# Glottometrics 10
# 2005

## Corpus Studies on Japanese Kanji

Guest Editor

**Katsuo Tamaoka**
*Hiroshima University, Japan*

**Hituzi Syobo and RAM-Verlag**

**Tokyo, Japan/ Lüdenscheid, Germany**

**ISSN 2625-8226**

# Glottometrics

**Glottometrics** ist eine unregelmäßig erscheinende Zeitdchrift (2-3 Ausgaben pro Jahr) für die quantitative Erforschung von Sprache und Text.
**Beiträge** in Deutsch oder Englisch sollten an einen der Herausgeber in einem gängigen Textverarbeitungssystem (vorrangig WORD) geschickt werden.
Glottometrics kann aus dem **Internet** heruntergeladen werden (**Open Access**), auf **CD-ROM** (PDF-Format) oder als **Druckversion** bestellt werden.

**Glottometrics** is a scientific journal for the quantitative research on language and text published at irregular intervals (2-3 times a year).
**Contributions** in English or German written with a common text processing system (preferably WORD) should be sent to one of the editors.
Glottometrics can be downloaded from the **Internet (Open Access)**, obtained on **CD-ROM** (as PDF-file) or in form of **printed copies.**

## Herausgeber – Editors




**Bestellungen** der CD-ROM oder der gedruckten Form sind zu richten an

**Orders** for CD-ROM or printed copies to RAM-Verlag RAM-Verlag@t-online.de

**Herunterladen/ Downloading:** https://www.ram-verlag.eu/journals-e-journals/glottometrics/

Die Deutsche Bibliothek – CIP-Einheitsaufnahme
Glottometrics. 10 (2005), Lüdenscheid: RAM-Verlag, 2005. Erscheint unregelmäßig.
Diese elektronische Ressource ist im Internet (Open Access) unter der Adresse
https://www.ram-verlag.eu/journals-e-journals/glottometrics/ verfügbar.
Bibliographische Deskription nach 10 (2005) **ISSN 2625-8226**

# Contents
# Glottometrics 10, 2005



**Abstract:** This paper traces the historical development of kanji, the Chinese characters used in the Japanese orthographic system. The paper outlines the structural principles which underlie their composition, both in respect to single kanji and to their combination in compound words. Discussion also pays attention to their usage and frequency, as well as to the various script reforms that have affected their number and deployment. Lastly, commentary on their role in the development of Japanese psycholinguistics and the relevance of this work to psychological studies of language in general is offered.



**Abstract:** The present study investigates the relationship between of Japanese kanji strokes and their printed-frequencies of occurrence, compositional asymmetry and kanji multiple readings. First, distributions of kanji strokes in both samples of the 1,945 basic kanji and of 6,355 kanji appearing in the *Asashi Newspaper* published between 1985 and 1998 followed a negative hypergeometric distribution as demonstrated by Figure 1. The distribution of strokes of the 1,945 kanji with their printed-frequencies is rather rhapsodic, as shown in Figure 2, but a rough-fitting model is drawn in Figure 3. Mathematical modelling for kanji strokes with lexical compositional asymmetry reveals the interesting tendency of *regressive compounding*; that is, that the greater the number of strokes in a kanji, the more it tends to produce two-kanji compound words by adding a kanji on the right side of the target kanji, as shown in Figure 4. A kanji may often have multiple readings; this study also examines the number of readings in relation to the number of kanji strokes. As shown in Figure 6, the greater the number of kanji strokes, the fewer the number of readings. In other words, the more visually complex the kanji is, the more specialised its reading becomes. As such, kanji strokes, as one of the central characteristics of kanji, are closely related to other properties such as frequency, asymmetry and readings. The present study uses mathematical modelling to indicate these relations.



**Abstract:** One of the most fundamental issues for all models of the mental lexicon is how to represent essential information about the morphological structure of polymorphemic words. This paper describes the construction of a large-scale database of two-kanji compound words, which supplements a central component of data relating to 78,426 compound headwords from the Kōjien dictionary with several components focusing on morphological family, morphological structure, and semantic category data. The database will be a particularly valuable resource in terms of supporting and extending research into the lexical retrieval and representation of two-kanji compound words within the Japanese mental lexicon from the perspective of compound word morphology, such as the series of constituent- morpheme

priming experiments (Joyce, 1999, 2002, 2003a, 2003b, 2004; Joyce & Masuda, 2004) that are discussed briefly.



**Abstract:** The present study investigates differences between Japanese prefixes and suffixes using editions of the *Asashi Newspaper* published between 1985 and 1998 (Amano & Kondo, 2000). The *right-hand head rule* (e.g., Kageyama, 1982; Kageyama, 1999; Namiki, 1982; Nishigauchi, 2004; Williams, 1981) predicts that prefixes would be attached to a wide variety of nouns while suffixes would be regularly attached to a smaller group of nouns. Twenty-four frequently-used affixes consisting of 12 prefixes and 12 suffixes were compared according to 7 corpus features, including printed-frequency, productivity, accumulative productivity, commonality, coalescence degree, Herdan's logarithmic function of type-token ratio (log TTR), and entropy. Although a series of Mann-Whitney *U*-tests calculated for the six corpus features of printed-frequency, productivity, accumulative productivity, commonality, coalescence degree and log TTR did not reveal any differences between the 12 prefixes and the 12 suffixes, the *t*-test for entropy indicated a significant difference. This suggests that the prefixes were more randomly or chaotically attached to nouns than the suffixes. Although the present findings are limited only to the selected 24 affixes, the result supported the *right-hand head rule.*



**Abstract:** Various ways are explored in this study of using kanji frequency lists derived from multiple corpora to characterise kanji usage within the corpora. First we discuss the scope of, and issues in processing, four corpora derived from commercially available CD-ROMs: two encyclopedias, a database of newspaper articles, and a four-CD-ROM collection of the texts of mostly fictional paper back books. Next a summary of the kanji frequency data is given, and it is pointed out that the frequency distribution is noticeably different from a classic Zipf's law distribution. A comparison is made between the standard set of Jōyō kanji and high-frequency kanj in the corpora, and the degrees of similarity among the corpora are obtained with the Chi ($\chi^2$) By Degrees of Freedom (CBDF) measure proposed by Kilgarriff (1997). Finally a simple method is tried and evaluated for identifying kanji that have a high frequency in a particular corpus compared to their cross-corpus frequency.



**Abstract.** In the Japanese language, the light verb *–suru* can be attached to various two-kanji compound words containing a *verb-like* feature (or aspects) to allow them to be used as a verb. Using a large sample of the 2,000 two-kanji compound words, encompassing a little less than 80 percent of the total two-kanji compound words printed in 14 years of *Asahi Newspaper* issues, the present study investigates how much the light verb attachment is predicted by four aspects: *inchoative*, *durative*, *telic* and *stative*. A binary logistic regression analysis indicates that all four aspects are significant predictors. Among them, the *telic* aspect shows an overwhelmingly high predictive power. The quantitative theory type III analysis further demonstrates that, in contrast to the *stative* aspect, the *inchoative*, *durative* and *telic* aspects

share a similar semantic feature of *time series*. Nevertheless, since the *telic* aspect overlaps not only the *time series* feature of the *inchoative* and *durative* aspects, but also the *stative* aspect, it is the most effective single predictor for light verb attachment, showing an extremely high prediction percentage of 93.64 with 1.05 percent error.



**Abstract.** For cognitive scientists investigating the nature of lexical knowledge, one essential task is to map out the rich networks of associations that exist between words. This paper reports on a project to construct a large-scale database of word association norms for basic Japanese vocabulary and, utilizing the database, to develop lexical association network maps that tap into important aspects of words and their connectivity. The Japanese word association database will complement existing databases concerning the lexical features of Japanese vocabulary, such as familiarity ratings and frequency counts (Amano & Kondo, 1999; Yokoyama, Sasahara, Nozaki & Long, 1998), and the kanji corpus research highlighted in this special issue. Part 2 of this paper outlines the construction of the database, by detailing initial collections of word association responses from two major questionnaire surveys and the current state of the database. Part 3 introduces the lexical association network maps that will be developed based on the word association norm data and discuses some particularly promising applications of the database and the network maps in the areas of cognitive science and Japanese lexicography and language instruction.