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