



## Current Issues on Japanese Corpus Linguistics



Takehiko Maruyama (NINJAL / University of Oxford)

19<sup>th</sup> March 2015, 10.30-12.00, 16.20-17.50

Over the last decade various Japanese corpora have been developed rapidly, and the number of linguistic studies based on them is also increasing. Japanese corpora include a large amount of real examples taken from texts and/or utterances, and thus are useful as "Language Resources" for many aspects of Japanese linguistics, Japanese language teaching, language processing, and related fields.

It was the 1950s in which we can find pioneering works based on Japanese "corpora". Sixty years later, Japanese corpora today can be classified into written, spoken, learner, historical, multi-modal corpora, etc., and we can choose any corpora according to our research goals. Also, various search tools to retrieve linguistic expressions from the corpora have been developed.

In this talk, I will introduce the 60-year history of Japanese corpora, how they were developed and used. I will also show what kind of corpora we can access now, how to use them, and what it will bring to the new era of Japanese corpus linguistics.

**Dr Takehiko Maruyama** is an associate professor at the Department of Corpus Studies, National Institute for Japanese Language and Linguistics. He is currently a Visiting Academic at the Faculty of Oriental Studies, University of Oxford. His research interest is on Japanese corpus linguistics, especially grammatical studies of spoken Japanese.

### Selected Publications:

1. Maruyama, Takehiko (2014). A Corpus-based Study of Colloquial Japanese: Retrospect and Prospect. The 14th International Conference of EAJS, Ljubljana, August 2014.
2. Maekawa, Kikuo, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, Yasuharu Den (2014). Balanced corpus of contemporary written Japanese, *Language Resources and Evaluation*, 48, 345-371. Springer.
3. Maruyama, Takehiko (2013). Analysis of Parenthetical Clauses in Spontaneous Japanese. *Proceedings of DiSS 2013 The 6th Workshop on Disfluency in Spontaneous Speech*, 44-48, Stockholm, August 2013.
4. Maruyama, Takehiko (2012). Speech segmentation by clausal and non-clausal boundaries in Japanese. *The Fifth International Conference on Cognitive Science, Spoken discourse corpora as a window on cognitive mechanisms of speech production*, 787-783, Kaliningrad, June 2012.



## コーパス日本語学の現在

丸山岳彦（国立国語研究所・オックスフォード大学）

2015年3月19日（木） 10:30~12:00、16:20~17:50

この10年の間に日本語コーパスの整備が急速に進んだことにより、コーパスを利用した日本語研究の数が飛躍的に増えてきました。実際に書かれたり話されたりした日本語の用例を大量に収録した日本語コーパスは、言語研究のみならず、言語教育や言語処理など、さまざまな分野において有益な「言語資源」としての役割が期待されます。

1950年代という極めて早い時期に開始された日本語のコーパス言語学は、当時から約60年が経過した現在、書き言葉コーパス、話し言葉コーパス、学習者コーパス、歴史コーパス、マルチモーダルコーパスなど、多様な種類が林立する状態となり、研究や利用の目的に応じてコーパスを選ぶことが可能になりました。また、コーパスを検索するための各種ツールの開発も進んでいます。

本講義では、コーパス日本語学の過去・現在・未来についてお話ししたいと思います。日本語コーパスがどのような歴史の中で開発されてきたのか、現在利用できる日本語コーパスにはどのようなものがあり、どのように使えばよいのか、といった点について、実際の検索デモンストレーションを交えながら、示したいと思います。

**丸山岳彦（まるやまたけひこ）** 国立国語研究所言語資源研究系准教授。2000年、神戸市外国語大学大学院国際交流専攻修了。博士（学術）。2000年、ATR音声言語通信研究所研究員。2004年、国立国語研究所研究開発部門研究員。2011年より現職。2014年、オックスフォード大学客員研究員。専門は日本語コーパスを用いた現代日本語文法の研究。特に「非文」や「不適格な文」が散見される話し言葉のデータから、一定の秩序や規則を見出す作業を通して、話し言葉の文法研究の可能性を探っている。

**【主要著書】**『講座 日本語コーパス 3 話し言葉コーパス —設計と構築—』（共著、2015年、朝倉書店）、『講座 日本語コーパス 2 書き言葉コーパス —設計と構築—』（共著、2014年、朝倉書店）、『話し言葉と書き言葉の接点』（共著、2014年、ひつじ書房）、『日本語複文構文の研究』（編著、2014年、ひつじ書房）、『講座 日本語コーパス 1 コーパス入門』（共著、2013年、朝倉書店）、『はじめて学ぶ日本語学』（共著、2011年、ミネルヴァ書房）、『講座 ITと日本語研究 5 コーパスの作成と活用』（共著、2011年、明治書院）、『シリーズ 文と発話 第3巻 時間の中の文と発話』（共著、2007年、ひつじ書房）

**【最近の主要論文】**『『日本語話し言葉コーパス』に基づく挿入構造の機能的分析』（『日本語文法』14巻1号、日本語文法学会、2014年）「作文の文体情報 —『現代日本語書き言葉均衡コーパス』から見えるもの—」（『日本語教育』140号、日本語教育学会、2009年）