'just as I thought' within the layered structure of UniDic.

take into consideration official orthography as well, which is not guaranteed to match our data-driven results.

In this research, we use the adverb list extracted from UniDic, which includes 7,432 orthographic forms of adverbs. However, due to the classification API not distinguishing between words having different pronunciation but sharing orthographic and lemma forms, 29 entries are removed, leaving 7,403 entries. Among these, we further exclude 878 orthographic forms which could not be found in any corpus, and 3,606 forms of onomatopoeic words, leaving the final number of adverbs used in the evaluation at 2,919. We then apply the register misuse identification method explained below to classify each into acceptable, unacceptable or unknown classes. In order to evaluate these predictions, we requested a Japanese language education expert separately evaluate the list.

3.3 Comparison between positive and negative corpus sets

Figure 2 shows the differences within the top 30 most frequent adverbs in the whole corpus set, the positive corpus set and the negative corpus set. The values in the figure represent PPM (parts-per-million), which corresponds to how many times an adverb occurred within a million-word long span of text. Based on the figure and statistics from the whole dataset, we can make several observations on three levels: adverb variety, magnitude of use, and adverb preference.

Firstly, the variety of adverbs employed differs greatly between the positive and negative sets: 1,023 used in the positive set compared with 2,253 used in the negative corpus set and 2,887 in the whole set. Just 73 of the most frequent adverbs in the positive

corpus set account for more than 90% of all adverbs occurrences when compared with 174 adverbs for the negative set and 215 for the whole set. Secondly, overall adverb use is 4.6 times more frequent in the negative corpus set than in the positive corpus set. While the whole corpus set displays an overall higher adverb use that is 1.9 times higher than the negative corpus set, the negative corpus set has a more skewed distribution of high frequency adverbs within the top 30. Thirdly, among the top 30 adverbs in all sets, the positive corpus set features the most distinct variety and ordering of adverbs when compared with the negative and whole sets, which follow a similar pattern. When looking at the overall overlap of adverbs, all but 40 adverbs from the positive set are contained within the negative set. Among these low-frequent adverbs are a few like 別して *besshite* 'especially' that are appropriate for the academic register. However, most others are onomatopoeic adverbs which had likely originated within natural language processing research papers dealing with various aspects of onomatopoeia.

When comparing between the top 30 adverbs across the three corpus sets, we found that 18 out of the top 30 adverbs were common to the whole and positive corpus sets. Adverbs missing from the positive set include ones commonly used in informal speech, while those particular to the positive set include formal spoken adverbs that find their use in lectures and meetings such as *tatoeba* 'for example', *mottomo* 'the most', and

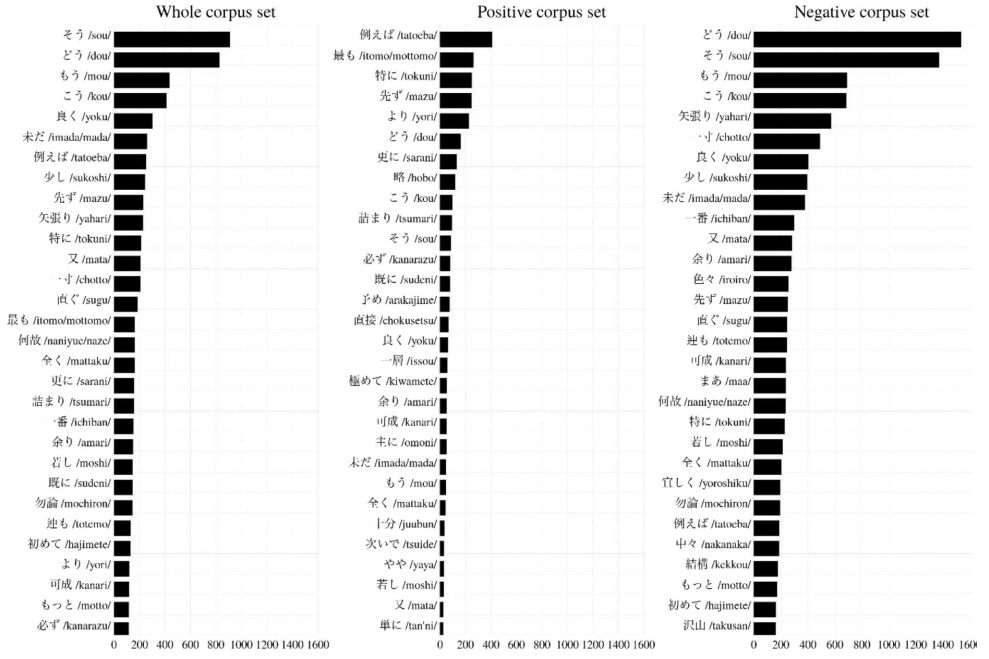


Figure 2. PPM of the top 30 adverbs in all corpora and within the positive and negative corpus sets.

tokuni 'especially'. As for the negative corpus set, 23 adverbs were shared with the whole corpus set, with the rest including adverbs such as *iroiro* 'many', *totemo* 'very', *yoroshiku* 'please', *nakanaka* 'hardly', *kekkō* 'much', *zutto* 'always, more', all of which are commonly used in everyday speech. Additionally, there were 17 adverbs not common to the positive and negative sets, including *ichiban* 'first place', *motto* 'more', *chotto* 'a bit', *mochiron* 'obviously', *yahari* 'as I thought', all of which were commonly used within learner writing and are a common source of signaling the wrong register.

3.4 Classification of adverbs with respect to suitability in the academic register

Nutmeg shows learners whether input words are acceptable or unacceptable in the academic register. Hodošček (2011) and Hodošček & Nishina (2011) proposed and described details on the basic idea of using the chi-square test on corpus data to identify expressions salient to a particular register. As the method described is the same one used in this research, we will only briefly explain the classification procedure by using two examples. The first example is *mottomo* or *itomo* 'the most', which is classified as an acceptable adverb for the academic register. The second is *chotto* 'a bit', which is classified as unacceptable. Table 3 shows the statistical data of the lemma and the orthographic forms of *mottomo* and *chotto* among adverbs from the whole corpus set, the positive corpus set and the negative corpus set. The system determines whether a word is acceptable or unacceptable for the academic resister by calculating how far its frequency within the target register deviates from the frequency distribution within all corpora by using the chi-square (χ^2) test. A word will be classified as acceptable if both the frequency of the positive set is significantly *higher* than that of all corpora and that of the negative corpus set is significantly *lower* than that of all corpora.

Table 3: Comparing classification results between the positive and negative corpus sets.

Lemma	Orth. Form	System Verdict	Frequency Whole	Frequency Positive	Frequency Negative	PPM Whole	PPM Positive	PPM Negative
最も mottomo/ itomo	最も	AC	17,492	7,189	1,274	121.83	**241.67**	**41.23**
	もっとも	UK	5,449	618	228	37.95	20.78	7.38
一寸 chotto	ちょっと	UA	27,677	193	13,975	192.77	**6.49**	**452.24**
	チョット	UK	277	2	237	1.93	0.07	7.67

Note: AC=acceptable, UA=unacceptable, UK=unknown; chi-square test: P(chi-square value > 17.275; $\alpha = 0.1$)

Conversely, it will be classified as unacceptable if both the frequency of the positive set is significantly *lower* than that of all corpora and that of the negative corpus set is

significantly *higher* than that of all corpora. In the examples, the lemma *mottomo* (最も) has two different orthographic forms: 最も, which is written in kanji, and もっとも, which is written in hiragana. The kanji form is classified as acceptable for the academic register, but the hiragana form is classified as unknown because no significant difference was observed. Similarly, the lemma *chotto* (一寸) also has two different orthographic forms: *chotto* (ちょっと) written in hiragana is classified as unacceptable, while *chotto* (チョット) written in katakana (the counterpart of the hiragana syllabic character pair) is classified as unknown, due to the χ^2 test not finding a significant difference between the opposing corpus sets.

3.5 Results of the differences in system and L2 expert judgments on the adverb list

Table 1 shows the number of adverbs classified as either unacceptable, acceptable or unknown by both the language expert and the system. For the purposes of this evaluation, classifications of the class unknown by the language expert were treated as insufficient grounds for identifying an adverb's use as unacceptable. We therefore treat these adverbs as acceptable for the purposes of the evaluation.

Table 1: Differences in system and L2 expert judgments on subset of UniDic adverb list for chi-square significance cutoff 0.1.

	Unacceptable	Acceptable	Unknown	Total
Expert	445	196	2,278	2,919
System	74	5	2,840	2,919

The system achieved a precision of 0.670, recall of 0.029, and F1 score of 0.055 when evaluated against expert's classifications. While the precision was shown to be better than a random baseline, the recall was very low, as the system identified only 74 adverbs as unacceptable, while the expert identified 445. Additionally, the system only identified 5 adverbs as particularly salient to the academic register, which also differs greatly to the 196 adverbs classified as acceptable by the expert. One reason for this is that there are few adverbs that are truly particular to the academic register, but many are acceptable simultaneously in both the academic register and other registers not represented in the positive corpus set. It should be noted, however, that the above performance numbers treat acceptable and unknown evaluations as the same—after all, the purpose of the system is to identify incorrect use, not point out if an expression is particularly well chosen. Finally, as the number of acceptable adverbs are quite low in either evaluation, we are able to presume that the overall use of adverbs in academic fields is quite limited.

4 Analysis of adverbs with erroneous classification results

In order to better understand the differences in judgement between the system and L2 expert, this section takes a detailed look at the adverbs where the judgments between the system and expert differed. As shown in Table 4, the two differ in several aspects. In order to analyze these differences, we examine the following two items in detail:

1) Complex lemma structure with several orthographic variations
2) Treatment of high frequency adverbs including KOSODO compounds

Table 4: Classification of different orthographic forms within frequently occurring adverbs.

Adverbs	System	L2 Expert	PPM Whole	PPM Positive	PPM Negative
例えば（例えば）*tatoeba*	UK	AC	14.4	330.6	146.4
例えば（たとえば）*tatoeba*	UK	AC	10.7	62.5	48.7
先ず（まず）*mazu*	UA	AC	22.4	237.6	154
先ず（先ず）*mazu*	UA	AC	0.5	2.5	7.9
特に（特に）*tokuni*	UK	AC	16.5	228.5	228.3
特に（とくに）*tokuni*	UK	AC	4.7	12.3	8.1
どう*dō*	UA	AC	43.7	154.6	1557.9
更に（さらに）*sarani*	UA	AC	14.3	99.6	98.8
更に（更に）*sarani*	AC	AC	1.5	27.4	18.7
こう（こう）*kō*	UA	AC	40.7	93.8	213.7
詰まり（つまり）*tsumari*	UA	AC	15.7	90.6	92.1
詰まり（詰まり）*tsumari*	UA	AC	0	0.1	0.6
必ず（必ず）*kanarazu*	UA	AC	10.5	71.3	122.3
必ず（かならず）*kanarazu*	UK	AC	0.8	0.9	3
そう（そう）*sō*	UA	AC	87.6	68.2	1368.3
良く（よく）*yoku*	UA	AC	28.3	55.8	385.8
良く（良く）*yoku*	UA	AC	0.9	3.4	30.1
最も*mottomo*	AC	AC	121.8	50.0	41.9
もっとも *mottomo*	UK	AC	37.9	4.3	7.7
可成（かなり）*kanari*	UA	AC	11.7	49.5	244
可成（可成）*kanaru*	AC	AC	0	0.4	0
より（より）*yori*	AC	UA	0.1	46.9	60.0
もう（もう）*mō*	UA	UA	43.3	44.4	712.7
予め（あらかじめ）*arakajime*	AC	AC	27.6	55.0	11.8

Adverbs	System	L2 Expert	PPM Whole	PPM Positive	PPM Negative
予め（予め）*arakajime*	AC	AC	6.3	4.0	3.7
（一層）一層 *issō*	AC	AC	25.2	11.5	20.5
（一層）一そう *issō*	UA	UK	0.2	0.0	0.1
矢張り（やはり）*yahari*	UA	UA	13.7	10.8	336.1
矢張り（矢張り）*yahari*	UK	UA	0.1	0.5	0.6
（主に）主に *omoni*	AC	AC	28.5	10.5	22.1
（主に）おもに *omoni*	UK	AC	3.4	0.1	0.8
（次いで）次いで *tsuide*	AC	AC	11.5	6.3	2.4
（次いで）ついで *tsuide*	AC	AC	2.7	0.3	0.3
（依然）依然 *izen*	AC	AC	14.9	5.3	7.7
（総じて）総じて *sōjite*	AC	AC	3.2	1.3	1.8
（総じて）そうじて *sōjite*	UK	UK	0.0	0.0	0.0
（概して）概して *gaishite*	AC	AC	2.7	0.6	0.9

Note: AC=acceptable, UA=unacceptable, UK=unknown

4.1 Complex lemma structure with several orthographic variations

As mentioned in Section 3, the lexical data used in the system is based on a subset of UniDic, namely the lemma and orthographic base forms of morphemes extracted using MeCab, which is an open source Japanese morphological analysis engine. We analyzed *yoku* and *yahari*, which are adverbs that both have a number of orthographic forms and were classified as 'unacceptable'. In order to explain the reasons behind this classification result, it is necessary to examine them from two viewpoints: lemma and orthographic form.

As mentioned in Section 3, the Japanese orthographic system has not yet been standardized. The Kōbunsho Manual is regarded as a sort of standard for writing Japanese official documents. The manual generally recommends hiragana notation for describing adverbs. As such, we assume that adverbs in the positive corpus set (White papers and Law documents, specifically) tend to conform to these standard guidelines. Hence, we will refer to the manual when analyzing the lemmas of 良く *yoku* and 矢張り *yahari* in our data.

4.1.1 良く *yoku 'well'*

The frequency distribution of the lemma 良く *yoku* is 300.5 PPM in the whole corpus, 62.9 PPM in the positive corpus and 404.0 PPM in the negative corpus. As the

frequency in the negative corpus is significantly higher than in the whole corpus, and the frequency in the positive corpus significantly lower than in the whole corpus, it is classified as unacceptable.

However, when considering the frequency of the lemma 良く *yoku* within the positive corpus set, we find it ranks 15th most frequent and cannot thus be considered low. Usage of *yoku* may be considered unacceptable within the academic register depending on the semantic context it is used in. On conferring the Digital Daijisen Japanese dictionary (Matsumura et al. 1998) and other dictionaries, we assume that *yoku* has six different meanings:

1. Frequently, in quantity. Synonyms: しばしば *shibashiba* 'frequently', しきりに *shikirini* 'often'
2. Adequately, enough. Synonyms: 十分に *jūbun-ni* 'adequately', 徹底して *tettei-shite* 'thoroughly'
3. [Subjective use] With high ability. Synonyms: 上手く *umaku* 'well'
4. Highly, widely. Synonyms: 極めて *kiwamete* 'extremely', 非常に *hijōni* 'very', 高度に *kōdo-ni* 'to a high degree'
5. Completely. Synonyms: 十分に *jūbun-ni* 'thoroughly'
6. [Subjective use] Favorably. Synonyms: 好意をもって *kōi wo motte* 'with goodwill'

As meanings 1, 2, 4, and 5 of *yoku* can be used in objective contexts, and express the meaning of a high frequency, their use is permissible in academic documents, although paraphrasing them with other expressions is still preferable. At the current stage, the system cannot clearly distinguish between these meanings. Therefore, without a way of automatically disambiguating the exact sense used within the text, the system can only point to the objective uses of *yoku* as found in the positive corpus set, and these can serve as examples for the learner to reflect upon. For example, given the sentence この計画はよく考えられている *Kono kēkaku wa yoku kangaerarete iru* 'This plan is well thought out', it is possible to suggest the following alternative: この計画は十分に考えられている *Kono kēkaku wa jūbun ni kangaerarete iru* 'This plan is sufficiently thought out'.

As for the orthographic frequency, 良く *yoku* has a PPM of 0.9, while よく *yoku* has a PPM of 28.3 in the whole corpus. The figures are 3.4 PPM and 55.8 PPM respectively in the positive corpus set, and 30.1 PPM and 385.8 PPM respectively in the negative corpus set. *yoku* is classified as unacceptable in all cases. The hiragana orthographic form よく *yoku* is used 92.7% of the time in the negative corpus set, and 96.7% of the time in the whole corpus set. On the other hand よく *yoku* is only used 62.1% of the time in the positive corpus set. For comparison, the Kōbunsho Manual mandates the use of よく *yoku*.

These results show that academic documents use more mixed orthography even though the Kōbunsho Manual does not condone such use. For documents of corpora

in which such standards do not apply, adverbs are also written using kanji (Chinese characters) and katakana.

On the other hand, an examination from the perspective of an expert in L2 Japanese language education classified these cases as acceptable within the academic register. It is therefore reasonable to assume that the choice of which orthographic form to use in academic writing depends on the intended meaning.

4.1.2 矢張り yahari 'as I thought'

As can be seen in Table 5, the lemma *yahari* can be written using 14 different orthographic forms. It should be noted that the last six all represent rare variations occurring less than ten times in the whole corpus set. In total, there are five word forms: *yahari* (やはり、矢張り), *yappari* (やっぱり, やつぱり, ヤッパリ), *yappa* (やっぱ), *yappashi* (やっぱし), and *yapa* (やぱ). As the system classifies at the orthographic form level, we are able to compare the results between different orthographic varieties of the same lemma.

The classification results for the lemma 矢張り give two orthographic forms (やはり, やっぱり) for which the verdict is unacceptable for the academic register, with the

Table 5. Orthographic form variations and their associated system classifications of the lemma 矢張り ordered according to their PPM in the whole corpus set.

Orthographic Base	System Verdict	Frequency Whole	Frequency Positive	Frequency Negative	PPM Whole	PPM Positive	PPM Negative
やはり	**UA**	**19,688**	**335**	**9,997**	**137.13**	**11.26**	**323.51**
やっぱり	**UA**	**11,130**	**48**	**6,357**	**77.52**	**1.61**	**205.72**
やっぱ	UK	1,502	10	1,210	10.46	0.34	39.16
矢張り	UK	107	16	19	0.75	0.54	0.61
やっぱし	UK	99	2	53	0.69	0.07	1.72
ヤッパリ	UK	49	13	32	0.34	0.44	1.04
やぱ	UK	36	-	35	0.25	0.00	1.13
やつぱり	UK	22	-	-	0.15	0.00	0.00
矢っ張り	UK	8	3	-	0.06	0.10	0.00
矢つ張り	UK	5	-	-	0.03	0.00	0.00
矢ッ張り	UK	4	-	-	0.03	0.00	0.00
やッぱり	UK	3	-	1	0.02	0.00	0.03
矢っ張	UK	1	-	-	0.01	0.00	0.00
矢つ張	UK	1	-	-	0.01	0.00	0.00
Total		32,655	427	17,704	227.45	14.35	572.92

Note: AC=acceptable, UA=unacceptable, UK=unknown

rest classified as unknown. The lemma, as a whole, occurs at a rate of 572.92 PPM in the negative corpus set, 14.35 PPM in the positive corpus set and 227.45 PPM in the whole corpus. Thus, according to the system classification and large discrepancy between PPM rates, the adverb 矢張り is clearly not appropriate for use in the academic register.

The Kōbunsho Manual recommends the use of the hiragana やはり over the Chinese character (kanji) variant 矢張り. The orthographic variation やはり is used in 79.02% of the positive corpus set, 60.97 % of the whole corpus set, and 56.48% of the negative corpus set. While the Minutes of the Diet sub-corpus is a part of the negative corpus set, it is edited from transcribed speech data, a process which strictly follows the governmental guidelines and, as such, contains less orthographic variations than its sibling corpora of Yahoo! Q&A and Yahoo! Blogs.

In conclusion, we find that the hiragana variant of the orthographic form of the lemma *yahari* most commonly appears in the positive corpus set, which is also the form recommended by the Kōbunsho Manual. However, the system classified even this usage as unacceptable.

4.2 KOSOADO (こそあど) demonstrative words

The Japanese KOSOADO demonstratives have either *ko, so, a,* or *do* as the first syllable and are most commonly represented by the adverbs *kō*, *sō*, *ā*, and *dō*. These adverbs occur frequently in the whole corpus set and, with the exception of *ā*, also occur frequently in the positive corpus set. However, the system classifies them all as unacceptable for the academic register. Across the whole corpus set as well as the negative corpus set, *sō, dō* and *kō* are respectively the first, second and fourth most frequent adverbs. Even in the positive corpus set, *sō, dō* and *kō* are the eleventh, sixth, and ninth most frequent adverbs. The existence of *kono-yō-ni*, *sono-yō-ni*, and *dono-yō-ni*, formal counterparts to *kō*, *sō*, and *dō,* within the STJC is a possible reason for their relatively high rank. Finally, though less frequent than the rest, *ā* does appear in the negative corpus set, while its use within the positive corpus set was observed only within linguistic examples or language data in scientific articles and are otherwise absent from the main body of text. The inappropriate use of *ā* can also be found in the error annotations of the Natane learner corpus. The present system is able to advise learners that *ā* is unacceptable in academic documents.

4.2.1 こう *kō*

The lemma こう *kō* has no orthographic variation other than its hiragana form. As shown in Table 6, its frequency is much higher in the negative corpus set (687.7 PPM) than in the positive corpus set (97.8 PPM). As such, the lemma こう *kō* is classified as unacceptable for academic writing. However, if we look at the frequency of the

compound adverbs, we find that the overall frequency in the positive corpus set is higher than in the negative corpus set. In order to uncover the reasons behind this shift in relative frequency between the positive and negative corpus sets, we analyze the usage of some of these compounds.

The compound adverb *kōshite* and compound noun modifier *kōshita* frequently appear in spoken language as well as written texts. These are paraphrased as *konoyōni* and *konoyōna* in the formal and academic texts as shown in the examples below. In addition, *kon'nani* and *kon'na* are casual expressions not found in the positive corpus set. Hence, it is possible to recommend the compound *konoyōni* for use in the academic register.

The following examples (2-4) show compound adverbs found in both the positive and negative corpus sets.

Ex. 2) たくさん問題をこなしているうちに，パターンが身につきます．こうして身についたパターンは，忘れることがなくなり，本当の学力につながりますよ． *Takusan mondai wo konashite iru uchi ni, patān ga mi ni tsukimasu. Kō-shi-te mi ni tsuita patān wa wasureru koto ga nakunari, hontō no gakuryoku ni tsunagarimasu yo.* 'You will never forget the patterns you have mastered this way, and this will lead to real learning.' (Yahoo! Q&A: OC12_05972)

Ex. 3) 今回のこうした不幸な事件を引き起こした大きな原因は、やはり外交上の問題があったと思うのです. *Konkai no kō-shi-ta fukō na jiken wo hikiokoshita ōkina genin wa, yahari gaikō-jō no mondai ga atta to omoun desu.* 'The main reason which caused such an unfortunate accident on this occasion is due to diplomatic problems.' (Minutes of the Diet: OM21_00010)

Ex. 4) 「今の世の中では，大学に進むのが当たり前だから」と答える親は極めて少ない。このように，親の側には，大学教育の役割について理想的なイメージがあるといえる。 *"Ima no yononaka dewa, daigaku ni susumu no ga atarimae dakara" to kotaeru oya wa kiwamete sukunai. Kono-yō-ni, oya no gawa ni wa, daigaku kyōiku no yakuwari ni tuite risō teki na imēji ga aru to ieru.* 'There are extremely few parents who would answer that "it is natural for their children to go to university in today's world". From this we can say that parents have an ideal image about the role of university education.' (White paper on public lifestyle: OW2X_00000)

Next, the *kōshite* in examples 5 and 6 is used as a direct deictic and not as a contextual demonstrative, making its use unsuitable for academic writing. *Kōshite* in example 5 indicates the way in which the speaker wants the food to be cut. Similarly, *kōshite* in example 6 indicates an ambiguous object, which cannot be determined from the context.

Ex. 5) 食べやすい大きさにこうしてちぎってください。 *Tabeyasui ōkisa ni kōshite chigitte kudasai.* 'Tear it into bite size pieces in this way, please.' (Nishida et al. (2003). Ryōri kyōji hatsuwa no kōzōkaiseki [Structural analysis of recipe

instructions utterances]. *Proceedings of the 9th Annual Conference of the Association of Natural Language Processing*, 601-604.)

Ex. 6) 「もっとこうしてほしい」っていうのは彼に伝えた方がいいと思います。 *'Motto kōshite hoshi' tte iuno wa kare ni tsutaeta hou ga ii to omoimasu.* 'I think you should tell him "I want you to do it more in this way"' (Yahoo Q&A: OC09_06241)

We suggest that the deictic usage of *kō*—including in the compound adverbs as mentioned above—should be discouraged in academic writing. Consequently, we have to divide the usages of *kō*, including its compound variants, into those suitable for academic writing and those unsuitable based on these observations.

It is possible to say that *kōshita* and *kōshite* are acceptable because of their frequent use in the STJC corpus. We have to take into account both a word's current usage tendencies as well as its normative uses.

Table 6: Frequency of *ko* as part of compound expressions.

Adverb	Expression Type	Frequency Whole	Frequency Positive	Frequency Negative	PPM Whole	PPM Positive	PPM Negative
こう *kō*	SM	59,100	2,908	21,190	411.4	97.6	685.7
こうして *kōshite*	CM	6,407	260	590	44.6	8.7	19.1
こうした *kōshita*	CM	14,390	2,225	1,120	100.2	74.8	36.2
こういう *kōiu*	CM	18,788	41	12,096	130.8	1.8	391.4
こう言う *kōiu*	CM	390	3	139	2.7	0.1	4.5
こう云う *kōiu*	CM	34	0	7	0.2	0.0	0.2
このような *konoyōna*	CM	21,394	7,588	2,196	149.0	255.1	71.1
この様な *konoyōna*	CM	229	48	123	1.6	1.6	4.0
このように *konoyōni*	CM	11,406	3,308	1,406	79.4	111.2	45.5
この様に *konoyōni*	CM	58	19	20	0.4	0.6	0.6
このようにして *konoyōnishite*	CM	1,055	375	25	7.3	12.6	0.8
こうやって *kōyatte*	CM	784	2	268	5.5	0.07	8.7
こんな *kon'na*	SM	28,860	110	10,304	200.9	3.7	333.4

Note: SM=single morpheme, CM=compound morphemes

On the other hand, *kōyatte* and *kon'na* are scarcely found in the positive corpus set. Hence, we will add the former two compound words into the list of acceptable adverbs, but exclude the latter two compound words.

4.2.2 そう *sō*

The lemma *sō* has the highest frequency within all corpora. Additionally, it is also frequent in both the positive and negative corpus sets. However, our system classifies そう *sō* as unacceptable, even though its frequency is as high as that of こう *kō*. Next, comparing the compound words of *sō* and *kō*, we find that *kō* tends to occur more frequently in the positive corpus set, and *sō* in the negative corpus set. As can be seen from Table 7, the PPM value of *sō* is relatively higher for all the items in the negative corpus set.

Table 7: Frequency of *sō* showing the conjugated compound adverbs *sō-iu* and *sō-itta* used within the positive corpus set.

Adverb	Expression Type	Frequency Whole	Frequency Positive	Frequency Negative	PPM Whole	PPM Positive	PPM Negative
そう *sō*	SM	130,824	42,449	2,521	910.6	84.7	1,373.7
そうして *sō-shi-te*	CM	2,898	604	21	20.2	0.7	19.5
そうした *sō-shi-ta*	CM	8,182	1,186	302	57.0	10.2	38.4
そういう *sō-iu*	CM	32,907	17,176	79	229.0	2.7	555.8
そう言う *sō-iu*	CM	1,244	293	5	8.7	0.2	9.5
そう云う *sō-iu*	CM	76	9	1	0.5	0.0	0.3
そのような *sono-yō-na*	CM	7,847	1,637	1,433	54.6	48.2	53.0
その様な *sono-yō-na*	CM	117	92	8	0.8	0.3	3.0
そのように *sono-yō-ni*	CM	1,915	607	78	13.3	2.6	19.6
その様に *sono-yō-ni*	CM	22	15	0	0.2	0.0	0.5
そのようにして *sono-yō-ni-shi-te*	CM	180	24	7	1.3	0.2	0.8
そうやって *sō-ya-tte*	CM	948	199	4	6.6	0.1	6.4
そんな *son'na*	SM	45,427	13,689	152	316.2	5.1	443.0

Note: SM=single morpheme, CM=compound morphemes

Although these idiomatic patterns are found in academic texts, they are relatively less frequent than words in the *kō* group. Words in the *sō* group are noted for their use in anaphoric expressions such as *'A ga B de aru bāi, ippō, A ga sō de nai bāi'* (In case A is B, and, on the other hand, in case A is not so).

Ex. 7) ペアが含まれるなら真、そうでないなら偽. *Pea ga fukumareru nara shin, sōdenainara gi.* 'If the pair is present, it is true, and if it is not so, then it is false.' (STJC: Murawaki, Y. & Kurohashi, S. (2007). Jōhō bunseki no tame no jutsugo kōzō wo mochiita dōteki ontorojī kōchiku [Construction of a dynamic ontology for information analysis using predicate structure]. In *Proceedings of the 13th Conference of the Association of Natural Language Processing* (pp. 867-870)).

Sōdenainara in this example paraphrases the previous expression *pea ga fukumareru,* which is its general function. On the other hand, substitution with *sonoyōdenainara* is unacceptable for reasons of syntax, although this may be substituted with the compound word *sonoyōni* which is more academic and formal than *sō* as a single morpheme. For example, it is possible to rewrite the expression *sō kaishaku dekiru* 'it is possible to interpret in that way' into the expression *sonoyōni kaishaku dekiru* in academic discourse. Hence, we are able to say that expressions such as *sō, sōitta* and *sōiu* are rather uncommon in academic discourse. The following examples extracted from the positive corpus set (examples 8 and 10) and the negative corpus set (example 9) illustrate these general observations.

Ex. 8) 最後の第九グループは，脂肪族化合物でアミノ基を有する場合の挙動を探ったものであるが，末端にある場合と，そうでない場合で多少反応性が異なり，場合によっては阻害性も発現する傾向がある. *Saigo no daikyū grūpu wa, shibōzoku kagōbutsu de aminoki wo yūsuru bāi no kyodō wo sagutta mono de aru ga, mattan ni aru bāi to, sō de nai bāi de, tashō hannōsē ga kotonari, bāi ni yotte wa sogaisē mo hatsugen suru keikō ga aru.* 'The last ninth group is an exploration of the behavior of possessing an amino group with an aliphatic compound. The reactivity is different in the occasion in the end and the occasion which is not so. The obstruction also tends to be manifested by a case.' (STJC: Watanabe O., & Nagai K.. (2000). Effect of Additive Reagents on the Reactivity of Lacquer Tree Paint. *Journal of the Chemical Society of Japan*, (3), 211-216.)

Ex. 9) 「おはようメール」がたまに届いたりしてました。ですが、最近はそういったメールが入ってきません。*Ohayō mēru ga tamani todoitari shite imashita. Desu ga, saikin wa sō-itta mēru ga haitte kimasen.* 'I had been occasionally receiving "good morning mails". But I have not received such mails recently.' (Yahoo Q&A: OC09_06528)

Ex. 10) 上記のように極めて短期の需給見通し等の場合にはそのようなおそれがあるとみられる。*Jōki no yō ni kiwamete tan'ki no jikyū mitōshi nado no bāi ni-wa sono yōna osore ga aru to mirareru.* 'It seems risky in case of such an extremely short-term supply and demand outlook as described above.' (Anti-trust white paper: OW3X_00120)

Having observed the corpora, usage of *so* in compound adverbs and in adjectival expressions such as *sōshite, sōshita, sōitta, sonoyōni, son'nani,* and *son'na* is extremely frequent in the negative corpus set compared to the positive corpus set. Therefore, we have to admit that anaphoric usage of *sō* is permitted in academic discourse. Even though the adverb *sō* is classified as unacceptable according to the system classification, the human evaluator classified the anaphoric usage of *sō* with compound words such as *so de areba* 'if it is so' and *so de nakereba* 'if it is not so' as acceptable. Consequently, we need to re-examine the system's focus on processing morphemes in isolation; expanding the unit size and taking into account the compound expressions is a promising avenue for increasing the accuracy of the system.

4.2.3 どう *dō*

The basic usage of the lemma *dō* is as the interrogative word of a sentence. It is ranked as the second most frequent in the whole corpus set, the sixth in the positive corpus set and the first in the negative corpus set (see Figure 2). With respect to PPM values, however, *dō* is most frequent in the negative corpus set; its frequency in the positive corpus set is significantly smaller than the norm to mark it as inappropriate for the academic register. The reason for this is clear if we analyze words co-occurring with *dō*: the frequency of the compound word *donoyōni* in the positive corpus set is higher than in the negative corpus set.

As shown in example 11, some adverbial *dō* appear as a part of constructions where they are followed by a verb, *ka* and a closing phrase such as *dō miru ka* or *dō kangaeru ka*. As shown in Table 8, the frequency of *ka dō ka* is highest in the positive corpus set, which covers approximately 67% of all instances of *dō*. However, the system classified it as unacceptable for the academic register. Instead of *ka dō ka, ka ina ka* is often used in academic fields, and the system has classified it as acceptable for the academic register. From this, we must admit *ka dō ka* as an alternative choice for learners, particularly since *ka dō ka* is relatively frequent in academic documents. We still recommend using *ka ina ka* as the first choice.

Ex. 11) 「どう思うか教えて下さい。」 *Dō omou ka oshiete kudasai* 'Tell me what you think about it.' (Yahoo! Q&A: OC09_13396)

Ex. 12) 「この意見に対してしてどう思います？」*Kono iken ni taishite dō omoimasu?* 'What do you think about this opinion?' (Yahoo! Q&A: OC09_14216)

Ex. 13) 業界ごとの市場規模を調べるにはどういう手段がありますか？ *Gyōkai goto no shijō-kibo wo shiraberu ni wa dōiu shudan ga arimasu ka?* 'What means are availwable for researching the market size of each industry?' (Yahoo Q&A: OC03_02066)

Ex. 14) 明日初めてロンドンに行くのですがどういった服装でいけばいいですか？ *Ashita hajimete rondon ni iku no desu ga dōitta fukusō de ikeba ii desu ka?* 'I will visit London for the first time tomorrow, so what kind of clothes should I wear?' (Yahoo Q&A: OC13_02305)

Examples 11 and 12 illustrate the usage of *dō* in conversations. Example 13 illustrates the usage of the expression *dōiu*. These expressions also appear in the positive corpus set although they are not very frequent.

Table 8: Frequency of *dō* as a single morpheme and as part of compounds.

Adverb	Expression Type	Frequency Whole	Frequency Positive	Frequency Negative	PPM Whole	PPM Positive	PPM Negative
どう *dō*	SM	118,995	47,508	4,822	828.3	162.1	1,537.4
どうして*dōshite*	CM	17,304	5,405	127	120.4	4.3	174.9
どうした*dōshita*	CM	6,985	2,861	53	48.6	1.8	92.6
どういう *dōiu*	CM	11,158	5,488	153	77.7	5.1	177.6
どう言う *dōiu*	CM	112	86	0	0.8	0.0	2.8
どう云う *dōiu*	CM	31	2	0	0.2	0.0	0.0
どのような *donoyōna*	CM	9,680	2,052	2,723	67.4	91.5	66.4
どの様な *donoyōna*	CM	146	114	21	1.0	0.7	3.7
どのように *donoyōni*	CM	8,545	2,253	1,820	59.5	61.2	72.9
どの様に *donoyōni*	CM	161	130	13	1.12	0.4	4.2
どのようにして *donoyōnishite*	CM	867	183	91	6.0	3.1	5.9
どうやって *dōyatte*	CM	3,163	1,459	35	22.0	1.2	47.2
どんな *don'na*	CM	22,791	7,787	374	158.6	12.6	252.0
かどうか *ka dō ka*	CM	16,122	4,565	3,220	112.2	108.2	147.7
か否か *ka ina ka*	CM	2,728	168	1,411	19.0	47.4	5.4

Note: SM=single morpheme, CM=compound morphemes

Moreover, *dōitta* is used with the same meaning as that used in example 14. On the other hand, *donoyōni* is more formal and is the preferred substitution for *dō* in written academic Japanese. Lastly, with regards to *donoyōni*, we found that it is used more frequently in the positive rather than the negative corpus set.

In summation, we surveyed adverbs that include KOSODO demonstratives, and compared their respective frequencies in the positive and negative corpus sets. The results show that *KO* group adverbs are used more in the positive corpus. The frequency of *SO* group adverbs is comparatively lower in the positive corpus, although instances of anaphora usage seem to be permitted. *DO* group adverbs are relatively infrequent, except as the noun modifier *donoyōna* that was observed in the positive corpus set. Consequently, we must be careful when classifying cases of *sō* and *dō* usages; in most cases, *kō* is more acceptable. Also, we need to be especially aware of their compound usages, which are not immediately clear from the short-unit word morphological annotation of corpora.

5 Discussion and conclusion

This paper analyzed the distributional trends of adverbs within the corpora used for the automatic classification of register misuse with the goal of improving the classification rate of adverbs in the academic writing of L2 Japanese language learners. Register misuse was identified by comparing distributional trends between corpora representing the target register of academic writing and the opposing register of informal and spoken corpora against the backdrop of all the corpora combined. The results of classifying all adverbs extracted from the UniDic dictionary were compared against the classifications given by an L2 academic Japanese teaching expert. We summarize our results as follows:

I. Our original and general algorithm for classifying register misuse was found to work for the specific case of adverbs. By using the adverb list of UniDic, the system was able to find occurrences of 6,935 adverbs and classify each into adverbs into either acceptable, unacceptable or unknown groups. In total, it classified 121 adverbs as acceptable and 2,712 as unacceptable for use in academic writing.
II. We were able to clarify the tendencies of orthographic usage differences in each genre by taking into account the relationship between lemmas and their orthographic forms using UniDic. From this investigation, we also found that the existing orthographic standards in Japan are not comprehensive or widespread enough in their use. However, by basing our recommendations on the distributional tendencies of lemma within the positive corpus set, we were able to recommend the use of hiragana for most adverbs, with exceptions such as 最も *mottomo* and 極めて *kiwamete*, which are written using a mixture of kanji and hiragana.

III. We found some expressions that were classified as unacceptable but seem to be useful for academic writing when approached as compounds. Expressions such as *sō de nakereba* and *konoyōni* that contain demonstrative adverbs from the KOSODO word group are observed in academic writing. These kinds of compound adverbs should be either whitelisted or deferred to the classification dealing with longer word units at a deeper linguistic level.

On the other hand, we found the following problems with our classification approach:

IV. While the classification was based on orthographic forms, we also examined words from a lemma-centric viewpoint. From the perspective of learner writing in a setting without an official style guide, it is important to convey the fact that the same lemma may contain different orthographic forms, some acceptable and some unacceptable for use in the academic register. While the variation that exists within words that have multiple forms is often used to convey different nuances, especially within the more creative literary writing found in the Books sub-corpora of the BCCWJ, as the goals of the academic genre are to disseminate information in a standardized and clear way, this variation is undesirable and consequently, rarely employed in academic writing.

V. There are some considerable problems when using the present data. Firstly, several academic papers, predominantly from natural language processing journals, include examples of conversational language that skewed the results for some adverbs within the positive corpus set. In addition to the identification and deletion of these parts, the addition of more data from scientific and technical fields not related to language should also help alleviate this problem. As the treatment of collocations is related to the study of multi-morpheme compounds, further linguistic investigation along these lines is needed. At the same time, the treatment of orthographic variation under the lemma promises to be an interesting research area. As the Japanese orthography phenomenon is quite complicated for learners to grasp, we plan to consider supporting learners by introducing a new method focused on assisting orthographic choice.

VI. Because the classification algorithm compares the relative frequencies between the positive and negative corpus sets, adverbs having a high frequency within the positive corpus set may still be classified as 'unacceptable', although their frequency is quite high. We also found differences between the classifications of the L2 language education expert and the system. Further consideration of the algorithms in the system is needed.

VII. The current system classifies a few extremely low frequency adverbs as acceptable. However, it is possible to prevent this if we set a minimum threshold value for

classification with the end goal being to reduce the number of false positives (i.e. classifying correct expressions as incorrect). Also, decreasing the number of unknown classifications by lowering the significance threshold of the chi-square test could be used to improve the recall of the system. This will be left to further research.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP15K01114 (C) (PI: Abekawa Takeshi; 2015-2017) for the 'Study on effective information displays for error detection and revising within a Japanese writing support system'.

Literature

Abekawa, T., Yagi, Y., Hodošček, B., & Nishina, K. (2015). Improvement of Error Feedback Method in Japanese Composition Support System "NUTMEG". In *Proceedings of the 6th International Conference on Computer Assisted Systems For Teaching & Learning Japanese (CASTELJ)* (pp. 115–118). Honolulu, Hawaii.

Halliday, M. & Hasan, R. (1976). *Cohesion in English*. English Language Series. Longman.

Hodošček, B. (2011). Word class ratios and genres in written Japanese: Revisiting the Modifier Verb Ratio. *Acta Linguistica Asiatica, 1*(2), 53–62. Retrieved from http://revije.ff.uni-lj.si/ala/article/view/28/37

Hodošček, B. & Nishina, K. (2011). On the treatment of register in writing assistance systems. (Vol. 2, pp. 522–523). International Conference on Japanese Language Education 2011. Tianjin, China.

Hodošček, B. & Nishina, K. (2012). Japanese learning support systems: Hinoki project report. *Acta Linguistica Asiatica, 2*(3) Lexicography of Japanese as a Second/Foreign Language (Part 2), 95–124. Retrieved from http://revije.ff.uni-lj.si/ala/article/view/221

Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., … Den, Y. (2013). Balanced Corpus of Contemporary Written Japanese. *Language Resources and Evaluation*, 1–27. doi:10.1007/s10579-013-9261-0

Matsumura, A., Ikegami, A., Kaneda, H., Sugizaki, K., Suzuki, T., Nakajima, T., … Hida, Y. (1998). *Digital Daijisen (Japanese dictionary)* (2015th ed.). Shogakukan.

Ministry of Education, Culture, Sport, Science and Technology-Japan, Section for the Japanese Language (Ed.). (2014). *Kōyōbun no kakiarawashikata no kijun [Criteria for the writing of official documents]* (Revised Edition). Ōkurashō Insatsukyoku.

Mizumoto, T. & Komachi, M. (2012). Robust NLP for Real-world Data: 3. Why is Japanese so Hard to Learn?—A Preliminary Investigation on Realistic Japanese Learners' Corpus and Application of Natural Language Processing to Japanese Language Learning and Education—. *IPSJ Magazine, 53*(3), 217–223.

Ng, H. T., Wu, S. M., Briscoe, T., Hadiwinoto, C., Susanto, R. H., & Bryant, C. (2014, May). The CoNLL-2014 Shared Task on Grammatical Error Correction. In *CoNLL Shared Task* (pp. 1-14).

Ogiso, T., Ogura, H., Koiso, H., Miyauchi, S., Watanabe, R., & Den, Y. (2010). Keitaiso kaiseki jisho no benchimāku tesuto: IPAdic, NAIST-jdic, UniDic no janrubetsu seido hikaku [Benchmark tests on a morphological dictionary: Between-genre comparison for IPAdic, NAIST-jdic and UniDic]. In *Proceedings of the 16th Conference of the Association of Natural Language Processing* (pp. 326–329).

Srdanović, I., Hodošček, B., Bekeš, A., & Nishina, K. (2009). Extraction of suppositional adverb and clause-final modality form distant collocations using a web corpus and corpus query system and its application to Japanese language learning. *Journal of Natural Language Processing, 16*(4), 29–46.

Watanabe, S. (2010). Analysis of the use of adverbs in essays written by undergraduate international students and Japanese students. *Kyoto Sangyo University (Ronshū), 41*, 77–92. Retrieved from http://ci.nii.ac.jp/naid/110007523044/

Yagi, Y., Hodošček, B., Abekawa, T., & Nishina, K. (2014a). Nihongo sakubun suikō shien shisutemu "Natsumegu" ni okeru gakushū hyōka jikken no bunseki [Analysis of Learner Evaluations of Japanese Composition Support System "Nutmeg"]. In *International Conference on Japanese Language Education 2014.* Sydney: International Conference on Japanese Language Education 2014.

Yagi, Y., Hodošček, B., Abekawa, T., & Nishina, K. (2014b). Problems Found in a Learner Evaluation Experiment Using Japanese Composition Supporting System "Nutmeg". In *Dai rokkai kōpasu nihongo wākushoppu yokōshū [Proceedings of the 6th Workshop on Japanese Corpus Linguistics]* (pp. 229–232). NINJAL.

Yagi, Y., Hodošček, B., Abekawa, T., Nishina, K., & Murota, M. (2014). Analysis of Learner Responses to Errors Identified Using a Composition Support System. In *Proceedings of the Research Report of the JSET Conference* (Vol. 14, 5, pp. 151–156). Japan Society for Education Technology (JSET).

Internet resources

Hinoki Project (Natsume, Nutmeg, Natane): https://hinoki-project.org/ (25.8.2019)

MeCab Japanese morphological analyzer: (25.8.2019) https://taku910.github.io/mecab/

要旨（Abstract in Japanese）

「日本語作文支援システム「ナツメグ」を利用した作文に見られる副詞用法の適切さの分析」

ホドシチェック・ボル（大阪大学）
仁科喜久子（東京工業大学）
八木豊（株式会社ピコラボ）
阿辺川武（国立情報学研究所）

現在公開中の日本語学習者のため作文支援システム「ナツメグ」（http://hinoki-project.org/nutmeg/）は、自動添削を可能にすることを最終目的としている。本稿ではアカデミック日本語における学習者作文における副詞をレジスターの視点から観察し、科学技術論文を含む大規模な日本語コーパスを用いて、論文で用いる副詞と用いられない副詞を計量的に分けることを試みた。コーパスからはUniDicで定義された副詞2,919項目を抽出し、各副詞が論文としてレジスターに適切か否かを日本語教育あるいは日本語学の専門家による判定と、システム判定を比較した結果、専門家が科学技術レジスターで適切とした多くの副詞群が、システムでは「不適切」となった。その原因のひとつは、UniDicの副詞が短単位であるため、複数の単位からなる、短単位の複合形が抽出できないためと分かった。今後、複合形を含む副詞辞書の整備が必要であるものの、レジスター判定で学習者の論文作成を支援する可能性があることが明らかとなった。

IV

Corpus-based diachronic research

9 Stylistic differences across time and register in Japanese texts: A quantitative analysis based on the NINJAL corpora

OGISO Toshinobu
National Institute for Japanese Language and Linguistics

Abstract
The construction of the Corpus of Historical Japanese (CHJ) is currently being prioritised at the National Institute for the Japanese Language and Linguistics. Thus far, the parts of the Corpus of Historical Japanese available to the public are the Heian period series and the Muromachi period series I: *Kyōgen*. The corpus of *Sharebon* (comprising data from a type of 18th to 19th century novel) and the *Kindai zasshi* corpus (comprising data from magazines of the 19th to 20th centuries) are currently under construction. These corpora are in the form of a full-text database, and are fully annotated with morphological information, such as parts of speech, lemma, and word form. Thus, multidirectional analysis of this data is possible.

As for historical Japanese documents, existing materials are limited, and we can usually use only the documents of a specific genre in a particular period. Therefore, it is often difficult to determine, for any particular characteristic discovered by observation of this corpus, whether it is the result of historical language change or due to a difference between genres.

In this paper, what will be demonstrated is an examination of the characteristics identified by the enumeration of the morphological information in the corpus of each historical period, and these will be compared to the characteristics of the various text genres of contemporary Japanese drawn from the Balanced Corpus of Contemporary Written Japanese. By these means we aim to investigate the characteristics of the documents constituting the CHJ more thoroughly and reliably, so that it may be used for historical study of the Japanese language.

Keywords: historical Japanese, corpus, language change, BCCWJ, CHJ

1 Background

When we study the history of the Japanese language, a fundamental problem is that the documents of each period are limited to particular genres. There are a few written documents from each historical period as most were scattered and lost over time. Furthermore, only a few of the available documents are suitable for the study of Japanese. For example, the most important material from the Nara period is limited to poetry from the *Man'yōshū* collection, and the most important documents for the study of language

in the Heian period are *kana* literary works such as *monogatari* novels and diaries. Although documents from non-literature genres have been preserved, most are written in the classical Chinese style and are not suitable for the study of Japanese. For periods after the Middle Ages, the situation is somewhat better and there are more remaining documents. However, documents that reflect spoken language are fewer in number. Thus, there are generally few historical documents suitable for thorough linguistic study.

In light of the above, we are forced to use documents of a specific genre in a particular period for studying Japanese. As a result, it is difficult to determine whether the characteristics observed in a historical corpus for any given period result from historical language change, or to differences among text genres. Mistaking stylistic differences based on text genre for difference due to time can be a serious problem for historical linguistic study. In this regard, it is important to understand the origin of any characteristic of the language in a corpus.

What will be proposed here is that the problem described above may be addressed by using data from both the Balanced Corpus of Contemporary Written Japanese (BCCWJ) and the current Corpus of Historical Japanese (CHJ). Differences in the Japanese language across genres may be identified from various types of texts in the BCCWJ, and differences across time may be identified from the historical texts of the CHJ. Thus, what will be shown here is a basic quantitative analysis of the data in various historical and contemporary Japanese texts.

2 Data source: The NINJAL corpora

At the Corpus Development Centre of the National Institute for Japanese Language and Linguistics (NINJAL), various types of corpora of Japanese have been and are being developed. They include data from Japanese language materials from ancient to contemporary Japan, both spoken and written. In this section, we will present an outline of the NINJAL corpora used for this study. Detailed information on these corpora appears on the website of the NINJAL Corpus Development Centre[1].

These corpora comprise full-text databases and are fully annotated with morphological information, such as parts of speech, lemma, and word form. The morphological information is based on the definition of Short Unit Words, defined by NINJAL for the purposes of the development of Japanese corpora (Ogura et al. 2011). The definition of the Short Unit Word is explained briefly in Section 2.4. By using morphological information, statistical analysis of text is possible.

1 http://pj.ninjal.ac.jp/corpus_center/en/

2.1 The Corpus of Historical Japanese

The CHJ is currently under development at the NINJAL. The aim of this is that the CHJ becomes a large diachronic corpus that covers historical Japanese language materials from the Nara period to modern times. As parts of the CHJ, the Heian period series and the Muromachi period series I: *Kyōgen* have been released.

The Heian period series contains fourteen *kana* literature works, namely *Tosa Nikki* (Tosa diary), *Taketori Monogatari* (The Tale of the Bamboo-Cutter), *Ise Monogatari* (The Tales of Ise), *Ochikubo Monogatari* (The Tale of Ochikubo), *Yamato Monogatari* (The Tales of Yamato), *Makura no Sōshi* (The Pillow Book), *Genji Monogatari* (The Tale of Genji), *Murasaki Shikibu Nikki* (The Diary of Lady Murasaki), *Izumi shikibu Nikki* (The Diary of Izumi Shikibu), *Heichū Monogatari* (The Tales of Heichū), *Sarashina Nikki* (Sarashina diary), and *Sanukinosuke Nikki* (The Diary of Sanukinosuke). These works were written in the Heian period (794–1185) and play a key role as Japanese classics, as they are of great literary value. Furthermore, these works have been widely used as the most important source of data for the study of Japanese in the Heian era.

The Muromachi period series I: *Kyōgen* contains 236 scripts of *Kyōgen* written by Okura Toraakira in 1642. *Kyōgen* is a form of comic theatre that developed alongside *noh* in the Muromachi period (1337–1573). For studying the language of the Muromachi era, *Kyōgen* scripts, Christian documents such as *Esopo no Fabulas*, and documents from *Shômono* (a kind of correspondence course) are important materials. Among these, *Kyōgen* plays a major role, and these texts have been used to study the spoken language of the Muromachi era.

In addition to this, a corpus of *Sharebon* is under development as part of the CHJ. This corpus is planned as part of the Edo period series of CHJ. *Sharebon* is a kind of novel developed in the 18th and 19th centuries. Twelve works of *Sharebon* are currently available. Although many documents written in the Edo era remain today, there are few that reflect the language that was spoken at that time. Therefore, *Sharebon* books, which contain many conversations, are valuable in the study of Japanese in the Edo era.

2.2 The *Kindai Zasshi* corpus

The *Kindai Zasshi* corpus consists of some independent corpora consisting of magazines published in the 19th and 20th centuries, earmarked for incorporation into the CHJ in the near future. One of the corpora is the *Taiyō* corpus, published in 2005, which is a corpus of the general interest magazine *Taiyō* (太陽, The Sun) published by *Hakubunkan* from 1895 to 1928. This magazine was read widely throughout Japan and had a widespread influence in the *Meiji* and *Taishō* eras. The *Taiyō* corpus contains the full-text of approximately 14,500,000 characters published in five years, namely 1895, 1901, 1909, 1917, and 1925. Because the *Taiyō* corpus includes a range of texts

by a large number of authors, it provides suitable material for the study of Japanese at the time. Although the *Taiyō* corpus was originally published without morphological information, morphological annotation has been completed, and morphological information has been manually corrected in the part that forms the core data. The core data is approximately 2% of the corpus.

The *Meiroku Zasshi* corpus, published in 2012, is another part of the *Kindai Zasshi* corpus. Although its scale is not large, it contains all issues of *Meiroku Zasshi* (明六雑誌, *Meiroku* Magazine) published by *Meiroku-sha* since 1874. It was the first modern magazine in Japanese that played a significant role in spreading Western ideas and thought. Its influence on Japanese cultural history was substantial, and it forms important material for the study of Japanese in modern times. The corpus comprises approximately 180,000 word tokens and its morphological annotation has been completely manually corrected.

Finally, the *Kokumin no Tomo* corpus, published in 2014, forms part of the *Kindai Zasshi* corpus. This is a corpus of *Tokumin no Tomo* (国民之友, The Nation's Friend) published by *Min'yū-sha* since 1887. This magazine corpus falls between the *Taiyō* and the *Meiroku Zasshi* period corpora. It was a widely read general interest magazine at the time. This corpus contains 101 million word tokens and a part has been manually corrected as the core data.

For the present study, we used the core data of the *Taiyō* corpus (222 thousand word tokens), all the data of the *Meiroku Zasshi* corpus (34 thousand tokens), and the core data of the *Kokumin no tomo* corpus (180 thousand tokens).

2.3 Balanced Corpus of Contemporary Written Japanese

The BCCWJ is a large-scale corpus of Japanese, containing more than 1,000 million word tokens. It includes contemporary written Japanese texts of various genres, media, and registers (Maekawa et al. 2014). The BCCWJ consists of both core data and non-core data. The morphological annotation of the core data has been manually corrected and its accuracy is higher than 99%. In contrast, the non-core data of the BCCWJ has been digitally analysed, but its accuracy is about 98%.

The BCCWJ is far larger than the historical corpora mentioned above. Therefore, the core data alone were considered sufficient for this study. The core data consist of six registers, namely books (PB: Publication+Books), magazines (PM: Publication+Magazines), newspapers (PN: Publication+Newspapers), white papers (OW: Out-of-population+Whitepapers), web data of the Q&A service "Yahoo! Chiebukuro" (OC: Out-of-population+*Chiebukuro*), and blogs (OY: Out-of-population+Yahoo! Blog). In this context, "out-of-population" means that texts of these sub-corpora were not sampled from a designed statistical population of contemporary written Japanese, but were collected for certain specific purposes.

2.4 The Corpus of Spontaneous Japanese

The Corpus of Spontaneous Japanese (CSJ), published in 2004, is a collection of large quantities of audio recordings of Japanese speakers, and is annotated with information for phonetic and phonemic studies. The CSJ also includes approximately 661 hours and 7,520 thousand transcribed tokens Although these data consist mainly of monologues, such as lectures, they serve as valuable contemporary spoken Japanese data.

In this study, we used approximately one million words from the text of the CSJ, called core data, as a sample of contemporary spoken Japanese (1,011 thousand tokens). The morphological information of the core data has been manually corrected.

2.5 Sizes of the corpora

Table 1 shows the sizes of the corpora described above.

Table 1 Sizes of the corpora

Target	Corpus	Sub-corpus	Size (tokens)
Diachronic (historical)	CHJ	Heian	871,477
		Kyōgen	277,424
		Sharebon	94,125
	Kindai	Meiroku	180,654
		Kokumin	34,279
		Taiyō	222,479
Synchronic (contemporary)	BCCWJ	OC_core	110,071
		OW_core	227,766
		OY_core	116,806
		PB_core	234,206
		PM_core	238,857
		PN_core	360,182
	CSJ	CSJ_core	1,011,681

Note: The names of corpora are abbreviated as follows:
Heian: CHJ, Heian period series (14 works)
Kyōgen: CHJ, Muromachi period series I
Sharebon: CHJ, Edo period series (12 works, under development)
Meiroku: Meiroku Zasshi corpus
Kokumin: Core data of the *Kokumin no Tomo* corpus
Taiyō: Core data of the *Taiyō* corpus

OC_core: BCCWJ core data of Yahoo! *Chiebukuro* data (Q&A web service)
OW_core: BCCWJ core data of government White Papers
OY_core: BCCWJ core data of Yahoo! Blog service
PB_core: BCCWJ core data of published Books
PM_core: BCCWJ core data of published Magazines
PN_core: BCCWJ core data of published Newspapers
CSJ_core: core data of the Corpus of Spontaneous Japanese

Except the case of *Sharebon* (now under development), the morphological information of all these sub-corpora has been manually corrected. In order to prioritise quality over quantity, the data for the present study were limited to manually corrected parts of the corpora, decreasing their size.

2.6 Historical periods and textual styles of the corpora

The table below shows the historical periods of the NINJAL corpora used for this study. Data were drawn from six historical and seven contemporary corpora.

Table 2 Historical periods of the corpora

Corpus	**Sub-corpus**	**Periods**	**Historical stage of Japanese Language**
CHJ	Heian Series	10-11c	Early Middle Japanese (*Chūko-go*)
	Kyōgen	15-16c	Late Middle Japanese (*Chūsei-go*)
	Sharebon	18-19c	Early Modern Japanese (*Kinsei-go*)
Kindai	Meiroku	1874-1875	Modern Japanese (*Kindai-go*)
	Kokumin	1887-1888	
	Taiyō	1895-1925	
BCCWJ	OC_core	2005	Contemporary Japanese (*Gendai-go*)
	OW_core	1976-2005	
	OY_core	2008	
	PB_core	1971-2005	
	PM_core	2001-2005	
	PN_core	2001-2005	
CSJ	CSJ_core	1999-2005	

The sub-corpora of the CHJ contain conversation-rich literary work, and are regarded as reflecting the colloquial language of the time. On the other hand, the sub-corpora of the *Kindai Zasshi* corpus contains a more refined text of the editorial writing style at the time.

The seven sub-corpora of contemporary Japanese contain a wide range of styles. OC and OY texts originate from web-based media, and are written in a light colloquial style. Conversely, OW texts are written in the formal style required for government white papers.

The terms "genre", "style", and "register" may be confusing and difficult to define. In this paper, the term "genre" is used to indicate the traditional classification of the document, such as a poem, novel, diary, playbook, editorial article, etc. The word "style" is used to refer to characteristics of the text, which derive from the differences among genres, entailing aspects such as historical era, tone of writing, etc. Finally, the word "register" is used for the various contemporary corpora and sub-corpora examined here. In the BCCWJ, sub-corpora such as OC, OW, PB, and PN are regarded as registers, and the CSJ is also regarded as a register of contemporary Japanese. The word register is purposely not used to refer to the historical corpora here.

2.7 Definition of word segmentation in the NINJAL corpora

In the Japanese writing system, no open space occurs between words, and agreement on how to segment words is generally lacking. Thus, it is necessary to define the manner in which words are segmented for annotation in the Japanese corpora.

The BCCWJ is annotated with morphological information in terms of units of two sizes, namely Short Unit Words (SUWs) and Long Unit Words (LUWs) (Maekawa et al. 2014). A SUW is a word of small size based on morphological unity. On the other hand, a LUW is a larger unit of words based on sentence structure (*bunsetsu*). A LUW consists of a combination of SUWs. For example, the string 国立国語研究所 (*Kokuritsu-kokugo-kenkyūsho*, National Institute for Japanese Language and Linguistics) is segmented as follows:

In SUWs: 国立 (*kokuritsu*, national) / 国語 (*kokugo*, Japanese language) / 研究 (*kenkyū* research) / 所 (*sho*, institute); four words

In LUWs: 国立国語研究所 (*Kokuritsu-kokugo-kenkyūsho*, literally: National-Japanese language-research institute); one word

The use of each of these two word units depends on the purpose of the study. However, except in the case of the BCCWJ and the *Heian* series of the CHJ, LUW annotation has not yet been completed. Thus, SUWs were used for the present investigation.

3 Measurement indices and results

In this section, we discuss the characteristics of the corpora on the basis of widely used indices, namely the parts of speech ratio, the type-token ratio (TTR), the modifier/verb ratio (MVR), and the *goshu* ratio.

Table 3. Numbers of tokens by parts of speech in the corpora

Corpus	Sub-corpus	Par-ticle	Verb	Copula	Aux. verb	Noun	Pro-noun	Adverb	Adj.	Adj. noun	Conj.	Prefix	Suffix	*Total*
CHJ	Heian	219058	163999	22831	73012	151792	12278	27127	35084	6394	521	14389	10874	*737359*
	Kyōgen	74818	47125	6665	20434	52999	8567	5220	5329	2193	1062	2512	2460	*229384*
	Sharebon	26279	13927	2310	5950	23416	2813	2948	2316	849	260	1402	1928	*84398*
Kindai	Meiroku	52274	28450	5225	10462	58422	4060	5824	2304	1456	2335	1063	2039	*173914*
	Kokumin	9624	4666	1077	2364	9189	679	1043	483	421	298	230	612	*30686*
	Taiyō	63616	29515	7132	12780	58238	4164	5049	3659	2468	1254	1330	4564	*193769*
BCCWJ	OC_core	30392	13404	2827	10378	26352	1425	1976	2182	1158	204	699	2101	*93098*
	OW_core	49157	21912	3537	5790	96297	433	811	1050	2510	1630	1715	11012	*195854*
	OY_core	27096	11335	2800	7891	31965	1375	2180	1851	1164	370	840	2592	*91459*
	PB_core	66616	30514	7520	14597	59821	3655	4208	3591	2675	871	1283	6064	*201415*
	PM_core	59876	25834	6276	11413	76812	2540	3568	3117	2613	553	1351	6446	*200399*
	PN_core	83719	33696	5594	14529	142680	1332	2154	3023	2882	602	2753	13992	*306956*
CSJ	CSJ_core	308060	129836	32002	87648	240674	21377	29414	14741	12729	11757	6079	20589	*914906*

Whereas the part of speech and type-token ratios are popular indices internationally, *goshu* is a unique index used for Japanese text analysis. *Goshu* refers to the origin of Japanese words (such as Chinese, Japanese, other foreign words, mixture of origins) and is similar to the strata of the English lexicon, which include words of Anglo-Saxon, French, Latin, etc. origin. The MVR is an index proposed by Kabashima (1965), and remains commonly used for Japanese text analysis (Koiso et al. 2008). The data relating to these indices were extracted from the NINJAL Morphological Information Database (Ogiso and Nakamura 2014) using SQL database queries.

3.1 Part of speech ratios

Table 3 shows the parts of speech ratios of the corpora. Although the corpora contain insubstantial tokens, such as signs, supplementary symbols, blanks, and particular un-analysable words, these are excluded from the data in the table. (A total of 469,000 tokens were excluded and approximately 96% of these were supplementary symbols and blanks.)

In all the corpora, the definitions of the parts of speech were based on the rules for SUWs. However, note that the auxiliary verbs *da* and *nari* were classified as verbs here. In Japanese, the behaviour of *da* (in colloquial language) and *nari* (in literary language) is similar to that of *be* verbs in English. They are therefore considered appropriate to be classified as copular verbs.

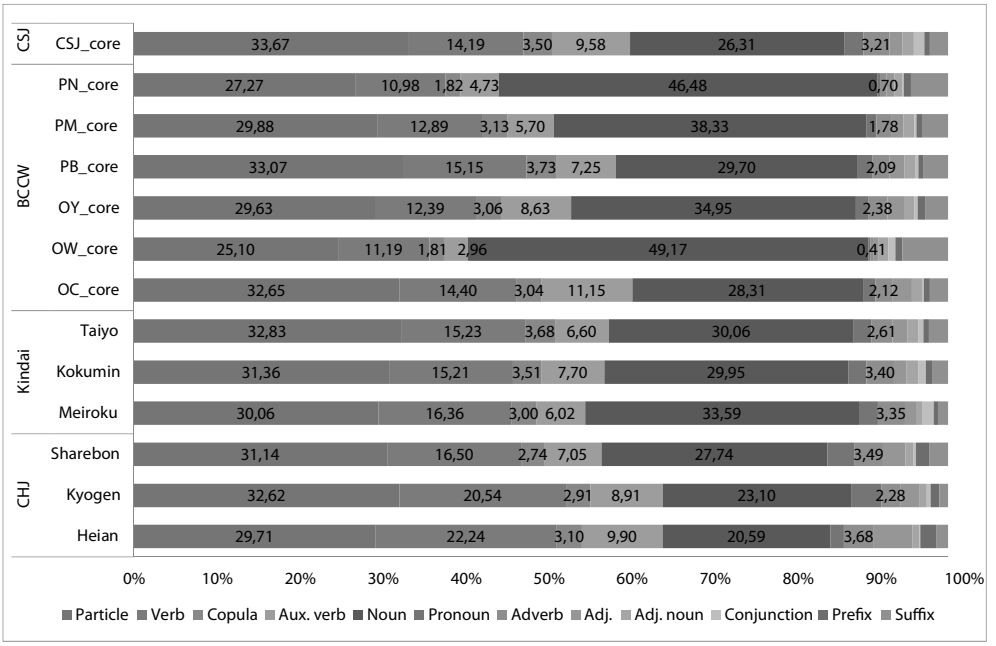


Figure 1: Parts of speech ratios based on tokens

Figure 1, which was generated with the data in Table 3, shows the parts of speech ratios of the corpora based on word tokens. For purposes of clarity, only the percentages of particles, verbs, copulas, auxiliary verbs, nouns, and adverbs are indicated.

With regard to parts of speech, it is known that the noun ratio reflects the characteristics of a text (Kabashima 1965, Koiso 2008). In the BCCWJ, sub-corpora written in in a refined style, such as OW and PN, have high noun ratios. However, in the corpora of *Kindai Zasshi,* also written in a formal style, the noun ratio is not as high. The high noun rates in OW and PN may be due to the multitude of compound nouns. In SUW analysis, compound nouns are divided into multiple nouns, resulting in an increase of the noun ratio.

From an historical perspective, the noun ratio increases over time: *Heian* series: 21%, *Kyōgen*: 23%, *Sharebon*: 28%, *Kindai*: 30-34%, and BCCWJ: 28-49%. Noun ratios including pronouns are as follows: *Heian*: 22%, *Kyōgen*: 27%, *Sharebon*: 31%, *Kindai*: 32-36%, and BCCWJ and CSJ: 29-49%. This finding may be due to historical changes, reflecting a consistent increase in the use of compound nouns. On the other hand, there remain considerable differences among the BCCWJ registers. Since a lot of nouns are divided into nouns and suffixes or prefixes and nouns in SUW annotation, this may influence the decrease in the number of nouns as well, which requires additional analysis based on LUW (in future).

3.2 Modifier/Verb Ratios

As mentioned above, the MVR is an index proposed by Kabashima (1965), and is calculated as follows:

$$\text{MVR} = \frac{\text{Number of modifiers (adjective + adjectival noun + adverb)}}{\text{Number of verbs}} \times 100$$

A high score means that the text contains many modifiers (words denoting manner), whereas a low score means that the text contains many verbs (words denoting movement).

The MVR scores of the corpora are as follows: *Heian*: 41.83, *Kyōgen*: 27.04, *Sharebon*: 43.89, *Kindai*: 34-42, and BCCWJ and CSJ: 20-45. These scores reflect straightforward historical change. The MVR scores are discussed in relation to the noun ratios in Section 0 below.

3.3 *Goshu* ratio

As mentioned above, *goshu* refers to the origin of Japanese words. Generally, Japanese words are classified in terms of three origins, namely *wago* (native Japanese words), *kango* (Chinese or Sino-Japanese words), and *gairaigo* (words of foreign origin other

than Chinese). Although most compound words consist of words with the same origin, some compound words consist of a mixture of *wago*, *kango*, and *gairaigo*. Such mixed origin words are labelled as *konshugo* (of hybrid origin). Such hybrid origin words are observed even in SUWs.

In the Japanese language, all particles and most adjectives (い形容詞) and auxiliary verbs are native Japanese, as are many basic words. On the other hand, Sino-Japanese and foreign words are common among nouns and adjectival nouns (形容動詞 / な形容詞). The origins of proper nouns are difficult to determine and are of less importance in the linguistic analysis of these texts. Therefore, they are simply marked as proper nouns without information about their origin.

Table 4 shows the numbers of *goshu* in the corpora. In this table, the "Other" column reflects errors of morphological analysis (i.e. unknown words) and words not defined for various reasons.

Table 4: Numbers of *goshu* in the corpora

Corpus	**Sub-corpus**	**Foreign (*gairaigo*)**	**Sino-Japanese (*kango*)**	**Hybrid (*konshugo*)**	**Native (*wago*)**	**Proper noun**	**Sign**	**Other**
CHJ	*Heian*	151	20696	2541	709814	4911	133309	61
	Kyōgen	207	18746	6197	207264	2434	42365	249
	Sharebon	149	7257	1668	72925	4606	7032	747
Kindai	*Meiroku*	558	43976	3677	127716	2655	2060	55
	Kokumin	59	6971	383	23256	643	2922	55
	Taiyo	617	50947	4060	150972	3298	4982	265
BCCWJ	OC_core	3423	16169	884	71717	1226	16761	102
	OW_core	4610	92214	2649	93712	2786	32118	83
	OY_core	4219	19135	1041	65281	2584	24782	200
	PB_core	4007	37755	1832	155227	5096	30431	53
	PM_core	10493	48367	2275	132597	7655	37999	60
	PN_core	9590	110045	2991	168325	16631	52865	79
CSJ	CSJ_core	24137	164910	7973	788933	10302	6027	13294

Figure 2 shows the *goshu* ratios in the corpora graphically. For purposes of simplicity, only the four main classes of *goshu* (*wago, kango, gairaigo*, and *konshugo*) are shown.

The ratio of words of native Japanese origin decreases over time as follows: *Heian*: 81%, *Kyōgen*: 75%, *Sharebon* 77%, *Kindai*: 68-71%, and BCCWJ and CSJ: 41-78%. On the other hand, the ratio of words of Sino-Japanese origin increases by the end of

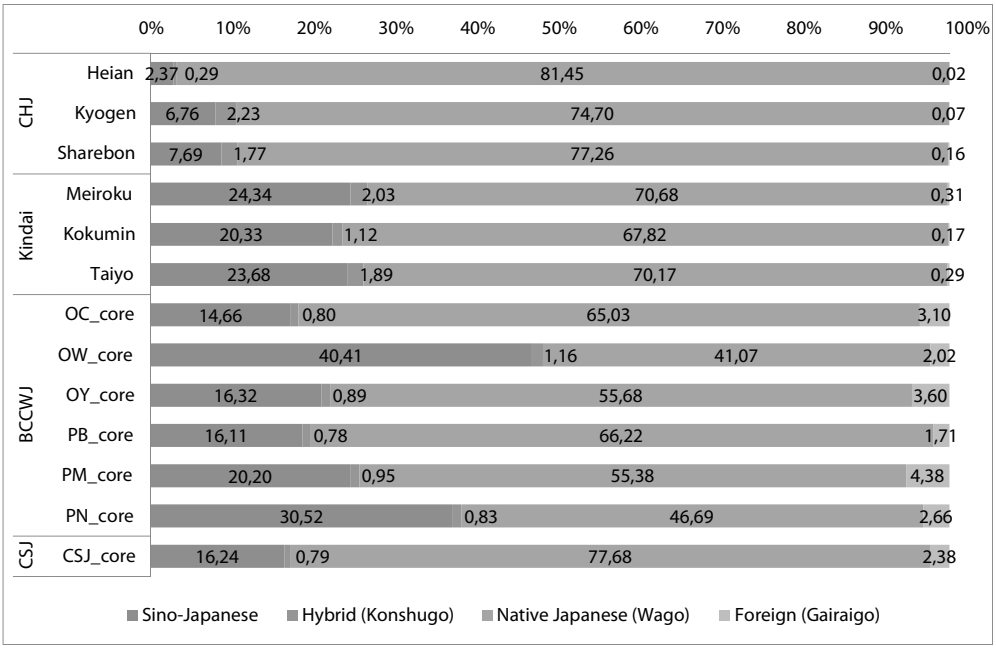


Figure 2: *Goshu* ratios in the corpora based on tokens (simplified)

modern period. However, the ratio of Sino-Japanese words may differ depending on registers in contemporary corpora. The presence of foreign words sharply increases in the contemporary corpora.

Thus, it appears that *goshu* may provide an indication of the historical period of a corpus. However, differences between registers remain considerable.

3.4 Type Token Ratio

The TTR is an index that explains the lexical density of a corpus. The value of the TTR is influenced by corpus size; thus, care is required in comparing TTRs across corpora. To deal with this problem, Baayen (2001) proposed indices that are more representative. However, for the present study, we simply took a designated number of words from the beginning of each corpus and calculated the TTR for that sample. As each text is automatically extracted from the beginning, the selected text represents a mixture of independent samples.

For example, the TTR indices for the Heian series were calculated as follows:

TTR (N=1000) = 299 (types) / 1000 (tokens)
TTR (N=10000) = 1235 (types) / 10000 (tokens)
TTR (N=100000) = 4192 (types) / 100000 (tokens)

Table 5 shows the results for the corpora studied here. The designated numbers of tokens for extraction were 1,000, 10,000, and 100,000, respectively.

Table 5: Type Token Ratios in the corpora

Corpus	Sub corpus	TTR		
		N=1,000	N=10,000	N=100,000
CHJ	*Heian*	29.90	13.25	4.19
	Kyōgen	29.30	11.76	4.56
	Sharebon	41.30	21.86	8.35
Kindai	*Meiroku*	38.00	19.41	9.59
	Kokumin	41.70	20.92	N/A
	Taiyō	39.30	20.69	12.97
BCCWJ	OC_core	33.30	18.57	8.12
	OW_core	27.20	13.43	4.86
	OY_core	36.40	22.08	9.85
	PB_core	32.20	15.85	8.38
	PM_core	31.30	18.25	9.83
	PN_core	31.90	21.00	10.31
CSJ	CSJ_core	28.50	10.37	3.30

The TTR scores in this table are difficult to evaluate in isolation. Similarly, small TTRs may be due to a number of different reasons in SUW analysis. For example, a small TTR may reflect the redundancy of spoken language. Thus, because data from historical literature (*Heian* and *kyōgen*) and the CSJ corpus are based on spoken language, their TTRs are small. However, if each text sample is long and includes many technical compound words and fixed phrase, the TTR will also be small. For example, the register OW leads to a small TTR because of the repetition of the same content words and similar expressions.

4 Analysis

Using the data presented in Section 3, we performed several statistical analyses, including an analysis of the relationship between the MVR and noun ratio, cluster analysis by *goshu* and parts of speech, and principal component analysis of synthetic indices.

4.1 Relationship between the MVR and noun ratio

In the field of the statistical analysis of Japanese texts, what has been accepted is that the relationship between the MVR and noun ratio (based on tokens) indicates characteristics of the text (Kabashima 1965). Kabashima states the following with regard to the MVR and noun ratio in contemporary Japanese:

1) If the ratio of the noun is big, and the MVR is small, it is an abstract text.
2) If the ratio of the noun is small, and MVR is big, it is a descriptive text.
3) If the ratio of the noun is small, and the MVR is small, it is a text tending to describe movement.

The above proposals were checked in terms of the historical data compared to the contemporary data in this study. Figure 3 shows a graph in which the MVR scores and noun ratios of the various corpora are mapped.

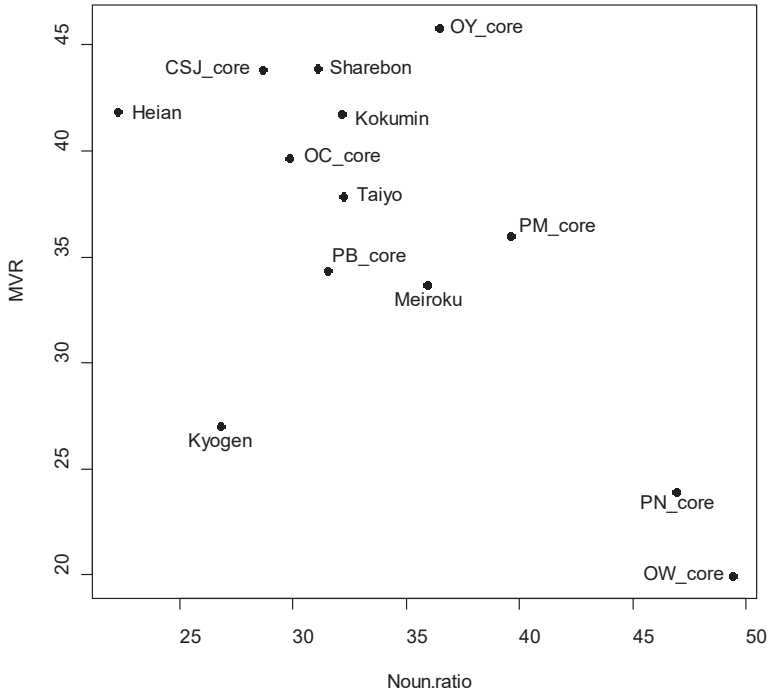


Figure 3. MVRs and noun ratios of the corpora

In Figure 3, the historical corpora and various contemporary corpora are mixed. In terms of Kabashima's (1965) proposal, OW and PN has a large noun ratio and

small MVR, which denotes them as most abstract texts in line with proposal 1. *Heian* is in line with proposal 2 as most descriptive data and CSJ is closest as well, as most descriptive and colloquial data. It is also understandable that *Kyōgen* describes movement.

However, there are some points more difficult to interpret. For example, the *Kokumin*, with the most refined editorial writing style, has been positioned near the colloquial *Sharebon* novels. Furthermore, although the noun ratios seem to explain, to a certain extent, the degree of formality or the casualness of the writing style, the *Kindai zasshi*, which contains mainly formal editorial texts, is located in the middle of the scale. The treatment of compound nouns in SUWs may have caused high scores in OW and PN, as pointed out in Section 3.1.

Thus, although Kabashima's (1965) explanation of the relationship between the MVR and noun ratio fits, to a certain extent, these corpora, it does not seem fully applicable to the present data, in which we analyse data across different periods based on SUWs.

4.2 Cluster analysis by *goshu* and part of speech

In this section, we will discuss the findings of cluster analysis using some of the indices discussed above, showing how the types of corpora are grouped together. As mentioned in Section 0, the *goshu* ratio reflects changes over time. The increase of both Sino-Japanese (*kango*) and Japanese native (*wago*) words are historically consistent. Furthermore, the differences across genres in contemporary Japanese are also considerable. However, words of foreign origin (*gairaigo*) suddenly increase in contemporary Japanese in comparison with *Kindai* corpora. It seems that the data may be grouped according to these features. To investigate this possibility, we performed cluster analysis using the *goshu* ratio. The cluster analysis was completed with the *hclust* function of R by using a Euclidean distance measure and Ward's methods.

Figure 4 shows the results of *goshu* cluster analysis. Only four main *goshu* features were used for this analysis, namely native, Sino-Japanese, foreign, and hybrid. The graph in Figure 3 shows that the corpora are clearly grouped into five clusters:

(1) Contemporary editorial writing style [OW and PN]
(2) Historical literary works [*Heian*, *Sharebon*, and *Kyōgen*]
(3) Contemporary text including conversations [CSJ, OC, and PB]
(4) 18th-19th century magazines (*Kindai Zasshi*) [*Kokumin*, *Meiroku*, and *Taiyō*]
(5) Other contemporary texts (contain no conversations, less bookish) [OY and PM].

Thus, it is confirmed that the *goshu* ratio is a useful index for classifying historical texts and those of various genres.

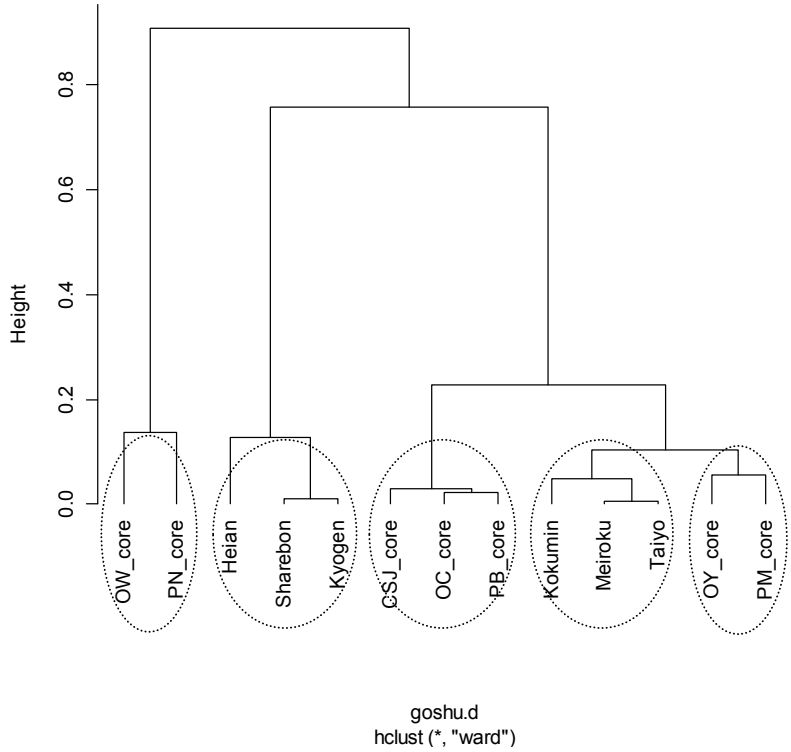


Figure 4. Cluster analysis by *goshu* ratio

Next, we performed a cluster analysis of the parts of speech ratios of the corpora, and compared the results to those of the *goshu* ratio. Figure 5 shows the results of the part of speech cluster analysis, which included 12 parts of speech, namely particle, verb, copula, auxiliary verb, noun, pronoun, adverb, adjective, adjectival noun, conjunction, prefix, and suffix. The analysis was performed in the same way as that for the *goshu* ratio above.

Whereas OW and PN, as well as OC and CSJ, are classified in the same way as in the *goshu* cluster analysis, the remaining corpora are clustered differently. In terms of characteristic writing style, the interpretation of this clustering seems difficult. Moreover, the clustering does not seem to reflect the expected characteristics across time periods.

As for the groupings by parts of speech data, it seems that various factors are overlapped, and classification in terms of stylistic differences cannot be made on the basis of parts of speech alone. As pointed out in Section 3.1, this may be due to the use of SUWs in the present study.

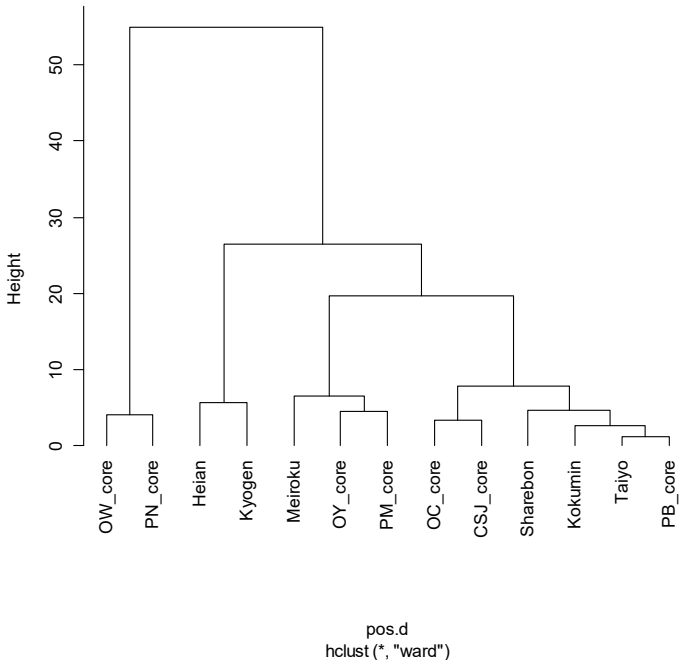


Figure 5. Cluster analysis by part of speech ratio

4.3 Principal component analysis using synthetic indices

Except in the case of the *goshu* ratio, individual indices were limited in their ability to explain the differences across the current text types. Therefore, we attempted to integrate the various indices discussed above, performing principal component analysis using general indices. Various combinations of characteristic features were analysed, including *goshu*, parts of speech, MVR, and TTR. Among them, the pair with the most explanatory power was *goshu* (*kango* and *gairaigo*) and parts of speech (nouns, modifiers, and verbs).

Figure 6 shows the results of the principal component analysis with the *goshu* and part of speech ratios. In this graph, the first principal component (PC1) appears on the X-axis and the second principal component (PC2) on the Y-axis. PC1 consists of noun: 0.4836604, modifier: -0.4735281, verb: -0.4753652, *kango*: 0.4585035, and *gairaigo*: 0.3250326. PC2 consists of noun: -0.2058790, modifier: 0.1899043, verb: -0.2150301, *kango*: -0.4068681, and *gairaigo*: 0.8424789.

The features of large absolute values contribute substantially to the axis. The arrows in the graph show the degree of the contributions to PC1 and PC2. In this analysis, the cumulative proportion of the principal components (PC1 + PC2) is higher than 0.934, and PC1 and PC2 explain most of the variation. Therefore, Figure 6 describes the data quite well.

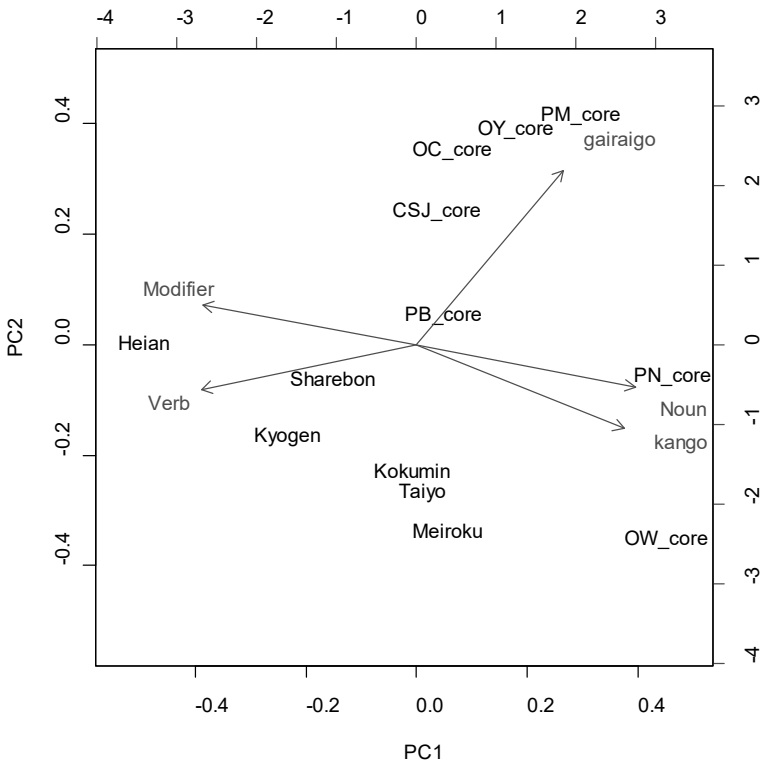


Figure 6. Results of principal component analysis

Note the almost exact chronological order of the corpora in relation to the PC1 axis: *Heian, Kyōgen, Sharebon, Kindai Zasshi*, and BCCWJ. Hence, PC1 may be taken as a component that describes historical differences. In addition, PC1 is also accurate in reflecting differences between colloquial texts (including conversation) and the editorial writing style in the BCCWJ registers. The PC1 axis leads to conversation-rich historical literary works. This axis is composed from parts of speech (+nouns, -modifiers and -verbs) and *goshu* (*kango* ratio).

The PC2 axis also distinguishes contemporary from historical texts. The percentage of *gairaigo* is the major component of PC2. Except for PN and OW, contemporary texts are located on the upper side and historical texts lie below. Moreover, for the BCCWJ registers, this axis shows the opposition between new and old in their lexicon.

In comparison with the wide range of the BCCWJ registers, the historical corpora are gathered together. Thus, the dispersion of the BCCWJ registers is greater than that of the historical corpora. As for the materials useful for the historical study of Japanese,

these tend to be conversation-rich in content, as colloquial language is regarded as important. As a result, whereas PC1 clearly showed differences across time, it also reflected the quantity of conversations in the respective corpora.

We also performed principal component analysis with the TTR. In this case, the cumulative proportion of the principal components (PC1 + PC2) was lower, around 0.82. The TTR has been rejected as the major variable composing PC2. Thus, the TTR seems to be independent from the parts of speech and *goshu* ratios.

5 Summary and conclusion

The present study showed that, for the classification of the mixed corpora of historical texts and various contemporary texts based on SUWs, the *goshu* ratio is the most effective index. The part of speech ratio alone was not enough for the classification of historical texts, as difference across genres greatly influenced the ratio. However, the parts of speech ratio may be more effective for an analysis based on LUWs, most historical texts are not yet annotated for this. Similarly, the MVR and TTR in isolation are not very effective indices. Classification based on both the *goshu* and parts of speech ratios was shown to be effective by principal component analysis.

Let us consider the position of Japanese historical documents in terms of these research results. The texts of the CHJ, examined here, are limited to *kana* literature works in the Heian period, script books of *Kyōgen*, *Sharebon* novels, and magazines from the 18th and 19th centuries. The results of this study showed these historical Japanese texts to be partial and one-sided in comparison with a variety of contemporary Japanese genres. Conversely, the dispersion of each register of contemporary Japanese is substantial.

In the study of historical Japanese, the content of documents is apt to be biased or partial. This must be borne in mind when we study the history of Japanese with corpora of such historical texts, and we should aim to include a variety of documents in the historical corpus after the Middle Ages, if such are available.

Finally, note that this research was based only on SUWs, which is not sufficient. We will strive to continue research with the entire vocabulary of the corpus, and to make these historical documents available for the study of the Japanese language.

Literature

Kabashima, T. and Jugaku A. (1965) *Buntai no kagaku*文体の科学 *(Science of the writing style).* Kyoto: Sogeisha. [In Japanese]

Baayen, R. H. (2001) Word frequency distributions, Kluwer Academic Publishers.

Koiso, H., Ogiso, T. and Ogura, H. (2008) Analysis of style in various genres based on *Short-Unit Word*, *Proceedings of the 2008 General Meeting of the MEXT Grant-in-aid for Scientific Research Priority Area Program "Japanese Corpus"*: 99-106. [In Japanese]

Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., Koiso, H., Yamaguchi, M., Tanaka, M. and Den, Y. (2014) Balanced corpus of contemporary written Japanese, *Language Resources and Evaluation* 48(2):345-371.

Ogiso, T., Nakamura, T. (2014) Design, implementation, and operation of annotation support system for morphological information of BCCWJ, *Journal of Natural Language Processing* 20(1): 301-332. [In Japanese]

Ogura, H., Koiso, H., Fujiike, Y., Miyauchi, S., Konishi, H. and Hara, Y. (2011). Rule book of the morphological information for "Balanced Corpus of Contemporary Japanese" (『現代日本語書き言葉均衡コーパス』形態論情報規程集). Tokyo: National institute for Japanese language and linguistics. [In Japanese]

[Internet resources]

Center for Corpus Development, NINJAL: http://pj.ninjal.ac.jp/corpus_center/en/ (12.1.2016)

Corpus of Historical Japanese: http://pj.ninjal.ac.jp/corpus_center/chj/overview-en.html (12.1.2016)

要旨（Abstract in Japanese)

「各時代やジャンルにおける日本語文章のスタイル変遷–国立国語研究所の諸コーパスを利用した数量的研究—」

小木曽智信（国立国語研究所）

現在、国立国語研究所では『日本語歴史コーパス』の構築が進められている。これまでに「平安時代編」（10世紀から12世紀の仮名文学作品）、「室町時代編 I 狂言」（17世紀成立の狂言台本）のコーパスが公開されており、さらに、「洒落本」（18〜19世紀の小説の一種）、近代雑誌コーパス（19〜20世紀の雑誌）などのコーパスも構築されている。これらのコーパスは全文テキストを含むだけでなく、品詞や語彙素などの形態論情報が付与されているため、多角的な分析が可能になっている。

これらの古い時代の日本語資料は現存しているものが限られているため、一つの時代には特定のジャンルの資料しか利用できないことが多い。そのため、各時代のコーパスから得られる特徴が、その時代に特有のものなのか、そのジャンルに特有なものなのかを区別しづらい場合が少なくなかった。

そこで、本研究では、各時代のコーパスの形態論情報の集計結果から得られる特徴を、『現代日本語書き言葉均衡コーパス』から得られる現代語の多様なジャンルのテキストの特徴と比較して、時代的な差異と位相的な差異の両面から検討する。これにより、『日本語歴史コーパス』を構成するテキストの研究資料としての位置づけを考える。

10 On the possibility of a diachronic speech corpus of Japanese

MARUYAMA Takehiko
Senshu University / NINJAL

Abstract
This study investigates the possibilities of the establishment of a diachronic speech corpus of Japanese. After identifying the conditions and limitations that underlie the process of compiling a diachronic speech corpus, this study discusses its potentiality with analyses of intonation patterns and grammatical expressions.

Keywords: Diachronic speech corpus of Japanese, Okada collection, Danwago data, Intonation patterns, Grammatical expressions

1 Introduction

Since the Corpus of Spontaneous Japanese (CSJ) was released to the public in 2004, research on Japanese spontaneous speech, as opposed to reading aloud, has made drastic progress. CSJ, including 7.52 million words with 651 hours of spontaneous speech, not only provides new research data in the field of the linguistic study of speech such as phonetics, phonology and syntax, but also greatly contributes to a wide range of pursuits in linguistics and related fields, most notably in variation studies in sociolinguistics, and in the development of techniques for speech processing systems such as ASR (Automatic Speech Recognition) and NLP (Natural Language Processing). On the other hand, there is an ongoing requirement to develop a corpus of daily conversation, as most of the spoken data collected in CSJ are monologues. In response to this request, NINJAL (National Institute for Japanese Language and Linguistics) started a new project to establish a new corpus called CEJC, which contains 200 hours of daily conversation in various contexts (Koiso et. al. 2016). When development has been completed and after its release, it will be possible to establish linguistic resources focusing on contemporary spoken Japanese, of both monologues and dialogues, which may lead to further progress in studies of spoken language.

Taking the aforementioned research as a starting point, this study investigates the possibilities of the establishment of a diachronic speech corpus of Japanese. While diachronic corpora usually target written language, the compilation of chronological speech data for the diachronic study of speech holds great prospects. In what manner will it contribute to spoken language studies? After identifying the conditions and limitations that underlie the process of compiling a diachronic speech corpus, this study discusses its potentiality with some case studies.

2 Previous research

In the history of corpus linguistics, there have been very few attempts to compile diachronic speech corpora. The Diachronic Corpus of Present-day Spoken English (DCPSE) is an example of such an attempt, presenting spontaneous speech data of British English from the 1960s to the 1990s.

In 2006, the DCPSE (by the "Survey of English Usage" project at the University College London) was released to the public.[1] This diachronic speech corpus contains colloquial British English from the late 1960s to the early 1990s. The recorded resources between the 1960s and the 1970s were derived from the London-Lund Corpus, whereas the data from the 1990s is from ICE-GB. Each of these two corpora contain approximately 400 thousand words.

All these transcribed texts were annotated with morphological and syntactic information. Aarts et al. (2015) quantitatively clarified that (1) the usages of auxiliary verbs, "must," "may," and "shall" drastically declined within this 30-year period; (2) the usages of "would," "could," and "should" also declined; and (3) the usages of "will" and "can" conversely increased. Although there are some problems underlying this corpus (concerning the validity and representativeness of diachronic speech corpora as will be discussed in Section 3), this is an exemplary attempt at compiling a diachronic speech corpus for the purpose of quantitatively and diachronically analyzing the linguistic changes in spoken language.

Concerning research on existing recorded resources of old spoken Japanese, Shimizu and Kanazawa have collected and analyzed wax cylinder recordings and Standard Playing (SP) records (Shimizu 1988, 1994, 2011, 2014, Kanazawa 1991, 2000, 2015). However, these materials are mainly composed of old recordings of *rakugo* (Japanese traditional comic storytelling), and thus are different from natural speech data.

Old recordings of spoken Japanese can also be found in the "Okada Collection" archive, mainly consisting of political speeches recorded in the early first half of the 20th century. In the early 1950s NINJAL began recording daily conversations with a tape recorder, resulting in approximately 80 hours of recorded material. The contents of this material will be described in Section 4.

3 Conditions and limitations of a diachronic speech corpus

This section will consider the necessary conditions for the realization of a diachronic speech corpus. Three conditions, according to the following key terms: "diachronic," "speech," and "corpus", are identified here.

1 http://www.ucl.ac.uk/english-usage/projects/dcpse/

Concerning the first condition, "diachronic," the corpus must be a collection of speech data from various time periods. It should be well organized in order to enable an analysis of the changes in spoken Japanese. Second, in terms of the "speech" condition, the recorded resources must be preserved so that playback and listening are possible. The essential resource of any speech corpus is the recorded data itself; thus, a collection of transcriptions alone cannot be called a speech corpus in the truest sense. Furthermore, the quality of recorded data should be as clear as possible; particularly concerning conversation, it is preferable to have a multiple track recording. A condition necessary for a true "corpus" is that "a corpus should contain vast range of digitized examples with various information for linguistic retrieval" (Maekawa 2013). This means that it must include not only a variety of electronic recorded data, but various annotations such as transcriptions, morphological information including POS, syntactically parsed information, and various metadata such as information on speakers, recorded time, speaking style, etc.

In reality, however, it is extremely difficult to fulfill all the aforementioned conditions. For instance, the condition of "speech data recorded various periods" is limited due to the fact that recording devices were only developed in the late 19th century, only becoming broadly available in and after the 20th century. This means that a diachronic speech corpus can only target spoken data collected in and after the 20th century.[2] While a diachronic corpus of written language enables us to collect written Japanese in and after the 8th century, a diachronic speech corpus must be extremely limited in scope as well as in quantity. Second, the condition of "good quality recording" is also limited as the quality of recorded materials from an earlier period is comparatively poor and not well-preserved. As explained in Section 4, the conversational speech data collected by NINJAL in the 1950s are sometimes insufficient in volume and marred by noises. Furthermore, the speech data that are not yet digitized may become inaudible in the near future, as the original media will inevitably deteriorate. Therefore, digitizing them is an urgent matter, but also time-consuming and costly. Finally, the condition of "a good quantity of data from various contexts" is also difficult to achieve as the amount of existing speech data is severely limited. Therefore, with respect to quantity and quality of data, a diachronic speech corpus cannot be expected to have the same quantity and quality of data as a large-scale corpus. In addition to the fundamental difficulty of achieving a balance in speech corpora generally (Maekawa 2013), problems are exacerbated when we target historical spoken data. Furthermore, even when original recorded data are extant, difficulties in using them broadly may arise due copyright.

Considering the limitations of preserved spoken data, we must accept that a diachronic speech corpus must be restricted in quantity and quality in its balance and

2 According to Shimizu (2014), the oldest recorded material of Japanese speech (discovered so far) is a reading of the Bible by Ichitaro Hitomi, which was recorded by the Societe d'Anthropologie de Paris on July, 1900, at the Expositions universelles de Paris.

diversity. In other words, it is crucial to collect the existing data in as wide an area as possible. This condition is the same as that faced by linguists working on Old Japanese texts from 8th century, who must conduct research with limited linguistic resources. As long as researchers are using old data, it is unavoidable for them to face the problem of quantitative limitation.

Taking these conditions into account, it is necessary to collect as many types of recorded material as possible and annotate the "metadata" required to categorize them (Maruyama 2012) for the purpose of compiling a single diachronic speech corpus. With regard to material, for example, public speeches and lectures in the collection of "Historical Speech Data"[3] made public on the NDL (National Diet Library) website can be one source of valid data. Annotating metadata requires an investigation into how to categorize the spoken data according to their characteristics. The categories of metadata annotated in the CSJ, such as monologue/dialogue, situation of speech, speaker (gender, age, and place of birth), speaking styles (high or low), and the degree of spontaneity, should be useful. In this manner, determining the criteria is essential to analyze and categorize the various types of speech data in multiple ways.

4 Data

In this section, we will consider the possible types of linguistic research when using a diachronic speech corpus such as the one proposed, also considering several concrete examples of available data to inform the discussion. In addition to the CSJ previously mentioned, the following two spoken resources will also be used as data for analysis:

1. Okada Collection I, *Kichō Ongen* Collection, *Sōryū-sha* Academic Resource Series
2. Recorded data in *Danwago no Jittai* (Research in Colloquial Japanese)

The first set of data will be henceforth referred to as the Okada Collection, and the second will be referred to as the *Danwago* Data. The Okada Collection is a set of spoken data recorded in SP vinyl from the late *Meiji* era (1867–1912) to the beginning of the *Shōwa* era (1926–1989). In total, 18.5 hours of speech data, comprised of 165 original speeches, were digitized and published[4] from among 35 thousand vinyl recordings collected by Mr. Norio Okada. All data are monologues, categorized into political speeches, general lectures, Buddhist sermons, recitations, and so on. Although several unclear segments exist due to low sound quality, the Okada Collection is undoubtedly important as a rare collection of spoken data from approximately 100 years ago.

3 http://rekion.dl.ndl.go.jp/

4 http://www.nichigai.co.jp/database/sp/

Professor Hiroyuki Kanazawa transcribed the Okada Collection in a NINJAL project titled "An Analysis for the Dynamic State of the Contemporary Japanese from Multifaceted Approach" (2009–2015, leader: Professor Masao Aizawa). A collection of papers with the results of this project has been published (Aizawa & Kanazawa 2016). In this study, I have used 109 lectures categorized as "political speech" or "lectures": 14.5 hours of speech data overall. The number of speakers is 76. Table 1 shows examples of speech data.

Table 1: Examples of Speech Data in Okada Collection

Years	**Speaker (Year of Birth)**	**Title of Speech**	**Recorded Time**
1915	Yukio Ozaki (1858)	A Speech by the Minister of Justice, Yukio Ozaki	0:28:10
1916	Shigenobu Ōkuma (1838)	The Power of Public Opinion in Constitutional Politics	0:17:14
1926	Shimpei Gotō (1857)	Ethics of Politics	0:12:54
1931	Tsuyoshi Inukai (1855)	Necessity of a Strong Cabinet	0:04:09
1937	Senjurō Hayashi (1876)	Announcement to Japanese Citizens	0:06:13
1941	Fumimaro Konoe (1891)	Concerning the Conclusion of the Tripartite Pact	0:10:25

The latter data, the *Danwago* Data, is a collection of resources from the 1950s to the 1960s recorded at NINJAL. Since its establishment in 1948, NINJAL has been investigating colloquial Japanese, including the Tokyo dialect and other local dialects. The earliest research using a recording device was done in October 1950 at Shirakawa city in Fukushima prefecture. Beginning in 1952, NINJAL began collecting daily conversations in various contexts. They analyzed intonation patterns, vocabulary, *bunsetsu* (phrases in Japanese), the length and structure of sentences, types of words, and so on. The results were published in three reports, *Danwago no Jittai* (Research in Colloquial Japanese) in 1955 and *Hanashi Kotoba no Bunkei 1, 2* (Research of Sentence Patterns in Colloquial Japanese 1, 2) in 1960 and 1963 (NLRI 1955; 1960; 1963). Approximately 40 hours of speech were recorded, and approximately 30 hours of speech were analyzed in these reports.

Although most of the recorded materials have been currently digitized, they are not well organized for linguistic analysis. The author of this paper has transcribed some

of them, including 33 conversations (in total approximately 19.5 hours) and 21 monologues (approximately 17 hours). Figure 1 illustrates the examples of the spoken data included. Most conversations are chats among laypeople, whereas all the monologues were lectures or talks held at NINJAL.[5]

Conversations: *Kudan High School Students, Three Young People, Kamakura Housewives, Old Men and Women, Fish Shop's Son*

Monologues: *Particles and Auxiliary Verbs, Lecture on Japanese Language, Accent etc. in Japanese, Talk for the 10th Anniversary of NINJAL*

Figure 1. Examples of Speech Data in the *Danwago* Data

Table 2 indicates the data statistics. The Okada Collection is divided into three periods: *Taisho* (from 1915–1926), *Shōwa* 1–9 (1926–1934), and *Shōwa* 10–19 (1935–1944). The *Danwago* Data is separated into two categories: conversations and monologues. The total frequency of words is calculated from the result of morphological analysis (UniDic 2.1.2 + MeCab 0.996) by removing supplementary-symbols (punctuations, brackets, and so on).

Table 2: Data Statistics of the Okada Collection and the *Danwago* Data

	Okada Collection (1915–1944)			***Danwago* Data (1950s–1960s)**	
	1915–1926 (*Taisho*)	1926–1934 (*Shōwa* 1–9)	1935–1944 (*Shōwa* 10–19)	Conversations	Monologues
Number of Files	19	52	38	33	21
Number of Speakers	16	42	30	unknown	21
Time of Recording	3 hours	6 hours	6 hours	19.5 hours	17 hours
Total Number of Words	23,022 words	46,998 words	49,070 words	218,497 words	182,619 words

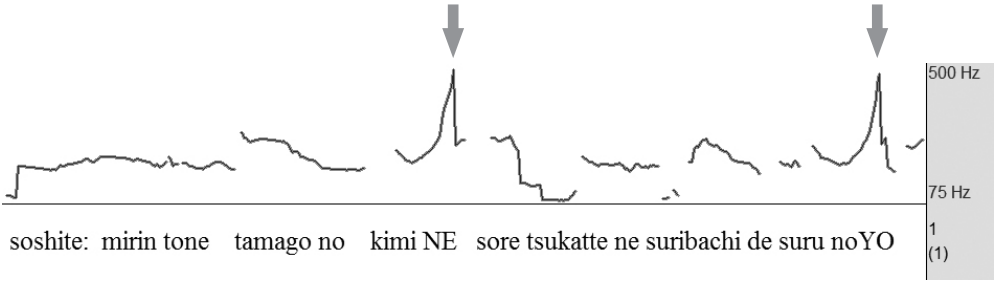
5 Concerning the conversation data, unfortunately, information such as the age of speakers, their occupation, place of birth, accurate dates and places of recording is partly not available.

Here we should note that these two collections do not necessarily cover a wide range of spoken Japanese. The Okada Collection is a collection mostly of political speech, while the *Danwago* Data includes daily conversation and lectures by researchers. It is needless to say that an optimal speech corpus should contain a wide range of speech in various situations, since distributions of intonation, vocabulary, grammatical expression and speaking style may vary in different situations. However, at least at this stage, we have to proceed with analyses using these limited data as a single corpus, since no other sizable collections of speech data have been found. The following sections present concrete case studies using the Okada Collection, the *Danwago* Data, and the CSJ to discuss the possibilities of a diachronic speech corpus.

5 Analysis of the Okada Collection and *Danwago* Data

5.2 Analysis of intonation

What is examined here is the characteristic pattern of intonation seen in the *Danwago* Data. Figure 2 shows a pitch contour of an utterance which appeared in an excerpt of recorded material called "Three Ladies" from 1957. The utterance is "*soshite: mirin tone tamago no kimi NE sore tsukatte ne suribachi de suru noYO*" (I added syrup and egg yolk, and then used a mortar to grind and mix them). We can see in the contour that the pitches on *NE* at the end of a phrase and on *YO* at the end of the utterance rise very rapidly. It is certain that this rising intonation does not signify a question addressed to the listener.

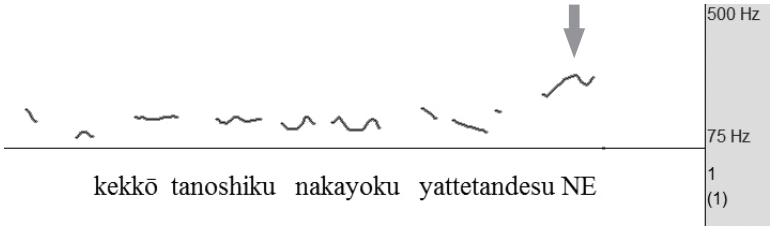


"*soshite: mirin tone tamago no kimi NE sore tsukatte ne suribachi de suru noYO*"

Figure 2. Rapid Rising Intonation (*Danwago* Data: "Three Ladies")

This rising intonation in Figure 2 reminded me of scenes spoken by actresses in old Japanese films from the 1950s. For example, in the film "Tokyo Story" (directed by Yasujirō Ozu in 1953), rising intonations like those in Figure 2 are frequently observed in the utterances by the actress Setsuko Hara. This suggests that such intonation patterns might have been natural for women in the 1950s.

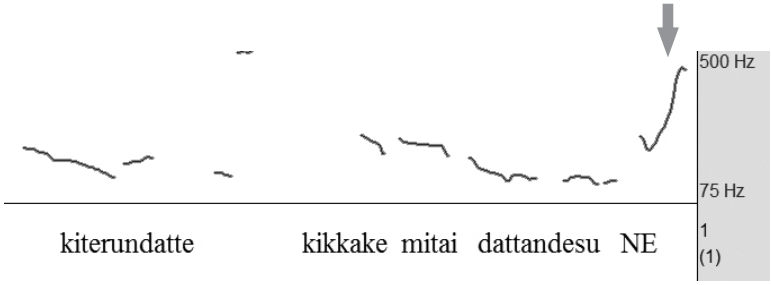
Rising intonations at the end of phrases in non-interrogative contexts can be seen even in the CSJ, which contains utterances by contemporary Japanese speakers. For instance, in Figure 3 the utterance "*kekkō tanoshiku nakayoku yattetandesu NE* (I had a fun and convivial time)" shows a rising intonation at the end. This, however, is not equivalent to the drastic rise seen in Figure 2.



"*kekkō tanoshiku nakayoku yattetandesu NE*"

Figure 3. Rising Intonation (CSJ:S05F1600)

On the other hand, Figure 4 shows a rising intonation at the end of a phrase seen in the CSJ, "*kiterundatte kikkake mitai dattandesu NE* (So it seems that this was the start of (it) coming, you know)," which seems to be very much similar to the intonation pattern in Figure 2.



"*kiterundatte kikkake mitai dattandesu NE*"

Figure 4. Rising Intonation 2 (CSJ:S01F1522)

Here, it is important to focus on the age gap between the speakers in Figures 3 and 4. The speaker in Figure 3 was in her late 20s (year of birth: the early 1970s) at that time of recording, whereas the age of the speaker in Figure 4 was approximately 50 (year of birth: the late 1940s). This means that there is an age difference of 25 years between them. At this point, we note that drastic rising intonations at the end of phrases like Figure 2 and 4 can be heard in the utterances of old women even in contemporary Japanese. It seems that these intonations occur in contexts in which older women, either of higher social

standing or with pretenses to such, speak elegantly. Furthermore, when my daughter listened to the utterance in Figure 2 she said, "It sounds like my grandmother speaking."

Supposing that the female speaker in Figure 2 was in her mid-20s at the time of recording in the 1950s, she would now be in her 80s. We can infer that the informants for the *Danwago* Data even now partially preserve and utilize the same intonation patterns from that era. For today's younger generation, however, an example like that in Figure 2 sounds like an utterance by an elderly woman or an actress in an old Japanese film.

Assuming a language change in which a new intonation pattern emerges among the younger generation thus replacing an older one, there is a point where rising intonations such as those in Figures 2 and 4 decline, and cannot be used by the younger generation. However, we must make recourse to more recorded resources and make quantitative and cross-sectional analyses in order to locate the era when such intonation patterns disappear.

5.2 Analysis of grammatical expressions: auxiliary verb *masuru*

This section focuses on grammatical expressions employing the auxiliary verb *masuru*. Hattori (2011) pointed out that the form of *masuru* began to change to *masu* at the beginning of the early modern period (from the *Azuchi-Momoyama* period (1568–1600) to the *Edo* period (1603–1867)), but it can still be frequently observed in the Okada Collection. The examples below show usages of *masuru* appearing at the end of a sentence (EOS) in (1), and appearing in verb phrases in both a *to*-clause and a *ga*-clause in (2).

(1) *Meiji 17 nen, sentē hēka no gorē o-itsutsu no koro to kioku o itashite orimasuru*
In Meiji 17th (1884), I remember that it was the time when the emperor was 5 years old.
(Akinobu Manabe, "*Taiko tenno go-yōji o shinobi tatematsurite,*" 1927)

(2) *Konnichi, shinbun nazo o mimasuru to, makoto ni nagekawashī koto ga takusan arimasuru ga, hitotsu ni ryōshin o kaeriminaide akuma no koe ni damasarete…*
Nowadays, while reading the newspaper, we find many lamentable matters, this is firstly because we do not care about conscience so that we are deceived by the devil…
(Motojiro Makino, "*Ryoshin undō no daiissei*", *Shōwa* 10s (1935–1944))

Such usages of *masuru* are also seen in the *Danwago* Data in the verb phrases of *kara*-clauses, *to*-clauses, and *keredomo*-clauses, as exemplified below.

(3) *Hijō ni yosan no kyūkutsu na, a:, jidai de arimasuru kara, e:, soredemotte….*
Now, it is the time that we must be very tight in budget, so…
(Yūzō Yamamoto, "Talk for the 10th anniversary of NINJAL," 1959)

(4) *Rajio news no kakikata toyū yōna hon o mi<u>masuru</u> to, e:, news niwa…*
Reading through a book titled "how to write a script for radio news" well, news…
(Kanji Hatano, "Talk for the opening anniversary of new office," 1962)

(5) *Atarashii jibiki ga 20-man go o shūsaisuru to kaite ari<u>masuru</u> keredomo, sononaka no 2 man go shika…*
Although it says that a new dictionary contains 200 thousand words, only 20 thousand of them…
(Ōki Hayashi, "Talk for the opening anniversary of new office," 1962)

During the lectures and talks in (1) to (5), the speakers also used an auxiliary verb, *masu*, such as "*kangeki itashiteoru shidai de ari<u>masu</u>* (I am very moved)" (Manabe), "*Rongo no uchi de attaka to omoi<u>masu</u> ga* (I think it was in the Analects of Confucius)" (Makino), "*muzukashiinde ari<u>masu</u> kara* (because it is difficult)" (Yamamoto), "*kiite ori<u>masu</u> to* (Listening to it,…)" (Hatano) and "*sa mo ari<u>masu</u> keredomo* (although there is a gap)" (Hayashi). Given that the two forms are functionally equivalent polite verbal endings, this means that *masu* and *masuru* are morphological variants. [6]

Table 3 shows the number of the auxiliary verbs *masu* and *masuru* that appeared in monologues in the Okada Collection, the *Danwago* Data, and the Core of CSJ (177 lecture talks, in total 41 hours).

Table 3: The number of *masu* and *masuru* that appeared in each data set

	Okada Collection						***Danwago* Data monologues**		**CSJ Core monologues**	
	1915–1926 (*Taisho*)		**1926–1934 (*Shōwa* 1–9)**		**1935–1944 (*Shōwa* 10–19)**					
masu	271	(86.6%)	752	(89.8%)	903	(92.9%)	3,918	(98.8%)	5,604	(100%)
masuru	42	(13.4%)	85	(10.2%)	69	(7.1%)	48	(1.2%)	0	(0%)

We can see that although *masuru* constituted 13.4% of all polite verbal endings in the *Taisho* era, it was eventually replaced by *masu*. The contemporary Japanese corpus, the CSJ, did not have any example of *masuru*.

I then proceeded to analyze the words that follow *masuru* in the Okada Collection and the *Danwago* Data. Table 4 presents the results for the top 10 expressions that appeared most frequently after *masuru*.

6 The years of birth for the speakers are Manabe (1878, Meiji 11), Makino (1874, Meiji 7), Yamamoto (1887, Meiji 20), Hatano (1905, Meiji 38) and Hayashi (1913, Taisho 2).

Table 4: List of Words that Appear after *masuru*

Okada Collection						*Danwago* Data monologues	
1915–1926 (*Taisho*)		1926–1934 (*Shōwa* 1–9)		1935–1944 (*Shōwa* 10–19)			
11	*to* (conjunctive particle)	28	*ba* (conjunctive particle)	16	。(EOS)	13	*keredomo* (conjunctive particle)
6	。(EOS)	11	*ga* (conjunctive particle)	13	*ga* (conjunctive particle)	8	*kara* (conjunctive particle)
5	*ga* (conjunctive particle)	10	noun phrase	8	*to* (conjunctive particle)	7	*to* (conjunctive particle)
4	*yueni*	7	*to* (conjunctive particle)	7	*ba* (conjunctive particle)	6	*shi* (conjunctive particle)
3	*naraba*	5	*ni* (case-marking particle)	5	noun phrase	6	noun phrase
3	*keredomo*	5	*kara* (conjunctive particle)	4	*no* (nominal particle)	6	*ba* (conjunctive particle)
3	*kara* (conjunctive particle)	5	。(EOS)	4	*kara* (conjunctive particle)	4	*ga* (conjunctive particle)
2	*no* (nominal particle)	4	*keredomo*	4	*ka* (sentence-final particle)	2	*ni* (auxiliary verb)
2	*ni* (case-marking particle)	3	*ya* (sentence-final particle)	3	*ya* (sentence-final particle)	1	*tameni*
1	*ba* (conjunctive particle)	3	*toiu* (quotation)	3	*ni* (case-marking particle)	1	*yueni*

While there are some instances of *masuru* appearing at EOS in the Okada Collection, no instance of *masuru* at EOS is observed in the *Danwago* Data. This result coincides with the view of Hattori (2011) that a prominent characteristic of *masuru* is that it never appears at EOS. This observation was obtained through his analysis of this auxiliary verb as it appears in the numerous amount of data of minutes of the National Diet. Further results reveal that although there is a low frequency of the conjunctive particle *keredomo* after *masuru* in the Okada Collection, *keredomo* achieves its greatest

frequency in the *Danwago* Data. It seems that in time, the form tended to be avoided for terminating a sentence, and became preferable as the style for connecting with the conjunction *keredomo*.

What we can infer from this is that the auxiliary verb *masuru* tended to be used in formal monologues (e.g., lectures, talks, and formal speeches) by a relatively small number of people. Here, the speakers' usages are strongly affected by their respective dates of birth. In time, the younger generations ceased to use *masuru* in their speech, yet it is impossible to know exactly when it began to disappear in monologues at this stage of research due to a lack of data.[7] To clarify changes in the process of alternation between *masuru* and *masu*, we need to supplement our data with more recording material from the blank period, from more registers and from a greater variety of speech situations.

5.3 Analysis of grammatical expressions: sentence final particles

This section analyzes the frequency and combination of sentence-final particles. We now take a look at examples of sentence-final particles in the conversations in the *Danwago* Data.

(6) a. *Watashi dattara Kyūshū ni ikitai<u>wa</u>.* ("Sagami Female College student")
If I were you, I would go to Kyushu.
b. *Harau to shitara, taihen desu <u>wane</u>.* ("*Tomonokai* member")
If I must pay it, it is hard.
c. *Anta n toko no o-sakana, oishii <u>wayo</u>.* ("Fish shop's son")
Your fish is delicious.
d. *Sensei to o-hanashi shite kimashita <u>noyo</u>.* ("Kamakura housewife")
I went to talk with a teacher.

All these examples contain sentence-final particles that are commonly understood as appearing at the end of a woman's utterance. Even though the utterances they attached to are not interrogatives, the particles *wa* and *noyo* have rising intonations. It is certainly extremely rare to observe analogous usages of *wa*, *wane*, *wayo*, and *noyo* in the conversations of younger people in contemporary Japanese. If they appear at all, these forms will most likely appear as features of a "role language" employed when its speakers take on the role of older women of higher social standing.

7 Hattori (2011) reports that the usages of *masuru* can be seen in the minutes of the National Diet even today, albeit in a smaller number.

In contrast, supposing the speakers of these utterances to be old women in the present day, the utterances all sound rather natural. To offer my own view, using *wa* and *noyo* with rising intonations is quite natural for older women when they speak elegantly.

I will now compare the conversations between the *Danwago* Data and the dialogue part of the CSJ (58 conversations, in total 12 hours). Table 5 shows which sentence-final particles appeared at the end of utterances, and indicates their frequency in both data for comparison.

Table 5: Sentence-Final Particles at the end of Utterances

	-wa	*-wane*	*-wayo*	*-noyo*	*-yo*	*-ne*
***Danwago* Data (conversations)**	153	296	116	296	1,675	5,752
CSJ (conversations)	4	2	0	0	391	4,165

Certainly, the numbers (and also the ratios) of instances of *wa, wane, wayo,* and *noyo* in the *Danwago* Data are much greater than those in CSJ. Since the intonation patterns are not considered in these totals, it is difficult to be absolutely certain, but the prediction is that instances with rising intonation are even less frequent in the CSJ particularly.

In any case, just as the instance of the rising intonation patterns shown in the Section 5, these (combinations of) sentence-final particles cannot be seen in the utterances of today's younger generations. This indicates that the usage of these grammatical expressions had gradually disappeared sometime earlier. However, at this stage it is quite difficult to identify the period when such a change in language occurred. To describe the dynamics of language change in accurate detail, it is necessary to supplement our data with more material from the 20th century, and from a greater variety of speech situations.

6 Concluding remarks

In this study, the possibility of compiling and analyzing a diachronic speech corpus of Japanese has been discussed. First, the conditions of a diachronic speech corpus were examined from the viewpoints of the key terms "diachronic," "speech," and "corpus." Also, the limitations of compiling a diachronic speech corpus of Japanese were identified; the amount of old recorded materials is limited, so only a corpus compiled from limited resources can be used for analysis. This is a general constraint which linguists must cope with when analyzing resources for the study of old language.

Following this, several case studies were presented, analyzing intonation patterns and grammatical expressions, auxiliary verbs and sentence-final particles, using three different recorded resources: the Okada Collection, the *Danwago* Data, and the CSJ. The analyses clarified some interesting linguistic findings in spoken Japanese, including

rapid rising intonation, the auxiliary verbs *masu* and *masuru*, and (combinations of) sentence-final particles.

When attempting a diachronic analysis in order to observe historical change in spoken Japanese, a serious problem arises due to the insufficiency and imbalance of existing recorded data, as seen earlier. In this study three different speech data sets were treated as a single diachronic speech corpus, however, the situations of speech and speaking styles vary among these three; the Okada Collection includes political speeches, the *Danwago* Data contains academic lectures and daily conversation, and the CSJ is mainly composed of academic and casual monologues.

In any case, there is no question about the significance of compiling a diachronic speech corpus and followed by analysis. To solve the problem of imbalance, it is crucial that we collect more varied recorded resources in order to establish a better diachronic speech corpus so that studies on spontaneously spoken Japanese in various eras can achieve full fruition.

Acknowledgement

This study is supported by the JSPS (Japan Society for the Promotion of Science), the Grant-in-Aid for Scientific Research (no. 24520523), and Grant-in-Aid for Collaborative Research Project of NINJAL "A multifaceted study of spoken language using a large-scale corpus of everyday Japanese conversation."

Literature

Aarts, B., Bowie, J., and Wallis, S. (2015) Profiling the English verb phrase over time: modal patterns. In: Taavitsainen, I., Kytö, M., Claridge, C., and Smith, J. (eds.) *Developments in English: expanding electronic evidence*, 48–76. Cambridge: Cambridge University Press.

Aizawa, M. and Kanazawa, H. (eds.) (2016) *Senzenki SP record ga hiraku nihongo kenkyū* (Japanese Language Studies by Analyzing SP Records Before World War II). Tokyo: Kasamashoin.

Hattori, T. (2011) *Washa no shussē nendai to hatsuwa jiki ni motozuku gengo henka no kenkyū: kokkai kaigiroku o riyō shite* (A Study on Language Change Based on the Speakers' Date of Birth and Years of Utterances: An Analysis of Minutes of the National Diet). *Kēryō Kokugogaku*, 28 (2): 47–62.

Kanazawa, H. (1991) *Mēji-ki Ōsaka-go shiryō toshite no rakugo sokkibon to SP record*. (Storybooks of Rakugo and SP Records as the Resources of Osaka Dialect in Meiji Era). *Kokugogaku*, 167: 15–28.

Kanazawa, H. (2000) *Rokuon shiryō no rekishi to sono kanōsē*. (The History of Recorded Resources and Their Possibilities). *Nihongogaku*, 19 (11): 197–208.

Kanazawa, H. (2015) *Rokuon shiryō niyoru kindai-go kenkyū no ima to korekara* (The Current Studies on Modern Language by Recorded Resources and Its Future). *Nihongo no Kenkyū*, 11 (2): 133–140.

Koiso, H., Tsuchiya, T., Watanabe, R., Yokomori, D., Aizawa, M., and Den, Y. (2016) Survey of Conversational Behavior: Towards the Design of a Balanced Corpus of Everyday Japanese Conversation, *Proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC 2016)*, pp.4434-4439.

Koiso, H., Tanaka, Y., Watanabe, R., and Den, Y. (2016) A Large-Scale Corpus of Everyday Japanese Conversation: On Methodology for Recording Naturally Occurring Conversations, *Proceedings of LREC 2016 workshop on casual talk among humans and machines*, pp. 9-12.

NLRI (The National Language Research Institute) (1955) *Danwago no jittai* (Research in Colloquial Japanese). NINJAL Report 8. NINJAL.

NLRI (The National Language Research Institute) (1960) *Hanashi kotoba no bunkē (1): taiwa shiryō niyoru kenkyū* (A Research for Making Sentence Patterns in Colloquial Japanese: On Materials in Conversation). NINJAL Report 18. Tokyo: Shūēshuppan.

NLRI (The National Language Research Institute) (1963) *Hanashi kotoba no bunkei (2): dokuwa shiryō niyoru kenkyū* (Research of Sentence Patterns in Colloquial Japanese: On Materials in Speech). NINJAL Report 23. Tokyo: Shūēshuppan.

Maekawa, K. (2013) *Corpus no sonzai igi*. (Raison d'être of Corpus). In: Maekawa, K. (ed) *Kōza Nihongo Corpus 1 Corpus Nyūmon* (Lecture Series: Japanese Corpus 1: Introduction): 1–31. Tokyo: Asakurashoten.

Maruyama, T. (2012) *Daikibo Corpus no riyō to meta data no yakuwari*. (Usage of Large-scale Corpus and the Role of Meta Data). The First Corpus Nihongaku Workshop Proceedings, 203–210.

Shimizu, Y. (1988) *Tokyo-go no rokuon shiryō*. (Recorded Resources of Tokyo Dialect). *Kokugo to Kokubungaku*, 65 (11): 129–143.

Shimizu, Y. (1994) *Rokuon shiryō ni kiku 20-seiki hajime no Tokyo-go*. (Tokyo Dialect in early 20th Century: An Analysis of Recorded Resources). *Annual Report of the Institute for Japanese Culture and Classics*, Kokugakuin University, 73: 191–230.

Shimizu, Y. (2011) *Ōbē no rokuon archives: shoki nihongo rokuon shiryō shozō kikan o chūshin ni*. (Recorded Archives in Europe and the USA: Institutes to Collect Recorded Resources in Early Japanese). *Kokubun Mejiro*, 50: 29–19.

Shimizu, Y. (2014) *Hyakunen mae no nihongo o kiku*. (Listening to Japanese in 100 years ago). Japan Women's University.

要旨（Abstract in Japanese）

「通時音声コーパスの可能性」

丸山岳彦（専修大学／国立国語研究所）

「話し言葉の通時コーパスは実現可能か」という問題について論じる。通時コーパスと言えば、通常、書き言葉を対象としたものが想定される。それでは、話し言葉を対象とした通時コーパスは実現可能であろうか。本稿では、「通時」「音声」「コーパス」という3つの条件について検討し、「通時音声コーパス」の実現によってどのようなことが明らかになるかを示す。大正から昭和前期にかけて録音されたSPレコードの音源資料、国立国語研究所において1950年代に録音された資料などを分析対象として、そこに見られるイントネーションの型、文法形式について分析の事例を示し、話し言葉の経年変化を追うための「通時音声コーパス」が持つ可能性について論じる。

Contributors

ABEKAWA Takeshi (Eng.D. Tokyo Institute of Technology 2006) is Associate Professor by Special Appointment in the Digital Content and Media Sciences Research Division of the National Institute of Informatics, Tokyo. His research interests include natural language processing and computational linguistics.

Andrej BEKEŠ (Litt.D. University of Tsukuba 1986) is Professor Emeritus of the University of Ljubljana and one of the co-founders of the Department of Asian Studies at the University of Ljubljana, Faculty of Arts. His research interests cover text linguistics, text pragmatics, language policies, writing systems and Japanese as a second language. His publications include *Tekusuto to shintakusu* [Text and Syntax] (Kurosio Publishers 1987), *Text and boundary: a sideways glance at textual phenomena in Japanese* (Ljubljana University Press 2008), and as a co-author *Nihongo bunkei jiten* [A handbook of Japanese grammar patterns for teachers and learners] (Kurosio Publishers 1998/2015). He is also the founding editor of the journal Acta Linguistica Asiatica.

HASEBE Yoichiro is Associate Professor of the Faculty of Global Communications, Doshisha University. His academic interests include cognitive linguistics and corpus linguistics. He also engages in development of computer systems that process language corpora for research and educational purposes. His publications include 'A Cognitive Approach to Compound *Kango* VNPs in Japanese' in T. Ishiguro and K-K. Luke (eds.) *Grammar in Cross-Linguistic Perspective: The Syntax, Semantics, and Pragmatics of Japanese and Chinese*, 9-42. (Peter Lang 2012) and 'Design and Implementation of an Online Corpus of Presentation Transcripts of TED Talks' *Procedia: Social and Behavioral Sciences* 198(24), 174-182, 2015.

Bor HODOŠČEK (Eng.D. Tokyo Institute of Technology 2013) is Associate Professor at the Graduate School of Language and Culture, Osaka University. He is currently working on the quantitative modeling of register in Japanese as well as exploring its role in writing assistance systems. His interests include quantitative linguistics, natural language processing, and educational technology.

LEE Jae-Ho (Ph.D. Kyoto University 2008) is Professor of the Graduate School of Japanese Applied Linguistics at Waseda University in Japan. His current research interests include a Corpus-Based Study of Japanese and a Methodology for E-learning. He has extensively published books such as *An Introduction to Corpus-Based Research for Japanese language education* (Kurosio Publishers 2018), and *ICTxJapanese Language Education: Theory and Practice* (Hituzi Syobo 2019).

MARUYAMA Takehiko (Ph.D. International Christian University 2013) is Professor of the School of Letters, Senshu University, and a visiting professor of the Spoken Language Division, National Institute for Japanese Language and Linguistics (NINJAL). He has worked on Japanese corpus linguistics, especially the methodology of how to design and compile Japanese written and spoken corpora. He also has an interest in grammatical studies of various disfluencies observed in Japanese spontaneous speech. His publications include *Kōpasu Nyūmon* [Introduction to Corpus Study] (Asakura Shoten 2013), *Kakikotoba Kōpasu: Sekkei to Kōchiku* [Written Corpus: Design and Compilation] (Asakura Shoten 2014), *Hanashikotoba Kōpasu: Sekkei to Kōchiku* [Spoken Corpus: Design and Compilation] (Asakura Shoten 2015).

NISHINA Kikuko (Ph.D. Tokyo Institute of Technology 1997) is Professor Emerita at the Tokyo Institute of Technology. She has worked on Japanese language teaching methods for second language learners specializing in Technical Japanese CALL (Computer Assisted Language Learning). Her publications include *Shokyū Bunkei de Manabu Kagaku Gijutsu no Nihongo* [Introduction to Technical Japanese] (3A Network 2007), *Construction of Speech Database for Second Language Learning of Japanese* (Publishing House Japan 2010), and *Gakushū Shien Shisutemu no Tame no Setsuzoku Hyōgen Jiten Kōchiku* [Construction of a Connectives Dictionary for Academic Writing Assistance System] (Mathematical Linguistics 31/2 2017).

OGISO Toshinobu (Eng.D. Nara Institute of Science and Technology 2011) is Professor at the National Institute for Japanese Language and Linguistics in Tokyo. His research interests are informatics (intelligent informatics) and linguistics, in particular Japanese linguistics. His publications include 'Development of Dictionaries for Morphological Analysis of Pre-Modern Japanese Aiming at Construction of the Diachronic Corpus' (*IPSJ SIG Notes* 2011(6) 1-41, 2011), 'Morphological Analysis of Kana Literature in Early Middle Japanese' (*Nihongo no kenkyū* [Japanese Language Studies] 9(4) 49-62, 2013), *A Study on Morphological Annotation for the Japanese Diachronic Corpus* (Nara Institute of Science and Technology 2014), 'Japanese Grammar Projected from Natural Language Processing: Statistical Machine Learning and "Grammar"' (*Journal of Japanese Grammar* 16(2) 20-31, 2016).

SAKUMA Mayumi is Professor Emerita of Waseda University. Her research interests are in Japanese linguistics (discourse and text analysis) and Japanese-language pedagogy. She is the editor of *Bunshōkōzō to Yoyakubun no Shosō* [Japanese discourse structures and aspects of summary texts] (Kurosio Publishers 1989), the co-author of *Nihongo no Bunpō 4: Fukubun to Danwa* [Japanese grammar 4: Complex sentences and discourse] (Iwanami Shoten, 2002), the editor of *Asakura Nihongo Koza7:Bunshō・Danwa* [Asakura

library on Japanese linguistic studies 7: Japanese discourse and text analysis] (Asakura Shoten 2003), and the editor of *Kōgi no Danwa no Hyōgen to Rikai* [The Expression and comprehension of Japanese lectures] (Kurosio Publishers 2010).

Irena SRDANOVIĆ (Ph.D. Tokyo Institute of Technology 2009; Ph.D. University of Ljubljana 2015) is Associate Professor at the Faculty of Humanities Juraj Dobrila University of Pula (Croatia), where she established the Department of Asian Studies and the Japanese undergraduate and graduate degree study program. Her research interests are corpus linguistics and lexicography, Japanese language education, lexical semantics, sociolinguistics, and pragmatics. She is the author of *Kolokacije in kolokacije na daljavo v japonskem jeziku: korpusni pristop* [Collocations and distant collocations in Japanese language: corpus based approach] (Ljubljana University Press 2016), a co-author of *Udžbenik japanskoga jezika Ippo ippo* [Japanese language textbook Ippo ippo] (Juraj Dobrila University of Pula Press 2018) and a co-editor of *Digital resources for learning Japanese* (Bononia University Press 2018).

SUNAKAWA Yuriko (Ph.D. University of Tsukuba 2005) is Professor Emerita of the University of Tsukuba. She works on Japanese grammar, Japanese discourse analysis and Japanese as a second language. Her publications include *Bunpō to Danwa no Setten* [Linkage between Grammar and Discourse] (Kurosio Publishers 2005), *Nihongo Kyōiku Kenkyū e no Shōtai* [Invitation to the Research for Japanese Teaching] (Kurosio Publishers 2010) and *Shin Nihongo Kyōiku no tame no Kōpasu Nyūmon* [Revised edition of How to Use Corpora in Japanese Language Teaching] (Kurosio Publishers 2018).

Polly SZATROWSKI (Ph.D. Cornell University 1985; Ph.D. University of Tsukuba 1991) is Professor of Japanese Language and Linguistics at the University of Minnesota. Her research interests include conversation/discourse analysis, sociolinguistics, and the emergence of grammar in discourse. Her publications include *Nihongo no Danwa no Kōzō Bunseki - Kanyū no Danwa no Sutoratejī no Kōsatsu* [Structure of Japanese Conversation - Invitation Strategies] (Kurosio Publishers 1993), and three edited volumes: *Hidden and Open Conflict in Japanese Conversational Interaction* (Kurosio Publishers 2004), *Storytelling across Japanese Conversational Genre* (John Benjamins 2010), and *Language and Food: Verbal and Nonverbal Experiences* (John Benjamins 2014).

TAKASAKI Midori is Professor Emerita of Ochanomizu University. She has research interests in Japanese discourse analysis, especially in the area of cohesion. She is the co-author of *Nihongo Zuihitsu Tekusuto no Shosō* [Characteristics of Japanese Essays Texts] (Hituzi Syobo 2007), and the editor of *Taishō-ki "Chuōkōron", "Fujinkōron" no Gairaigo Kenkyū* [Study of Borrowed Words of "Chuōkōron" and "Fujinkōron" in Taishō-era] (Fuzanbō International Publishers 2019).

YAGI Yutaka (M.Eng. Tokyo Institute of Technology 2001) is a systems engineer at Picolab Co., Ltd. His research interests include natural language processing and language education.

Name Index