Poster session 2 (March 27, 2019, AWLL12@University of Cambridge)

A database of three-kanji compound words in Japanese, with particular focus on their morphological structures

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

Terry Joyce
Tama University, Japan
terry@tama.ac.jp

Background

Japanese lexicon and writing system

Lexicon: 4 vocabulary strata (Kageyama & Saito, 2016)
1 和語 /wa-go/ Native-Japanese (NJ),
2 漢語 /kan-go/ Sino-Japanese (SJ),
3 外来語 /gai-rai-go/ Foreign-Japanese (FJ),
4 Mimetics (states + physical sensations).

Writing system: 4 scripts (Joyce & Masuda, 2018)
1 漢字 /kan-ji/ Kanji (Chinese characters),
2 平仮名 /hira-ga-na/ Hiragana (syllabary),
3 片仮名 /kata-ka-na/ Katakana (syllabary),
4 ローマ字 /rōma-ji/ Roman alphabet (phonemic).

Compound morphology

Kanji function as morphographic script (Joyce, 2011), being associated with both NJ and SJ morphemes, which are phonologically referred to as 訓読み /kun-yo.mi/ and 音読み /on-yo.mi/, respectively.

Moreover, the number of morphemes associated with a given kanji varies (Joyce, Masuda, & Ogawa, 2014).

Thus, the morphology of Japanese compounds is especially interesting topic from the perspectives of writing systems and mental lexicon research.

[[[[新+[社屋]]+建設]+案]+[[発表]+会]]
/shin-sha-oku-ken-setsu-an-hap-pyō-kai/
'Gathering to present plan for the construction of a new company building' (Example based on Kobayashi et al., 2016)

Rationale and aims

Two-kanji words are most frequent word structure (Joyce et al., 2014), but many 3-kanji compounds words (3KCW) also exist, with diverse structures.

Primary aim of this research project has been to compile a database (DB) of scale for 3KCWs to contribute to both:

- Larger DB project on Japanese lexical properties (Joyce et al., 2014; Joyce, Hodošček, & Masuda, 2017)
- Stimuli preparation for psycholinguistic surveys and priming experiments (Joyce & Masuda, 2018).

3KCW-DB

Create analysis list

Stage 1: Extracted 3KCWs from **Corpus Word Lists** (CWL) (excluding proper noun lists), which Joyce, Hodošček, & Nishina (2012) extracted from the **Balanced Corpus of Contemporary Written Japanese (BCCWJ)** (Maekawa et al., 2013, Joyce et al., 2017)

→ **171,123** spreadsheet rows.

Stage 2: Reduced and cleaned extracted list
- Reduction criteria: Lemma frequency ≥ 10
- Due to automatic extraction methods of CWL source corpus, cleaning tasks needed for
  (1) non-words,
  (2) proper nouns, and
  (3) lemma replications.

→ **23,046 3KCW-lemmas**

Analysis focus

The list of 3KCWs at the core of this DB includes NJ, SJ and hybrid words, reflecting our criterion of orthographic representation.

Denoting the constituent kanji as **A**, **B**, and **C**, respectively, our analysis classifies 3KCWs according to their word structure (see result panels).

As Kobayashi et al. (2016) observe, however, with SJ morphemes in particular, it is often very unclear with regard to both a morpheme's status (free or bound) and the underlying word-formation process (compounding or affix-derivation).

Results

1. Analysis summary

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**

Understandably, 2,776 (12.0%) involve number kanji with various numerical units or classifiers.

As 3KCWs generally have transparent morphological structures, possible to confidently classify majority as either [AB]+C or A+[BC] structures.

2. Analysis of both A and C additions

For both dominant [AB]+C and A+[BC] structures, also analyzed A and C additional components according to their morpheme status.

However, given that a particular kanji can be associated with multiple morphemes (both multiple NJ and multiple SJ), it should also be noted that any given kanji can potentially be regarded as being free, bound or an affix, depending on the 3KCW.

Morpheme	[AB]+C			A+[BC]		
Status	Types	Tokens	%	Types	Tokens	%
Free	369	5,904	33.2	360	1,882	38.4
Bound	401	5,016	28.2	225	491	10.0
Affix	68	6,841	38.5	70	2,531	51.6
Total		17,761	100.0		4,904	100.0

3. [AB]+C

Top 10 Cs by type

Top 10 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
中	in [place/time]	352
化	verbal ending '-ization'	294
後	after	253
達	pluralizer	244
上	above; in terms of	239
人	person ending '-er'	227

Top 10 Cs by token

Top 10 [AB]+C 3KCWs by token counts

3KCW	Gloss	Meaning	Frequency
基本的	/ki-hon-teki/	basic	182,008
消費者	/shō-hi-sha/	consumer	97,209
可能性	/ka-nō-sei/	possibility	51,613
子供達	/ko-domo-tachi/	children	38,513
誕生日	/tan-jō-bi/	birthday	38,167
外国人	/gai-koku-jin/	foreigner	29,778
二十年	/ni-jū-nen/	twenty years	27,344
十二月	/jū-ni-gatsu/	December	23,480
三十分	/san-jup-pun/	thirty minutes	22,651
世界中	/se-kai-jū/	throughout world	22,050

4. A+[BC]

Top 10 As by type

Top 10 AC-additions by type counts

A	Meaning	Frequency
御	honorific prefix	430
大	large, big	313
各	each; every	152
不	negative prefix 'non-'	143
新	new	127
一	one	126
無	negative prefix 'un-', 'non-'	95
同	same	93
諸	various; several	90
全	all, whole	86

Top 10 As by token

Top 10 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-gyō/	large company	49,820
不可能	/fu-ka-nō/	impossible	38,170
一時間	/ichi-ji-kan/	one hour	10,752
無意識	/mu-i-shiki/	unconsciousness	9,929
二種類	/ni-shu-rui/	two kinds	8,695
小学校	/shō-gak-kō/	primary school	7,488
三箇月	/san-ka-getsu/	three months	7,170
各地域	/kaku-chi-iki/	every region	6,604
新製品	/shin-sei-hin/	new product	6,255

5. Other structures

[AC*]+[BC]

視聴覚 /shi-chō-kaku/ audiovisual [視覚 vision +聴覚 audition]

入出国 /nyū-shutsu-koku/ entering + leaving country [入国 entry to +出国 departure from country]

[AB]+[A*C]

国内外 /koku-nai-gai/ domestic and foreign [国内 domestic +国外 foreign]

十五六 /jū-go-roku/ 15 or 16 [十五 15 +十六 16]

A+B+C

産官学 /san-kan-gaku/ industry, government and academia [産 industry +官 government +学 academia]

衣食住 /i-shoku-jū/ necessities of life [衣 clothing +食 food +住 shelter]

Non-divisible

雰囲気 /fun-i-ki/ mood [atmosphere +surround +atmosphere]

食洗機 /shoku-sen-ki/ dishwasher ← of 食器洗浄機 /shok-ki-sen-jō-ki/ dishwasher [dishes +washing +machine]

Monomorphemic (熟字訓)

波止場 /hatoba/ wharf; quay [wave +stop +place]

五月雨 /samidare/ [5 +month +rain] → early-summer rain

Phonological transcription (当て字)

歌舞伎 /kabuki/ kabuki [song +dancing +art]

目論見 /mokuromi/ plan [eye +argument +see]

Concluding remarks

This presentation of 3KCW-DB project has focused primarily on our analysis of 3KCWs according to their word structures.

However, being extracted from CWLs (Joyce et al, 2012), which were, in turn, extracted from BCCWJ, 3KCW-DB automatically inherits many valuable data-fields, such as word class (mainly ordinary and adjectival nouns), lexical strata, token frequencies of lemma and orthographic base form, pronunciation(s), and similar information for all component 2KCWs.

3KCW-DB will be further refined with the results of future surveys relating to the structural transparency of 3KCWs and utilized in preparing experimental stimuli for visual word recognition research with the priming paradigm (Masuda & Joyce, 2018).

It will also be incorporated as a component of the larger DB of Japanese lexical properties in due course.

References

Joyce, T. (2011). *Written Language & Literacy*, 14, 1, 58–81. doi:10.1075/wll.14.1.04joy

Joyce, T., Hodošček, B., & Masuda, H. (2017). *Written Language & Literacy*, 20(1), 27–51. doi 10.1075/wll.20.1.03joy

Joyce, T., Hodošček, B., & Nishina, K. (2012). *Written Language & Literacy*, 15(2), 254–278. doi:10.1075/wll.15.2.07joy

Joyce, T., & Masuda, H. (2018). In H. K. Pae (Ed.), *Writing systems, reading processes, and cross-linguistic influences: Reflections from the Chinese, Japanese and Korean languages* (pp. 179–199). Amsterdam: John Benjamins.

Joyce, T., Masuda, H., & Ogawa, T. (2014). *Written Language &* Literacy, 17(2), 173–194. doi:10.1075/wll.17.2.01joy

Kageyama, T., & Saito, M. (2016). In T. Kageyama & H. Kishimoto (Eds.), Handbook of Japanese lexicon and word formation (pp. 11-50). Boston; Berlin: Walter de Gruyter.

Kobayashi, H. et al. (2016). In Taro Kageyama & Hideki Kishimoto (Eds.), *Handbook of Japanese lexicon and word formation* (pp. 93-131). Boston; Berlin: Walter de Gruyter.

Maekawa, K. et al. (2013). Language Resources and Evaluation, 1-27. doi:10.1007/s10579-013-9261-0

Masuda, H., & Joyce, T. (2018). In H. K. Pae (Ed.), Writing systems, reading processes, and cross-linguistic influences: Reflections from the Chinese, Japanese and Korean languages (pp. 221–244). Amsterdam: John Benjamins.