

## Distributional evidence for word class: Corpus data

Goal: To use corpus data to explore patterns of distribution for distinguishing word classes

- KOTONOHA (少納言/Syoonagon) corpus — the Balanced Corpus of Contemporary Written Japanese (BCCWJ) [<https://shonagon.ninjal.ac.jp/>] [[how to use](#)]
- Can narrow corpus search by year range or by text genre; all subcorpora included here

About the corpus counts reported here:

- Collected on February 26, 2019
- The Syoonagon corpus searches for exact-string matches only
- “Control cases” and/or percentages are provided below because raw numbers of hits are usually not useful; overall frequencies vary widely for different lexical items

(1) If a word is immediately followed by case markers *-ga/-o/-ni*, is it guaranteed to be a N?

	raw hits	% of total
行ったが itta ga 'go-PAST' + ??	348	93.0%
行ったを itta o	0	0%
行ったに itta ni	26	7.0%
<i>total</i>	<i>374</i>	

	raw hits	% of total
自転車が zitensya ga 'bicycle NOM'	279	19.5%
自転車を zitensya o	591	41.3%
自転車に zitensya ni	560	39.2%
<i>total</i>	<i>1430</i>	

- *-ga*: nominative marker for N, *or* 'but, although' (which can follow V and A)
- *-ni*: multiple uses, not all of which require it to follow N
- *-o*: accusative marker for N

(2) If a word is preceded by a demonstrative such as *sono* 'that', is it guaranteed to be a N?

	raw hits	% of total
高いです takai desu 'high COPULA'	649	95.2%
その高い sono takai	33	4.8%
<i>total</i>	<i>682</i>	

	raw hits	% of total
自転車です zitensya desu 'bicycle COPULA'	18	54.5%
その自転車 sono zitensya	15	45.5%
<i>total</i>	<i>33</i>	

- An adjective or other modifier can intervene between a demonstrative and its N

(3) X *no* N: What word class is X?

- "の報告": Examples from search results (報告 *hookoku* 'report')

故障の報告	故障	<i>kosyoo</i>	'breakdown, failure'
松村さんの報告	松村さん	<i>Matumura-san</i>	personal.name-HONORIFIC
他の報告	他	<i>hoka</i>	'other'
協会などの報告	協会	<i>kyookai</i>	'association'
	など	<i>nado</i>	'etc.; things like that'
次長からの報告	次長	<i>zityoo</i>	'vice-chief; assistant manager'
	から	<i>kara</i>	'from'
大坂への報告	大坂	<i>Oosaka</i>	'Osaka' (place name)
	へ	<i>e</i>	'to, toward'

(4) '*i*-adjectives' and '*na*-adjectives' with forms of the copula: Are the patterns different?

(oops! this is old data from Google — needs updating)

	raw G-hits	% of total
高いです takai desu 'high COP-FML'	45,800,000	54.1%
高いだ takai da ...COP-INFML'	1,030,000	1.2%
高いでした takai desita ...COP-PAST-FML'	3,680,000	4.3%
高いだった takai datta ...COP-PAST'	70,600	0.1%
高かったです takakatta desu 'high-PAST COP-FML'	34,000,000	40.2%
高かっただ takakatta da ...COP-INFML'	82,500	0.1%
<i>total</i>	<i>84,663,100</i>	

	raw G-hits	% of total
静かです sizuka desu 'quiet COP-FML'	2,320,000	7.7%
静かだ sizuka da	2,590,000	8.6%
静かでした sizuka desita	21,600,000	72.0%
静かだった sizuka datta	3,500,000	11.7%
—		
—		
<i>total</i>	<i>30,010,000</i>	

Key to color codes:

very rare	0– 4.9%
rare	5– 14.9%
frequent	15– 64.9%
very frequent	65– 100%