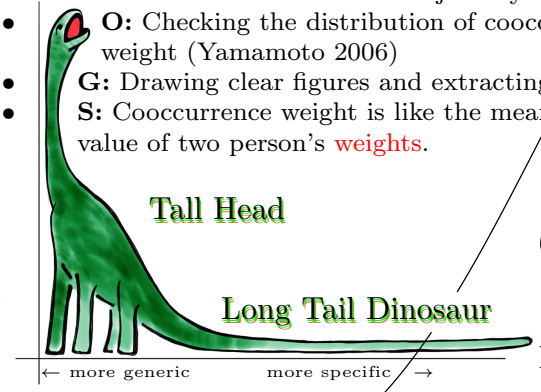
A Study on the Distribution of Cooccurrence Weight Patterns of Classical Japanese Poetry

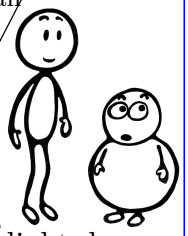
Hilofumi Yamamoto
Tokyo Institute of Technology

Bor Hodošček
Osaka University

Introduction

- **P:** Hairball effect or spoke effect (Yamamoto 2005)
- **P:** Difficult to observe all word adjacency features.
- **O:** Checking the distribution of cooccurrence weight (Yamamoto 2006)
- **G:** Drawing clear figures and extracting function words.
- **S:** Cooccurrence weight is like the mean value of two person's weights.





C. F. Gauss
(1777–1855)

Methods

Material: ***the Hachidaishū (ca. 905–1205)***

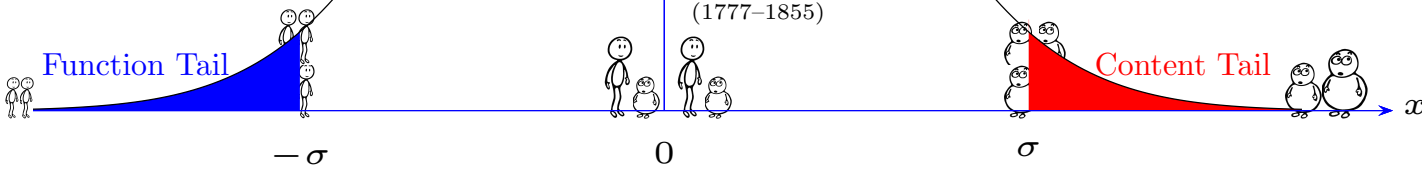
Calculation of Cooccurrence Weight: ***cw***

$$w(t,d) = (1+\log\ tf(t,d)) \cdot idf(t)$$
$$cw(t_1,t_2,d) = (1+\log\ ctf(t_1,t_2,d)) \cdot cidf(t_1,t_2)$$
$$cidf(t_1,t_2) = \sqrt{idf(t_1) \cdot idf(t_2)}$$
$$idf(t) = \log \frac{N}{df(t)}$$

ring!ring!

Distribution of *cw* becomes Bell curve.

- Over $\sigma \implies$ Content Tail.
- Under $-\sigma \Rightarrow$ Function Tail.



Result

Table 1: Upper cutoff patterns of *ame* (sakura): *cw* = co-occurrence weight; *z* = z-value (normalized value of frequency). word annotations: ari(be), ba(cond.), ha(topic.), hana(flower), hito(human), keri(past.), ki(past.), koso(emphatic.), miru(see), mo (also), nasi(no exist), nu(neg.), o(obj.), omou(think), ramu(aux.will), su(do), te(p.), to(and), ware(we), zo(emphatic.), zu(neg.)

	cw	*z*	pattern		*cw*	*z*	pattern		*cw*	*z*	pattern
1	0.62	-0.91	mo–keri	11	0.59	-0.96	nasi–ha	21	0.52	-1.05	nu–o
2	0.62	-0.92	hana–o	12	0.57	-0.98	o–ramu	22	0.52	-1.05	o–zo
3	0.62	-0.92	o–koso	13	0.57	-0.98	mo–ramu	23	0.52	-1.05	miru–o
4	0.60	-0.94	zu–keri	14	0.57	-0.98	ha–ki	24	0.48	-1.09	ba–mo
5	0.60	-0.94	su–ha	15	0.56	-1.00	zu–mo	25	0.48	-1.09	o–keri
6	0.60	-0.94	to–ba	16	0.56	-1.00	o–te	26	0.43	-1.16	zu–ha
7	0.59	-0.96	ari–ha	17	0.55	-1.01	hito–mo	27	0.43	-1.16	to–o
8	0.59	-0.96	ari–mo	18	0.54	-1.02	zu–te	28	0.43	-1.16	te–ha
9	0.59	-0.96	ware–mo	19	0.52	-1.05	zo–ha	29	0.34	-1.27	o–ha
10	0.59	-0.96	nasi–o	20	0.52	-1.05	omou–o	30	0.34	-1.27	o–mo

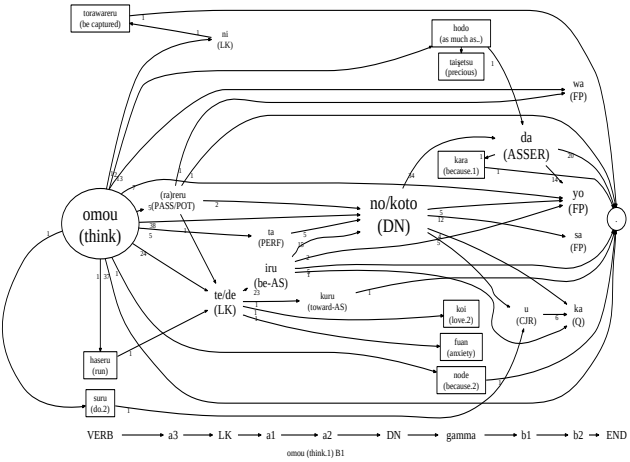


Figure 2: Construction of the predicate of *omou* (think) with Function Tail

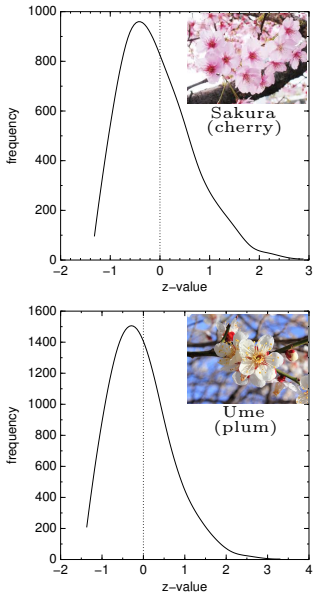


Figure 1: Bell curves

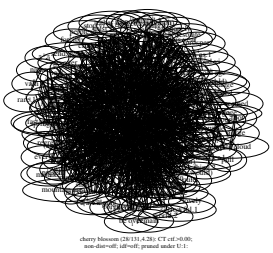


Figure 3: Hairball effect

Table 2: Lower cutoff patterns of *ame* (sakura) in Kokinshū: 30 out of 164 patterns extracted; *cw* = co-occurrence weight; *z* = z-value (normalized value of frequency) word annotations: ba(cond.), bakari(only), besi(should be), chiru(fall), fukakusa(deepgreen), hana(flower), isa(already), kakusu(hide), katu(win), koku(pull), komoru(go deep inside), magiru(mix), makasu(entrust), maku(wind up), manimani(as it is), masi(as), mazu(mix), me(eye), minami(south), miyako(city), mono(thing), nagara(even if), sakura(cherry), si(emphasic.), sumi(black ink), tatu(start,stand), tazumu(being around), tu(past.), uturou(change), watasu(give), yamakaze(mountain wind), yamu(stop), yanagi(willow), yononaka(world)

	cw	*z*	pattern		*cw*	*z*	pattern
1	3.86	3.18	yamu–manimani	106	2.38	1.31	si–fukakusa
2	3.75	3.04	minami–magiru	107	2.38	1.31	sakura–hana
3	3.67	2.93	minami–maku	108	2.38	1.31	sakura–isa
4	3.61	2.86	maku–magiru	109	2.38	1.31	sakura–ba
5	3.42	2.62	yanagi–koku	110	2.38	1.30	sakura–me
6	3.38	2.57	yamu–makasu	—			
7	3.38	2.56	mazu–koku	155	2.17	1.04	chiru–katu
8	3.27	2.43	yanagi–mazu	156	2.17	1.04	bakari–sumi
9	3.26	2.42	sakura–yamu	157	2.16	1.03	maku–besi
10	3.25	2.40	minami–yamakaze	158	2.16	1.03	tatu–maku
–				159	2.16	1.03	tatu–tazumu
101	2.40	1.33	uturou–komoru	160	2.16	1.03	tazumu–tu
102	2.40	1.33	sakura–watasu	161	2.16	1.03	miyako–sakura
103	2.40	1.33	katu–nagara	162	2.16	1.02	kakusu–si
104	2.39	1.32	sakura–masi	163	2.14	1.00	yononaka–sakura
105	2.39	1.31	sakura–makasu	164	2.14	1.00	mono–sakura

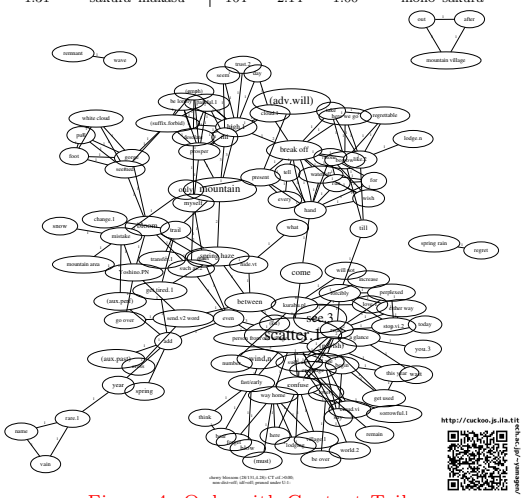


Figure 4: Only with Content Tail

Conclusion

1) the distribution of classical texts fits a Gaussian (Bell) curve as well as in modern texts (Hodošček and Yamamoto 2013);
2) the *cw* value can separate patterns into three layers (low-, mid-, and high-range) using inflection points (-1σ and 1σ);
3) of the three layers, the high-range could be extracted without a list of stop words;
4) the mid-range lexical layer might include mathematical traits not yet revealed in the present study.

Reference

- Yamamoto, H. (2005), Visualisation of the construction of poetic vocabulary using the database of the *Kokinshū*., Jinbun kagaku to dētabēsu (Humanities and Database) the 11th symposium, 81–8, The council of humanities and database.
- Yamamoto, H. (2006), Extraction and Visualisation of the Connotation of Classical Japanese Poetic Vocabulary, Symposium for Computer and Humanities, vol. 2006, 21–28, The information processing society of Japan.
- Hodošček, B. and H. Yamamoto (2013) "Analysis and Application of Midrange Terms of Modern Japanese", in *Computer and Humanities 2013 Symposium Proceedings*, No. 4, pp. 21–26.