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) [http://www.kotonoha.gr.jp/shonagon/]
- 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%
total	*374*	

	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%
total	*1430*	

- *-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%
total	*682*	

	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%
total	*33*	

- 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	*kosyoo*	'breakdown, failure'
松村さんの報告	松村さん	*Matumura-san*	personal.name-HONORIFIC
他の報告	他	*hoka*	'other'
協会などの報告	協会	*kyookai*	'association'
	など	*nado*	'etc.; things like that'
次長からの報告	次長	*zityoo*	'vice-chief; assistant manager'
	から	*kara*	'from'
大坂への報告	大阪	*Oosaka*	'Osaka' (place name)
	へ	*e*	'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%
total	*84,663,100*	

	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%
—		
—		
total	*30,010,000*	

Key to color codes:

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