Grapholinguistics in the 21st century (G21C 2020):
From graphemes to knowledge

Online conference; 17-19 June, 2020
https://grafematik2020.sciencesconf.org/

Constructing a database of Japanese compound words: Some observations on the morphological structures of three- and four-kanji compound words

Terry Joyce

Tama University, Japan
terry@tama.ac.jp

Hisashi Masuda

Hiroshima Shudo University, Japan
hmasuda@shudo-u.ac.jp

Opening remarks 1

As the principal component of multi-script Japanese writing system, kanji function as core building blocks in graphematic representation of considerable proportion of Japanese lexicon (Joyce & Masuda, 2018, 2019; Joyce, Masuda, & Ogawa, 2014).

Deeply entwined with the morphographic nature of Japanese kanji (Joyce, 2011), as Kobayashi, Yamashita and Kageyama (2016) observe, there are direct ramifications of this situation.

1. From practical and psychological perspectives,

> kanji play an important role in providing the readers of written Japanese with a visual aid for capturing the meaning of a word at a glance (p. 129)

2. From a morphological perspective, analyses of compound words can elucidate morphographic nature of kanji as linked to both native-Japanese (NJ) + Sino-Japanese (SJ) morphemes.

Opening remarks 2

This presentation reports on the construction of a database of Japanese compound words, with particular focuses on their graphematic representation + their morphological structures.

Prototypically, these are graphematically represented by kanji.

- Majority are SJ (音読み /on-yo.mi/ on-reading) compounds.
- Also NJ (訓読み /kun-yo.mi/ kun-reading) compound words.
- Also some hybrid combinations of SJ and NJ elements.

Consistent with common practice (Kobayashi et al, 2016), our database project is classifying and analyzing Japanese compound words according to overall length and constituents.

Accordingly, the main database components are currently:

- Two-kanji compound words (2KCWs).
- Three-kanji compound words (3KCWs) (Masuda & Joyce 2019).
- Four-kanji compound words (4KCWs) (focus of this presentation).

Opening remarks 3

Main aims of the project are to compile a database of scale to contribute to both:

- A larger database of Japanese lexical properties (Joyce, Hodošček, & Masuda, 2017; Joyce, Masuda, & Ogawa, 2014).
- Stimuli preparation for psycholinguistic surveys and priming experiments (Joyce & Masuda, 2018)
  - In particular, various surveys will be conducted to verify the psychological reality of the morphological analyses applied.

Against a background of growing research interest into how morphological information is represented within the mental lexicon, visual word recognition research, such as constituent priming, with Japanese compound words of various lengths represents a particular promising approach to explore.

Opening remarks 4

Analyses of both the 3KCWs and 4KCWs adopt similar conventions of denoting the constituent kanji:

- As either **A**, **B**, **C**, (3KCWs) + **D** (4KCWs), respectively
- Also using square-brackets, **[ ]**, to indicate internal structures.

The classification analysis is also based on checking for alternative structures within the compound words.

More specifically, all the compound words have been segmented into their consistent kanji, which have then been recombined in different ways, in order confirm the presence of all possible lexical elements.

3KCW analyses (Masuda & Joyce 2019) 1

23,046 most frequent **3KCW lemmas** (token frequencies ≥ 10, excluding proper nouns), extracted from corpus word lists (Joyce, Hodošček & Nishina 2012), compiled from Balanced Corpus of Contemporary Written Japanese (BCCWJ: Maekawa et al, 2013).

3KCW list includes SJ, NJ and hybrid words – this is due to focus on graphematic representation during extraction, but lexical stratum coded.

As Kobayashi et al (2016) note, with SJ morphemes, it is often difficult to discern both morpheme status (free vs. bound) and word-formation process (derivation vs. compounding).

3KCW analyses (Masuda & Joyce 2019) 2: Summary 1

Structure	Type counts	%
[AB]+C	17,761	77.1
A+[BC]	4,904	21.3
[AC*]+[BC] (*C of [AC] omitted)	154	0.7
[AB]+[A*C] (*A of [AC] omitted)	15	0.1
A+B+C	25	0.1
Non-divisible	93	0.4
Monomorphemic (熟字訓)	45	0.2
Phonological transcription (当て字)	64	0.3
Multiple types (Count adjustment)	-15	-0.1
Total	**23,046**	**100**

Dominant [AB]+C pattern (77.1%) and A+[BC] pattern (21.3%) both involve 2KCWs with an additional morpheme appended, underscoring the significance of 2KCWs (Joyce, 2011; Nomura, 1988).

3KCW analyses (Masuda & Joyce 2019) 3: Summary 2

Further analysis results for the [AB]+C structures

Top 4 **C**-additions by type counts

C	Meaning	Frequency
的	adjective ending ‘-ic’	873
者	person ending ‘-er’	685
等	etc.; and so forth	577
性	nature, ‘-ity’ ending	498

Top 4 **[AB]+C** 3KCWs by token counts

3KCW	Gloss	Meaning	Frequency
基本的	/ki-hon-teki/	bas**ic**	182,008
消費者	/shō-hi-sha/	consum**er**	97,209
可能性	/ka-nō-sei/	possibil**ity**	51,613
子供達	/ko-domo-tachi/	child**ren**	38,513

3KCW analyses (Masuda & Joyce 2019) 4: Summary 3

Further analysis results for the A+[BC] structures

Top 4 **A**-additions by type counts

A	Meaning	Frequency
御	honorific prefix	430
大	large, big	313
各	each; every	152
不	negative prefix ‘non-’	143

Top 4 **A+[BC]** 3KCWs by token counts

3KCW	Gloss	Meaning	Frequency
御意見	/go-i-ken/	**your** opinion	54,956
大企業	/dai-ki-kyō/	**large** company	49,820
不可能	/fu-ka-nō/	**im**possible	38,170
一時間	/ichi-ji-kan/	**one** hour	10,752

3KCW analyses (Masuda & Joyce 2019) 5: Summary 4

Notwithstanding certain challenges, given that most kanji are linked to multiple NJ + SJ morphemes, also analysed the additional **A** and **C** components according to their status, as either free, bound or affix morphemes.

Morpheme	[AB]+C				A+[BC]			
status	Types	%	Tokens	%	Types	%	Tokens	%
Free	369	44.0	5,904	33.2	360	55.0	1,882	38.4
Bound	401	47.9	5,016	28.2	225	34.4	491	10.0
Affix	68	8.1	6,841	38.5	70	10.7	2,531	51.6
Total	**838**	**100.0**	**17,761**	**100.0**	**655**	**100.0**	**4,904**	**100.0**

4KCW analyses 1

Adopting the same criteria for extracting the 4KCW lemmas from the same corpus word lists, Stage 1 yielded 298,944 spreadsheet rows.

Stage 2 cleaned the extracted list for classification analysis.

Due to the automatic extraction methods of CWL source corpus, cleaning needed for (1) non-words, (2) proper nouns, and (3) lemma replications → **23,159 4KCW lemmas**

As with 3KCW list, 4KCW list also includes SJ, NJ and hybrid words, due to focus on graphematic representation, and again coding of lexical stratum retained.

4KCW analyses 2: Summary 1: All 4KCW structures

Structure	Type counts	%
[AB]+[CD]	19,805	85.3
[ABC]+D	2,809	12.1
A+[BCD]	449	1.9
Non-divisible	23	0.1
[ACD*]+[BCD] (*CD of [ACD] omitted)	18	0.1
[AD*]+[BD*]+[CD] (*D of [AD] + [BD] omitted)	16	0.1
A+B+C+D	16	0.1
Phonological transcription (当て字)	14	0.1
[AB]+C+D	6	0.0
Monomorphemic (熟字訓)	2	0.0
[AD*]+[BCD] (*D of [AD] omitted)	1	0.0
Total	**23,159**	**100**

Dominant [AB]+[CD] structure, 85.3%, is followed by [ABC]+D pattern (12.1%) and by A+[BCD] (1.9%).

4KCW analyses 3: Summary 2: Dominant [AB]+[CD] pattern

Most frequent [AB] components of [AB]+[CD] structures

Top 4 **AB**-components by type counts

AB	Gloss	Meaning	Frequency
当該	/tō-gai/	respective, appropriate	112
経済	/kei-zai/	economic; finance	88
自己	/ji-ko/	self; oneself	82
生活	/sei-katsu/	living; life	79

Top 4 **[AB]+[CD]** 4KCWs, with the most frequent AB-components, by token counts

4KCW	Gloss	Meaning	Frequency
当該各号	/tō-gai-kaku-gō/	**relevant** article number	214
経済成長	/kei-zai-sei-chō/	**economic** growth	689
自己責任	/ji-ko-seki-nin/	**self**-responsibility	356
生活環境	/sei-katsu-kan-kyō/	one's **living** environment	822

4KCW analyses 4: Summary 3: Dominant [AB]+[CD] pattern

Most frequent [CD] components of [AB]+[CD] structures

Top 4 **CD**-components by type counts

CD	Gloss	Meaning	Frequency
関係	/kan-kei/	relation; connection	164
活動	/katsu-dō/	activity; action	156
以上	/i-jō/	.. and upwards	154
時間	/ji-kan/	time, hour, period	143

Top 4 **[AB]+[CD]** 4KCWs, with the most frequent CD-components, by token counts

4KCW	Gloss	Meaning	Frequency
人間**関係**	/nin-gen-kan-kei/	human **relations**	1,862
経済**活動**	/kei-zai-katsu-dō/	economic **activity**	519
必要**以上**	/hitsu-yō-i-jō/	**more than** necessary	504
労働**時間**	/rō-dō-ji-kan/	working **hours**	790

4KCW analyses 5: Summary 4: [ABC]+D pattern

Second most frequent pattern of [ABC]+D (12.1%)

Top 4 **D**-additions by type counts

D	Meaning	Frequency
等	etc.; and so forth	156
円	yen	152
条	article (in document), provision	116
的	adjective ending ‘-ic’	109

Top 4 **[ABC]+D** 4KCWs by token counts

4KCW	Gloss	Meaning	Frequency
高齢者等	/kō-rei-sha-ra/	**such as** the elderly	99
千五百円	/sen-go-hyaku-en/	1,500 **yen**	691
第十二条	/dai-jū-ni-jō/	**article** 12	636
中長期的	/chū-chō-ki-teki/	mid-to-long term-**ish**	249

4KCW analyses 6: Summary 5: A+[BCD] pattern

Third most frequent pattern of A+[BCD] (at 1.9%)

Top 4 **A**-additions by type counts

A	Meaning	Frequency
約	approximately	84
各	each	46
総	gross, whole, general	23
同	same	22

Top 4 **A+[BCD]** 4KCWs by token counts

4KCW	Gloss	Meaning	Frequency
約一時間	/yaku-ichi-ji-kan/	**approximately** 1 hour	241
各市町村	/kaku-shi-chō-son/	**each** city, town, village	113
総司令部	/sō-shi-rei-bu/	**head**-quarters	134
同委員会	/dō-i-in-kai/	**same** committee	116

4KCW analyses 7: Summary 6: Other structures 1

Non-divisible

炭水化物 /tansuikabutsu/ carbohydrate

[ACD*]+[BCD] (*CD of [ACD*] omitted)

歓送迎会 /kan-sō-gei-kai/ party to welcome (e.g. new employees) and to send off (e.g. retiring employees)

[歓迎会 /kan-gei-kai/ welcome party] + [送迎会 /sō-gei-kai/ sending off party]

[AD*]+[BD*]+[CD] (*D of [AD*] + [BD*] omitted)

陸海空軍 /riku-kai-kū-gun/ land, sea and air forces

[陸軍 /riku-gun/ land forces] + [海軍 /kai-gun/ navy] + [空軍 kū-gun/ air force]

A+B+C+D

春夏秋冬 /shun-ka-shū-tō/ spring, summer, autumn, winter

4KCW analyses 8: Summary 7: Other structures 2

Phonological transcription

滅茶滅茶 /me-cha-me-cha/ disorderly, absurd; excessive

[AB]+C+D

十二箇月 /jū-ni-ka-getsu/ 12 month (period)

Monomorphemic

再従兄弟 /hatoko; haitoko/ second cousin

[AD*]+[BCD] (*D of [AD*] omitted)

産婦人科 /san-fu-jin-ka/ maternity and gynaecology

[産科 /san-ka/ obstetrics] + [婦人科 /fu-jin-ka/ gynaecology]

Closing remarks 1

Results of analyzing the morphological structures of 3KCWs and 4KCWs reveal different dominant principles for the different lengths of compound words.

However, the findings underscore the immense significance of 2KCWs within the Japanese lexicon, not only as words in their own right, but as the basic blocks of longer compound words (Joyce et al 2014).

The next stage of this project will be to conduct various studies to further verify the psychological reality of the morphological analyses applied.

The analysis results will also be incorporated within the larger database of Japanese lexical properties, under gradual ongoing construction.

Closing remarks 2

The conducted analyses into the morphological structures of both 3KCWs and 4KWCs will also be utilized for the preparation of various visual word recognition studies using the constituent-priming paradigm to further investigate the involvement of morphological information within the Japanese mental lexicon.

Consistent with the morphographic nature of kanji (Joyce 2011), the present analysis results also clearly highlight how the concatenation of constituent kanji in graphematically representing the vast majority of Japanese compound words is primarily the province of the morphological processes that underlie the formation of Japanese compound words.

References 1

Joyce, Terry (2011). The significance of the morphographic principle for the classification of writing-systems. *Written Language and Literacy*, 14(1), 58–81. https://doi.org/10.1075/wll.14.1.04joy

Joyce, Terry, Hodošček, Bor, & Masuda, Hisashi. (2017). Constructing an ontology and database of Japanese lexical properties: Handling the orthographic complexity of the Japanese writing system. *Written Language and Literacy*, 20(1), 27–51. https://doi.org/10.1075/wll.20.1.03joy

Joyce, Terry, Hodošček, Bor, & Nishina, Kikuko. (2012). Orthographic representation and variation within the Japanese writing system: Some corpus-based observations. *Written Language and Literacy*, 15(2), 254–278. https://doi.org/10.1075/wll.17.2.01joy

Joyce, Terry, Masuda, Hisashi, & Ogawa, Taeko. (2014). Jōyō kanji as core building blocks of the Japanese writing system: Some observations from database construction. *Written Language and Literacy*, 17(2), 173–194. https://doi.org/10.1075/wll.17.2.01joy

Joyce, Terry, & Masuda, Hisashi. (2018). Introduction to the multi-script Japanese writing system and word processing. In Hye Pae (Ed.), *Writing systems, reading processes, and cross-linguistic influences: Reflections from the Chinese, Japanese and Korean languages* (pp. 179–199). Amsterdam: John Benjamins. https://doi.org/10.1075/bpa.7.09joy

References 2

Kobayashi, Hideki, Yamashita, Kiyo, & Kageyama, Taro. (2016). Sino-Japanese words. In Taro Kageyama & Hideki Kishimoto (Eds.), *Handbook of Japanese lexicon and word formation* (pp. 93-131). Boston, Berlin: Walter de Gruyter.

Maekawa, Kikuo, Yamazaki, Makoto, Ogiso, Toshinobu, Maruyama, Takeiko, Ogura, Hideki, Kashino, Wakako, Koiso, Hanae, Yamaguchi, Masaya, & Den, Yasuharu. (2013). Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation*, 1–27. doi:10.1007/s10579-013–9261-0

Masuda, Hisashi, & Joyce, Terry. (2018). Constituent-priming investigations of the morphological activation of Japanese compound words. In Hye Pae (Ed.), *Writing systems, reading processes, and cross-linguistic influences: Reflections from the Chinese, Japanese and Korean languages* (pp. 221–244). Amsterdam: John Benjamins. https://doi.org/10.1075/bpa.7.11mas

Masuda, Hisashi, & Joyce, Terry. (2019). A database of three-kanji compound words in Japanese, with particular focus on their morphological structures. Poster presentation given as the *'Diversity of writing systems: Embracing multiple perspectives': 12th International Workshop on Written Language and Literacy*, 26-28 March 2019, Faculty of Classics, Cambridge University, UK.

Nomura, Masaaki. (1988). Niji kango no kōzō [The structure of two-kanji Sino-Japanese words]. *Nihongogaku*, 7(5), 44-55.