Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Alvin C. Grissom II Date

SENTIMENT IN JAPANESE: A CORPUS-BASED APPROACH WITH SOCIOLINGUISTIC AND CROSS-LINGUAL IMPLICATIONS

By

Alvin Castillo Grissom II
Master of Science
Computer Science

Eugene Agichtein, Ph.D. _____

Advisor

James J. Lu, Ph.D. _____

Committee Member

Phillip Wolff, Ph.D. _____

Committee Member

Eugene Agichtein, Ph.D. _____

Committee Member

Accepted:

_____

Lisa A. Tadesco, Ph.D

Dean of the Graduate School

_____

Date

SENTIMENT IN JAPANESE: A CORPUS-BASED APPROACH WITH SOCIO-LINGUISTIC AND CROSS-LINGUAL IMPLICATIONS

By

Alvin Castillo Grissom II

B.A., Hendrix College, 2006

Advisor: Eugene Agichtein, Ph.D.

An abstract of

A thesis submitted to the Faculty of the Graduate School of Emory University in partial fulfillment of the requirements for the degree of

Master of Science

in Computer Science

2009

Abstract

SENTIMENT IN JAPANESE: A CORPUS-BASED APPROACH WITH SOCIO-LINGUISTIC AND CROSS-LINGUAL IMPLICATIONS

By Alvin Castillo Grissom II

Great progress has been made on sentiment analysis techniques for in the English language; however, for other languages, sentiment analysis is less well understood. This thesis reports on statistical analysis of sentiment in Japanese and English text. Salient features for each are analyzed to better understand how authors convey sentiment. In particular, socio-psychological and linguistic explanations are given for their usage. In addition to proposing linguistic insights and hypothesis, the foundation is laid for more effective automatic sentiment classification for non-English languages.

SENTIMENT IN JAPANESE: A CORPUS-BASED APPROACH WITH SOCIO-LINGUISTIC AND CROSS-LINGUAL IMPLICATIONS

By

Alvin Castillo Grissom II

B.A., Hendrix College, 2006

Advisor: Eugene Agichtein, Ph.D.

A thesis submitted to the Faculty of the Graduate School of Emory University in partial fulfillment of the requirements for the degree of

Master of Science

in Computer Science

2009

TABLE OF CONTENTS

TABLE OF FIGURES

1 INTRODUCTION

In general, when one writes or speaks, he or she may express some opinion or encodes some emotive content into what is being said. This is the *sentiment*. Sentiment analysis, in the context of computer science, has been extensively studied in the English language domain by using traditional classification methods. The sentiment analysis research for other languages, however, is much more sparse. Pragmatically, as English is the most widely used language among both the researchers and the users of the Internet, it is not difficult to imagine why this is the case. However, at least one survey[1] has suggested that Japanese may actually be the most popular language among blogs, being roughly tied with English with 37% and 36%, respectively, quite a feat for a language used in a country roughly half the size of the United States. Ahmed et. al. note that, save for work done by Katayama et. al. on Japanese text classification, there has been little work on sentiment analysis in the non-English domain, and even this study necessitated first translating the material to English.[2-3]

The research literature is, by and large, devoid of computational sentiment analysis in other languages, and especially those which use features specific to the language.

In general, East Asian languages, such as Chinese, Japanese, and Korean operate on a different paradigm of expression from that of western languages; there is no reason to assume, therefore, that identical techniques as those used for English, or even western languages in general, will be appropriate for these languages. One possible approach, such as the one attempted by Katayama et. al., involves automatically converting the source language to English, for which we have models which yield reasonable performance.

There would seem to be two ways of approaching the problem, and a bit of middle ground: on one hand, a strictly utilitarian approach to natural language processing necessitates only that the techniques which yield the best numerical results by pursued; at the other extreme, a purely theoretical approach concerns itself not with performance, necessarily, but with the insight gained as a result of the research. Here, I explore the insights which may be gleaned from reasonably large, labeled corpora in Japanese. What are some of the important linguistic patterns in Japanese text, as they pertain

to sentiment analysis? How do these patterns differ from their English counterparts?

Indeed, Japanese is structurally a very different language from English, but also important are the different ways in which Japanese is used in practice. I contend, first, that there *are* important differences in the ways that Japanese-speaking and English-speaking people express sentiment; consequently, to gain insight into the ways that Japanese speakers express themselves in practice, it is necessary to use the source language, not a translation.

With the large amount of product reviews available from sites such as Amazon.com (our corpus), we have a large amount of labeled data. A one-star reviewer is woefully displeased with his product and perhaps displeased in general, a five-star reviewer is the exact opposite, and a three-star reviewer, as we shall see, is likely attempting to balance the good and the bad. By exploring the *nature* the langue of these reviews, we gain insight into what a five-star review *means,* what a three-star review *means*. What does the language betray? Furthermore, is this consistent across cultures? That is not at all obvious. If, then, we are to classify sentiment in any way other than strictly numerically or very broadly: if we are to

go beyond predicting whether a reviewer is “expressing a positive sentiment,” we must delve deeper than predicting the labels and actually interpret the data linguistically.

Approach

To analyze cues of sentiment in respective classes of sentiment, I have used Amazon.com reviews. Each review has a rating, from one to five stars, which we shall use as our sentiment, with a one-star review being strongly negative sentiment and a five-star review representing strongly positive sentiment. The corpus consists of lists of tokens for each review class, described below. When pertinent to the discussion, examples from actual reviews will be used. My aim is to measure the relevance of certain tokens to particular rating classes and show that there are some specific ways in which individuals communicate in certain contexts.

Methodology

I have made use of the English and Japanese sections of UMass Amherst Linguistics Sentiment Corpora[1], used by [4-5]. This corpus consists only of lists of bigrams and unigrams for each language, collected from Amazon reviews. Each *n*-gram has a count for every review score. In the case of English, a unigram is, as usual, a single word, with all punctuation aside from exclamation points and question marks removed; for Japanese, a unigram is slightly more complicated: Since Japanese words are not segmented by spaces, preprocessing must be done in order to extract the individual words from the text. Japanese contains linguistic elements, such as particles, which are not considered to be individual words. The tokenization was done with MeCab[2], a morphological analyzer for Japanese text. It will often separate constituent parts of the words, separating the stem from added morphology. As a result, the bigrams are often complete "words," while the unigrams often contain stems, inflections, or other grammatical transformations which, by themselves, would never be used. As we shall see, however, they are nevertheless useful units of analysis. Moreover, often these tokens appear to be the only way to capture in isolation

---

[1] http://semanticsarchive.net/Archive/jQ0ZGZiM/readme.html Accessed October , 15, 2009

[2] http://mecab.sourceforge.net/

certain content: tenses, conditional statements, desire, and politeness, to name of a few.

The Japanese corpus *n*-gram list consists of book, movie, electronics, and music reviews from 12,747 authors. The following are the statistics from the source material:

JP	1 Star	2 Star	3 Star	4 Star	5 Star	Total
reviews	971	759	1609	3504	11031	17874
tokens	127049	123312	277857	636067	1805764	2970049
vocab	9574	9909	16247	24902	39948	2970049

Table 1: Japanese Token Statistics

Likewise, for the English Amazon material, we have the following statistics, taken from only book reviews by 40,625 authors:

EN	1 Star	2 Star	3 Star	4 Star	5 Star	Total
reviews	3323	2687	3994	8601	34952	53557
words	570687	512643	767958	1513776	4769921	8134985
vocab	27352	26239	32818	46036	80569	112323

Table 2: English Token Statistics

Since the data are noisy, I introduce an empirically-determined cutoff for feature reduction. For a term to be deemed significant, unless otherwise stated, it must occur in .06% of the rating for which it is being measured.

Unless otherwise stated, we use a weighted version of the log-odds metric, one which takes into account the wide variances in the number of tokens for each class. We compare the probabilities that the terms will occur in an instance of each respective class. For each calculation, we assume two classes: $R_i$and $\overline{R_i}$, where $i \in \{1,2,3,4,5\}$, corresponding to a specific rating class, and and $\overline{R_i}$ is union of all ratings except R_i. For each class, we calculate $p(i,n) = \frac{Count\ (n \in R_i)}{|R_i|}$ as the probability that a random token taken from an instance of class R_i will be n: it is the percentage of the given rating class that n comprises. (Simply calculating the probability that a token n occurs in a given class would skew the results toward those with more sample data; thus, we calculate individual percentages for each class in order to normalize). For the class $\overline{R_i}$, we average the individual percentages $p(j,n)$ for all $j : j \neq i$. Our modified odds will be the ratio of these two probabilities, and our version of log odds will be the natural log of this value.

$$ln\frac{p(i,n)}{Avg[p(j,n)]:j \neq i}$$

This is to estimate the following:

$$L \approx ln\frac{P(n\ occurrs\ in\ an\ instance\ from\ R_i)}{P(n\ occurs\ in\ an\ instance\ from\ \overline{R_i})}$$

While I have attempted to compensate for statistical inequity by averaging the percentages of the classes, there is still a potential bias due to the overwhelming inequity in tokens. While calculating a version of odds, instead of simply raw counts, ameliorates this somewhat, it is worth noting that there is a potential bias.

CONTRIBUTION

I address the following questions: What are cues of sentiment in Japanese, how do they function, and how does this usage compare to English usage?

I shall demonstrate, first, that, though the sentiment-carrying tokens in Japanese do not, in general, correlate with similar terms in English, there are some classes of tokens which show similar patterns in both languages. Often, however, some of the most statistically informative tokens in Japanese have no English

equivalent at all. Mapping the top terms from English to Japanese is much easier (or at least feasible) in most cases, whereas the important tokens in Japanese often have no correlates at all in English

In Japanese, I show that there are patterns of politeness, word usage, emphasis, honorifics, conjunctions, and psychological distance which correlate with sentiment in the data. I offer a qualitative analysis of these patterns and of some selected examples which exhibit them, and I offer linguistic, psychological, and practical explanations for many of the patterns. I then compare some of the overall trends with those that can be gleaned from English *n*-grams, as well.

In the final section, I briefly demonstrate that, as an alternative to the work pursued by Kanayama et. al., mentioned earlier, it is possible to achieve reasonable performance with standard statistical machine learning techniques by segmenting the text into tokens, without resorting to deep analysis and English translations.

2 RELATED WORK

Sentiment analysis in English has been studied extensively, often as a classification task in which sentiments are divided into positive and negative categories.[6-7] Mihalcea and Banea have explored methods of detecting subjectivity by in multilingual contexts using English as a bridge language.[8] In general, this is the approach that has been taken in the realm of multilingual sentiment analysis, due to the lexicons and methods that are known to work to some degree for English text. Bautin et. al. also used translation to perform multilingual sentiment analysis on news and blogs.[9]

As mentioned, there has been relatively little work published in the realm of Japanese sentiment analysis. Kanayama and Nasukawa successfully employed a novel machine translation framework for sentiment analysis, translating text from Japanese to English fragments by using deep syntactic analysis.[10] Kanayama et. al. also successfully used domain-specific *polar clauses* which convey positive or negative meaning in a specific domain.[3] They implemented an unsupervised method for the detection of these polar clauses. In so doing, they were able to achieve high precision on their data set or forum posts. Their approach purposefully included domain-specific information, and they noted that this has the benefit of automatically determining the positive and negative

attributes of particular products. However, since their method was designed to acquire domain-dependent properties, it is robust for various domains. Kanyama and Nasukawa furthermore make the observation that context polarity is easier to determine in Japanese, noting that indirect negation is very rare in Japanese. We shall see that direct negation is, indeed, relevant to sentiment analysis.

Insofar as linguistic analysis of the sentimental cues in multilingual corpora is concerned, Constant et. al. computed and demonstrated patterns in the log-odds ratio among the five Amazon ratings for some selected terms in English, Chinese, German, and Japanese. In particular, they tracked the occurrence of *expressives,*[11] a few words and phrases which, in their words, "pack a punch." As noted by Tsujimura, the *totemo* expressive "…modifies the majority of adjectives in Japanese."[12] Constant's study demonstrated some regularity in how these terms are used in their respective languages; however, due to the number of languages being studied, each language received shallow treatment. Analysis of term patterns in Japanese, for example – the primary focus of this research – was limited to a single term, *shimau*[3], and its

---

[3] しまう, shimau, is a Japanese antihonorific, attached to a verb and adding a negative connotation to the expression; it often connotes that something has been done unintentionally. . Potts and Kawahara draw a parallel to the English supplementary relative,

conjugations. Nakanishi theoretically examines the Japanese equivalents of “even” and “only” and their correspondence to negative linguistic polarity, which, as I shall show later, is correlated with negative sentimental polarity. [13]

Many of the units of analysis in this thesis will be particles, many of which are difficult to define. Among these are the “explanatory” particles *n* and *no*.

Takatsu[14] summarizes the problem as follows:

> The NO DA construction is one of the most common expressions in Japanese, and yet its precise function is rather difficult to define. A considerable amount of research has been undertaken on this construction, revealing in what environment NO DA tends to appear and what kind of inferences it gives in each case. It has been suggested, for example, that the function of the NO DA construction is, in some cases, "explanatory" (Alfonso 1966, Kuno 1973), in others "emphatic" (Alfonso 1966, Okatsu 1974, Martin 1975, Mizutani 1977, McGloin 1980), and in yet other instances serves

---

“which sucks.”[4] Another usage of しまう is to express that something has been done fully or completely.

> to "present new information as if it were already known" (McGloin 1984). However, while many linguists have demonstrated considerable insight in their discussion of this expression, and have contributed some extremely valuable comments, none of them has accounted for all the uses of the NO DA construction, or fully explained
> its apparently different roles in different environments.

We find, as well, that pronoun usage may be relevant to sentiment, such as in the case of the Japanese reflexive pronoun, *jibun.*

Kunishige[15] summarizes some of the previous study of *jibun* as follows:

> Kuno, and Kuno and Kaburaki characterize *jibun* as empathy expressions. They argue, for example, that when the reflexive *jibun* is used to refer to a participant in an event in a complex sentence…with its antecedent not in the same simplex sentence that it is in, the speaker empathizes and identifies himself with the participant…

3 POLAR TOKENS IN JAPANESE REVIEWS

That certain telltale terms are much more likely to be used in positive or negative contexts in unsurprising. This has been studied extensively in English, usually with “bags of words” machine learning approaches[6-7]. While this approach is not unreasonable for Japanese, unlike English, Japanese language consists of more than merely “words.” Politeness levels, honorifics, and particles, in many cases, have no English counterparts. Often, as in the case of particles, linguists do not even agree on the meaning of these elements of Japanese. I believe that it will become clear that, in Japanese, ignoring these tokens would be misguided. Though particles, for example, might be analogized, in many instances, to stop words in English, it is clear that there is significant relevance to their usage in terms of sentiment.

In the following section, I present the *L*-scores for the ratings 1 and 5 for the terms which exhibit interesting behavior and for which we find a reasonable qualitative explanation for this behavior. We are choosing the “polar” classes of one-star and five-star reviews for our initial analysis We will first consider the unigrams, which are

often not complete words, and bigrams, which are more often complete words and sometimes are two words or a combination of a word and a particle. It is worth noting that, due to duplicate character encodings for some characters, there are duplicates in the original lists. I do not consider these duplications to add anything qualitatively to the data and will generally omit them; I do, however, leave alternate “spellings” intact. Likewise, there are occasionally entries which do not appear to be valid character encodings. These are also removed. The versions in the appendices are unaltered for reference.

Table 3: Top Unigram Tokens
for 5-star Reviews

Rank	Unigram	1-Star L	5-Star L	Meaning	Translit.
1	！	0.025262	0.839447		
2	心	-0.38889	0.737261	heart/mind/feeling	kokoro
3	そして	-0.32712	0.716928	And	soshite
4	とても	-0.63121	0.605945	very	totemo
5	くれ	-0.04271	0.55008	[indicates strong command]	kure
6	度	-0.07918	0.507049	"time," as in "This time"	tabi/do
7	聴い	-0.01024	0.477925	listen/hear	
8	読み	-0.90352	0.443507	read	yomi
9	とき	-0.35258	0.438497	when/time	toki
10	お	-0.30217	0.426534	[honorific prefix]	o
11	読ん	-0.6725	0.409481	conjugation of "read"	yon
12	本	-0.23664	0.388759	book	hon
13	中	-0.18815	0.375586	middle, inside	naka
14	本当に	0.367065	0.374454	really, truly	hontou ni
15	今	0.203959	0.343071	now	ima
16	,	-1.1733	0.33349	[comma]	
17	ながら	-0.39576	0.3276	while (temporal or "although")	nagara
18	時	0.030414	0.318536	when, time	toki
19	自分	-0.3132	0.314465	one's self	jibun
20	いく	-0.70187	0.308293	to go, to surely plan to do some	iku
21	この	-0.0131	0.271619	This (near to the speaker)	kono
22	できる	-0.68978	0.263004	[indicates ability]	dekiru
23	い	-0.17836	0.259855	*unclear*	i
24	私	-0.14831	0.257847	I	watashi
25	年	0.315923	0.242232	year, years	nen
26	まし	-0.00861	0.238907	past tense token	nen
27	また	-0.19602	0.226394	again	mashi
28	み	-0.27009	0.225841		mi
29	一	0.195057	0.223721	one	ichi
30	なり	-0.10292	0.218504	polite token for become	nari

Table 4: Japanese 1-Star

Bigram Tokens Ranked by L

Rank	Bigram	1-star L	5-star L	Meaning	Translit.
1	！！	0.127095	1.08168		
2	この本	-0.31693	0.651441	this book	kono hon
3	本を	-0.42275	0.564636	[indicates "book" is direct object]]	hon o
4	てくれ	0.004122	0.551501	[indicates strong command]	te kure
5	の中	-0.32337	0.543773	"inside of [something]"	no naka
6	います	-0.42225	0.524992		imasu
7	読んで	-0.74612	0.507646	gerund, comd., or conj. form of "read"	yonde
8	ことが	-0.58593	0.439968	[a fact or abstract notion is the subject]	koto ga
9	いまし	0.103168	0.417583	past tense of (6)	imashi
10	になり	-0.26416	0.382356	becomes	ni naru
11	にも	-0.13185	0.346303	[combination particle]	ni mo
12	自分の	-0.2712	0.342082	one's	jibun no
13	。この	-0.04971	0.325743	. This	. Kono
14	見て	-0.11497	0.305384	gerund, comnd., or conj. form of "read"	mite
15	てい	-0.28242	0.281972	gerund, comnd., or conj. form of "look"	tei
16	いて	-0.63539	0.279866	gerund, comnd., or conj. form of "go"	ite
17	思って	-0.19145	0.267131	likely the gerund form of "think"	omotte
18	私は	-0.13589	0.2602	[indicates "I" (oneself) is the topic]	watashi wa
19	ます。	-0.31585	0.257522	[present/future tense (polite)	masu
20	ました	-0.01168	0.245771	[polite verb form of past tense]	mashita
21	てみ	-0.24328	0.236942	[likely indicates "trying" something]	temi
22	。	-0.59159	0.230849	[period]	.
23	。「	0.157987	0.227477	[period followed by a beginnign quotation	. "
24	なった	-0.08159	0.214837	became	natta
25	いる。	-0.4325	0.208059	[likely some living thing exists]	iru.
26	てき	0.05534	0.20468	[-al ending, as in "logical" educational," et	teki
27	」の	-0.39964	0.200609	[posessive form of a direct quotation]	
28	ことを	-0.01141	0.192587	[nominalized verb phrase is direct object (	" no
29	と思っ	-0.08257	0.186303	"I'm thinking/thought [something]"	to omo-
30	いた	-0.08918	-0.03504	probably past progressive tense	ita

Table 5: Japanese Tokens Ranked by *L* for 1-Star Reviews

Rank	Unigram	1-star L	5-star L	Meaning	Translit.
1	章	1.681706	-1.34053	chapter	shou
2	ベスト	1.494947	-1.09731	best	besuto
3	歌い	1.435387	-1.08762	sing	utai
4	売れ	1.398573	-1.14461	can buy	ure
5	出す	1.347816	-1.1695	put out	dasu
6	?	1.327914	-1.07569		
7	型	1.020542	-0.75065	model/type	gate
8	第	1.011832	-0.74193	[counter]	
9	ファン	0.984621	-0.61906	fan	fan
10	なんて	0.900958	-0.11952		nante
11	枚	0.878294	-0.35214	[likely counter for CDs]	mai
12	もう	0.856494	-0.26541	already/again/more/enough	mou
13	こんな	0.799946	-0.52419	This kind of…	konna
14	てる	0.7831	-0.54408	informal gerund	teru
15	じゃ	0.777999	-0.5531		ja
16	…	0.771053	-0.70402		
17	よ	0.7672	-0.18748		yo
18	出し	0.751304	-0.57087		dashi
19	のに	0.735456	-0.73807	despite	no ni
20	ばかり	0.672103	-0.17886	only	bakari
21	!	0.667302	0.4838		
22	買う	0.646588	-0.54691	to buy	kau
23	アルバム	0.641466	-0.08358	album	albamu
24	やっ	0.610232	-0.20992	did	ya[tta/tte]
25	CD	0.604939	-0.14809		
26	歌	0.550224	0.098848	song/sing	utai
27	ん	0.540745	-0.45133		n
28	って	0.527623	-0.421	likely quotation/command	tte
29	しか	0.510983	-0.4886	only	shika
30	言っ	0.495647	-0.2587	conjugation of "to say"	i[tta/tte]

	Bigram	1 -star L	5-star L	Meaning	Translit.
1	章の	1.724394	-1.44323	chapter's	shou no
2	か?	1.596696	-1.19649	[question]	ka?
3	ですか	1.565261	-1.21288	[formal question]	desu ka
4	を	0.935427	-0.40495	piece of music (d.o.)	kyoku o
5	よ。	0.882879	-0.25163		yo.
6	…。	0.877747	-1.1741		…
7	んだ	0.824407	-0.3695	exp. copula	n da
8	なん	0.691585	-0.39817		nan
9	んです	0.657586	-0.25371	exp. copula (formal)	n desu
10	のでしょ	0.652267	-0.71858	probably	no desho[u
11	の曲	0.650631	-0.21895	X's chapter	no kyoku
12	じゃない	0.64607	-0.50818	is not	ja nai
13	だから	0.58148	-0.42431	Therefore	dakara
14	言って	0.538955	-0.28532	say[ing]	itte
15	んでし	0.513857	-0.5294	explanatory version of "was"	n deshi[ta]
16	ないです	0.49199	-0.65842	is not (formal)	nai desu
17	か？	0.447833	-0.63982	[question]	ka?
18	ない。	0.427833	-0.73486	is not	nai.
19	ん。	0.425502	-0.53334		n
20	ません	0.405419	-0.47228	is not (formal)	masen
21	うか	0.400115	-0.47436		u ka
22	でしょう	0.384815	-0.36987	probably	deshou
23	うと	0.351392	-0.10951		u to
24	方が	0.339734	-0.70854	[indicates a comparison]	hou ga
25	ですね	0.320687	0.066777		desu ne
26	ね。	0.314235	-0.13917		ne.
27	をし	0.30762	-0.06509	[do X]	o shi
28	から、	0.25907	-0.10217	from/since	kara
29	購入し	0.25582	-0.02157	purchase	kyounyuu s
30	のか	0.240351	-0.46219	[question]	no ka

Table 6: 1 Star Top Bigram Tokens

3.1 EXPLICIT MODIFIERS

Of the unigram tokens ranked for five-star reviews, two -- *totemo*[4], ranked 4, and *hontou ni*[5] , ranked 14 -- are explicit intensifiers. The loan word word *besuto*[6] is an explicit modifier, though it is not clear why it is ranked so highly. That *hontou ni*, a relatively straightforward intensifier, is bipolar might lead one to believe that its usage carries with it no specific emotive polarity, only that it intensifies whatever emotive content is indicated in its context. Further inspection of its behavior, however, indicates that its usage *is* more likely to occur in a negative context than a positive.

Functionally, *totemo* is generally used in the same way as "very" in English, modifying verbs and adjectives, but primarily adjectives. Tsujimura further notes Bolinger's[16] claim that, in English, there are at least two types of intensifiers for verbs: those which modify the intensity of the event to which the verb refers, and those which emphasize the amount, rather than the degree. The adverb *totemo*, Tsujimura shows, functions in ways analogous to this usage in

---

[4] とても, totemo: very

[5] 本当に, hontou ni : truly, really

[6] ベスト, besuto: best. This is an English loan word.

English[7]. How *totemo* functions depends upon whether or not the verb is a degree verb. Tsujimura notes several classes of verbs which are prone to modification by *totemo*: psych verbs (pleased, suffer, surprised), verbs of emission (stinks, shone), and two subclasses of change-of-state verbs (widened, warmed, shrunk). He further states that, "…a large majority of the verbs outside of [these classes] resist totemo modification." My experiments suggest that *totemo* usage is very much skewed toward positive polarity.

Consider the following excerpts from five-star reviews:

---

[7] Tusjimura provides several examples, including the following:

太郎は金を**とても**借りた。*Tarou wa kane o totemo karita.* "Tarou borrowed **a lot of** money." In this example, totemo does not strictly translate to "very." Intuitively, it might be considered analogous to the sentence, "Tarou *very much* borrowed money." Such statements, often used ironically, intensify the action being done: in this case, borrowing. But how does one "borrow money 'verily?' In this case, it is by borrowing a significant amount of money. However, it could also conceivably be used in the purely intensive case. For example, "I very much stole his car." In this case, the intensity does not imply the amount: either he stole his car or he did not. This usage is often indicative of emphasis. "Did you steal his car?" "Yeah, I very much stole his car." Thus, it would seem, in English, whether the intensification of a verb implies an extension of degree depends upon the context and the verb being modified. The same is true in Japanese.

太郎は古本を**とても**売った。*Tarou wa furuhon o utta.* "Tarou sold a lot of used books." Here, again, *totemo* modifies the degree.

太郎はとても苦し nda。*Tarou wa totemo kurushi n da.* "Tarou suffered very much." Here, *totemo* modifies the degree. Of course, in English, this could also be phrased, "Tarou suffered a lot," which may or may not have the same meaning.

(1) 僕はとても面白いゲームだと思います

boku wa totemo omoshiroi da to omoimasu

"I think that this is a very interesting game."

Here, *totemo* modifies the i-adjective[8] 面白い[9], making it "very interesting."

(2) **とても**分かりやすい本でした。

totemo *wakariyasui hon deshita*

This was a very easy-to-read book.

The following is a one-star review:

(3) **とても**児童向けの作品ではない。

totemo *jidou muke no sakuhin de wa nai.*

This product is not appropriate for children at all.

The final example illustrates the *totemo…nai* construction, which can mean "not at all." Despite this construction, the fact that

---

[8] There are two kinds of adjectives in Japanese: i-adjectives and na-adjectives. I-adjectives end in *i,* whereas na-adjectives end with *na*.

[9] 面白い omoshiroi. This word has the general meaning of "interesting," but contextually can mean "enjoyable," "amusing," "funny," and other variations.

totemo skews so much toward positive polarity indicates that this construction is not particularly popular.

3.2 Scope Extension vs. Scope Restriction in Polar Cases

While sentences (1) and (2) are relatively straightforward modifications, sentence (3) is actually describing what is *not* done. In fact, the one-star tokens might yield clues regarding why *totemo* behaves in this way.

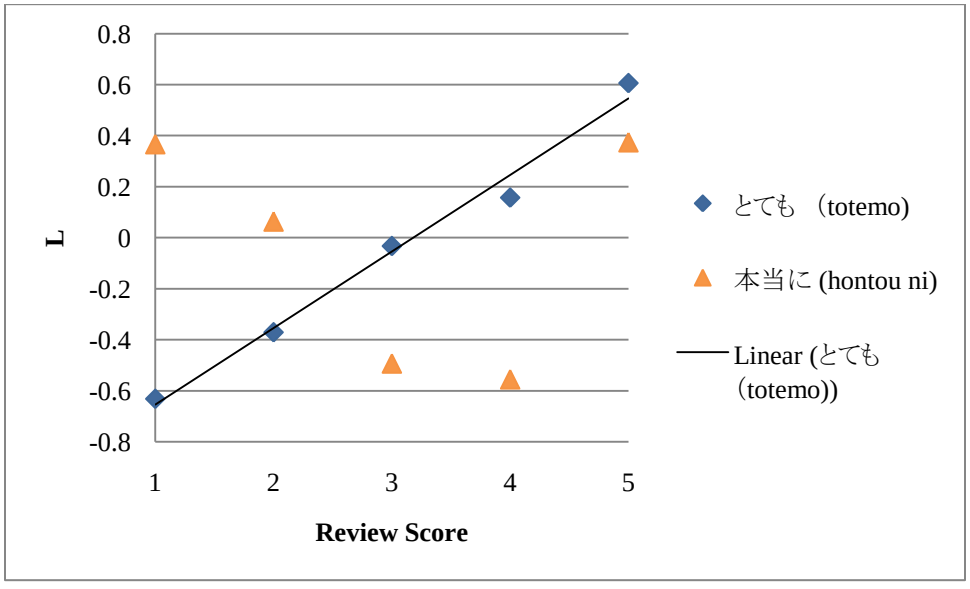


Figure 1: totemo ("very"), hontou ni ("really")

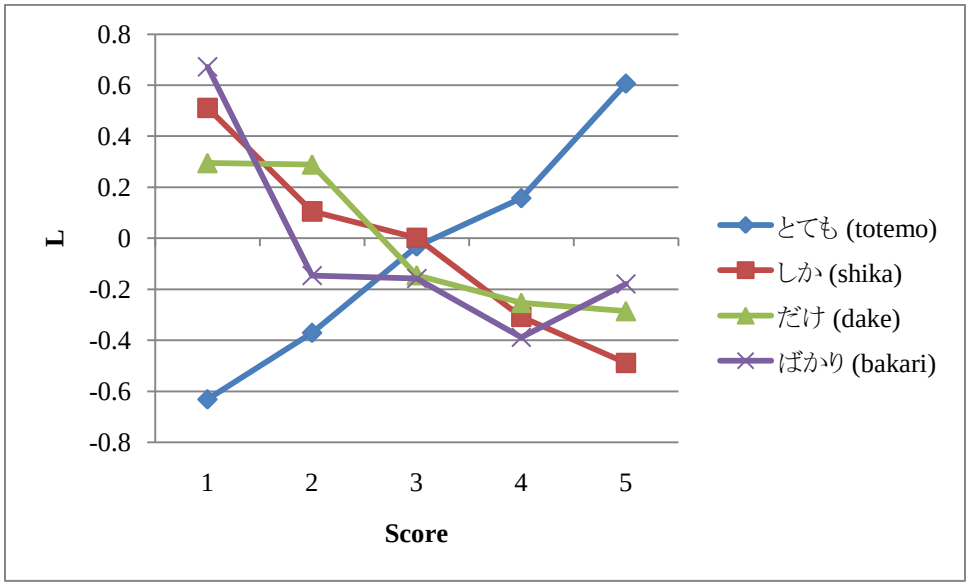


Figure 2: Restriction Words vs. Expansion Words.

In this figure, we see that totemo ("very"), usually a word of expansion, is inversely correlated to words of restriction, shika, dake, and bakari, which mean "only.

Comparing the tokens in Table 3 to those in Table 5 and *Figure 3*, there is evidence that, while five-star reviewers are more likely to describe what *is*, one-star reviews are more likely to describe what is *not.* Bigrams 12, 16, 18, and 20 in Table 5 all denote *absence* or negation in Japanese for non-living things, both the absence of a property and existential absence. As we can see clearly in *Figure 3*, all of the negation tokens which occur exhibit linear behavior.

Likewise, unigram 29, “only,” is always used with the negative form of a verb.

The inverse correlation shown by *Figure 2* is striking: *totemo* and *shika* are generally analogous to the English terms “very” and “a lot” vs. “only” in much of their usage, though we must caution from making too much of this, as many terms show a generally upward or downward trend. Indeed, all of the negation bigrams in the 1-star list exhibit this pattern, shown in Figure 3. It would, however, appear that, given the prominence of negation in strongly negative reviews – the propensity for these reviewers to describe what something is *not* or what it *lacks*– the usage of *totemo* would decline as this occurs. It is much less likely that one would describe either the extent or amount of something that is missing, especially given the alternate *shika…nai* construction[10], which excludes all other classes but the one marked by *shika* as having a property or participating in an action, rather than extending the intensity of an action vis-à-vis *totemo*, and which requires the negative form of a sentence, i.e., the absence of something. That they both stabilize is

---

[10] しか…ない s shika…nai: rougly “only” or “Except for…nothing…” This construction involves using *shika* with a negative verb to limit the extent to which something occurs. *Shika* is always used with the *nai* form of a verb or adjective, or its formal counterpart, *masen*。 Both *nai* and *masen* are also high on the 1-star list.

in sharp contrast to *shika*, which is roughly linear. All perform very differently from *tada*[11], which is treated in the section on neutral polarity.

3.3 SENTIMENT POLARITY AND DISTINCTIONS IN RESTRICTION

Also in Figure 2, we have *dake*[12] and *bakari*[13]. While, like *shika*, they are generally both translated as "only," they have qualitative differences in meaning to the overloaded term "only" in English[14]. Jorden notes the following regarding *shika* and *dake*:

> The question that immediately arises relates to the difference between shika and dake. X dake means that just X – no more, no less – is relevant. In contrast, X sika definitely implies an occurrence less than might be

---

[11] ただ: tada. However

[12] da ｲﾅ, dake: only, just. Unlike the NPI (see footnote 13) *shika*, which always takes the negative form of a verb and implies the absence of everything but that which it follows , dake is more robust, meaning simply "only" or "just" more broadly.

[13] bakari, bakari: only, "nothing other than." This could also be translated as "only," but it more specifically expresses that what is or is being done has precluded the possibility that anything else might be done; that is, the space of possibilities is entirely taken by that which bakarimarks. It is therefore more intuitively negative in sentiment in many classes. Jorden notes that bakari, "'only,' 'only just,' 'little else except for' occurs in a number of different patterns. Their variety reminds us of [dake] ."[12]

bakariis not in the top 30 unigrams for 1-star ratings, but it is nevertheless significant and pertinent to this discussion.

> expected: as indicated by the negative that follows, there was no occurrence with the exception of X. Thus, in reply to Oozee miemasita ka, 'Did many people attend?' if attendance was actually fifty and considered a small number under the circumstances, the reply must be: Iie, gozyuu-nin sika miemasen desita. Compare: dake miemasita. 'Just (= exactly) fifty people. Just fifty people attended.'[17]

Therefore, *shika* may very naturally connote a sense of disappointment. "I was expecting more than there is," may be an appropriate cognitive interpretation. While it is possible that there is a positive sentiment associated with such constructions, the quantitative evidence suggests that this is not usually the case.

Figure 2 illustrates some important divergent behavior among the varieties of restriction words: while the trends are all clearly downward, what happens in between is of note. Whereas *dake* is stable from 1-star to 2-stars and relatively stable for scores 3-5 (exhibiting a slight negative slope), the more preclusive *bakari* has significant *L* score only for 1-star reviews. The particle *bakari*, moreover, is relatively stable for scores 2-5, (the slight dip in the four star category notwithstanding), indicating that its usage is heavily

negative, as intuition suggests. We have, then, evidence of distinct trends in the usage of these restrictive formulations: we may differentiate between the usage expectations of *dake*, *shika*, and *bakari* as they pertain to sentiment.

Table 7: Negation and Exclusion (shika, ja nai, nai desu, masen, totemo, bakari, dake)

Token	Translit.	1	2	3	4	5
しか	shika	0.510983	0.105243	0.001792	-0.30797	-0.4886
じゃない	ja nai	0.64607	0.167333	-0.10161	-0.48014	-0.50818
ないです	nai desu	0.49199	0.346519	-0.12872	-0.31881	-0.65842
ない。	nai.	0.427833	0.501417	-0.05343	-0.49709	-0.73486
ません	masen	0.405419	0.218597	-0.06437	-0.23344	-0.47228
とても	totemo	-0.63121	-0.37048	-0.03253	0.157185	0.605945
ばかり	bakari	0.672103	-0.14681	-0.15779	-0.38859	-0.17886
だけ	dake	0.294822	0.289286	-0.14524	-0.25367	-0.28599

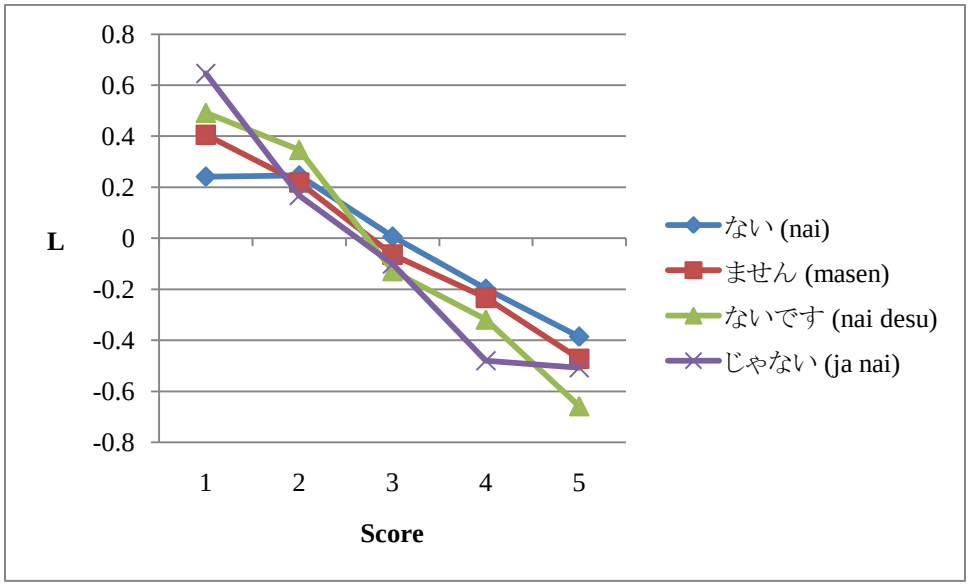


Figure 3: 1-Star Negation (not) Tokens:

All of these tokens are negative conjugations of verbs, and they all exhibit similar behavior, indicating that reviewers describe what is not more often in negative contexts than in positive ones.

3.4 Statements of Fact and "Explanatory" Statements

We have established that statements of negation are most prominent in the lower review scores, the L-scores decreasing as the review scores increase. The results, however, may at first seem duplicitous, as the bigram chart for one-star reviews is also dominated by *da* and its variant copulas, as shown by the

prominence of *n da*[15] (bigram 7), its more formal counterpart *n desu*[16](bigram 9), the less certain *deshou*[17] (bigram 22), its more explanatory counterpart *no deshou*[18] (bigram 10), *n deshi*[*ta*][19] (bigram 17), and *desu ne* [20](bigram 25). However, *desu* and its variants may be used (and often are used) in negative statements, as well as positive ones.[21] In all but the last case, the usage of *no* or *n* is employed. (We will refer to these as "explanatory" markers, though, as we shall see, it is not nearly that simple.)

Indeed, there are many complexities involved in the analysis of the usage of *n* and *no*, and further explanation is warranted at this

---

[15] んだ, n da. This is a shortened form of no da. While the informal (and often grammatically optional) copula da by itself denotes equivalency, the addition of the particle *no* or *n* before (or without) it adds an explanatory (or an emphatic tone) to the statement being made. The explanatory usage is not equivalent to "because" or "since" in English, which provide a direct, explicit causal connection. It is assumed participants in the conversation can put the pieces together from the context.

[16] んです, n desu. This is the shortened form of no desu. The copula desu is often described as a more formal version of da, though their usage does differ in certain situations. When used in the form of a question, it is seeking explanation. See the previous footnote.

[17] でしょう, deshou. This is an oft-used alternative to da or desu. A statement ending in deshou indicates that the speaker is less certain about the claim being made, though still believes is *probably* true. If deshou is said with a rising intonation (or, in written language, a question mark), it indicates a question, though the speaker believes that the answer to the question is the statement being made, similar to ne.

[18] のでしょう, no deshou. This is the explanatory or emphatic version of deshou

[19] n でした, n deshita. The past tense version of n desu. See footnote 16.

[20] ですね, desu ne. This combines the formal copula desu with the sentence-ending particle ne. Here, ne is a call for agreement, roughly equivalent to ending an English sentence with "right?" or "isn't it?"

[21] For example, *nai desu*. The copula *desu* may follow statements of negation without a change in meaning. This is common.

point. Jorden[17] claims that the difference between, *iki**masu*** and *iku **n** desu* is mostly stylistic, asserting that *iku n desu* , which makes use of *n* is "…a more indirect form, and hence is often described as softer and less abrupt. Often, the extended predicate with *n* is a pattern of familiarity…" Most, however, ascribe ***no*** *desu* primarily to explanation,[18-20] and it is generally accepted that there is no difference in meaning between *n* and *no*, the former being a shortened version of the latter. (The data indicate that, while this is usually the case, there is at least one important exception[22]). McGloin[20] attempts to "…account for various usages of no desu in a more unified way." He makes some point relevant to our analysis, including the following:

1. The addition of *yo* to *no desu* "…seems to add emphasis or [the] speaker's emotional involvement…" (p. 123)
2. The usage in question presupposes some assumed knowledge about what has occurred. Therefore, in such cases, the function is "…not to ask for the hearers explanation, but rather to indicate that the speaker assumes a certain event of state to be true." (p. 126)

---

[22] See *no desu* in *Figure 6*.

3. It is used when something is recognized to be true by both hearer and speaker. (p. 126) It indicates that "…the speaker has some knowledge of the truth value of a certain proposition." (p. 127) "Thus, *no desu* sentences are *not* [emphasis added] used in simple information giving or –seeking situations." (p. 128)

Takatsu[14] argues for understanding *no desu* in terms of *cohesion,* noting Halliday and Hasan's[21] definition:

> The concept of cohesion is a semantic one; it refers to relations of meaning that exist within the text, and that define it as a text. Cohesion occurs where the INTERPRETATION of some element in the discourse is dependent on that of another. The one PRESUPPOSES the other, in the sense that it cannot be effectively decoded except by recourse to it. When this happens, a relation of cohesion is set up, and the two elements, the presupposing and the presupposed, are thereby at least potentially integrated into a text (p. 4).

Takatsu furthermore uses the example of "and" in English: a word may have various interpretations depending on the context in which

it is used: for example, it may indicate a temporal sequence, cause and effect, both, or neither. It is, he argues, "…context (combined with real-world knowledge) which is responsible for the different interpretations in each case." He continues, saying, "So, while it is extremely important to acknowledge all of the possible interpretations of *and* in English, it would be more enlightening to discover what is in common to all instances of this construction." Thus he continues to do for *no da* in Japanese. He critiques McGloin's[20] assertion that the speaker introduces mutually understood information, in favor of the notion that the speaker is "…introducing information which is linked to what has preceded it." Takatsu believes that the use of *no da* and its formal counterparts, in cases "…where the speaker wishes to express reservation, it is natural that s/he will want to appeal to the addressee to 'fill in the gaps' as it were – to understand the message without having it fully spelt out. NO DA signals this expectation on the part of the speaker." He concludes, stating, "[he has] attempted to capture all of these cohesive properties in [his] proposal of the two basic semantic components of the NO DA construction": namely,

1. "In saying X, I am talking about something you know about."
2. "I assume you will understand why I say X now."[14]

Let us, then, consider the polarization of *n desu* and *n da* in this context. A number of possible explanations for this lopsided behavior present themselves:

1. Are strongly negative reviewers more apt to assume an extra-linguistic shared perspective with the reader of such reviews, in the manner described by Takatsu?
2. Do strongly negative reviewers feel more inclined (or obligated) to explain the reasons for their negativity?
3. If either (1) or (2) is true, is it the result of a psychological tendency to seek common ground when one is being negative: e.g., rather than strictly *assuming* a extra-linguistic shared perspective as in (1), *seeking* one ?

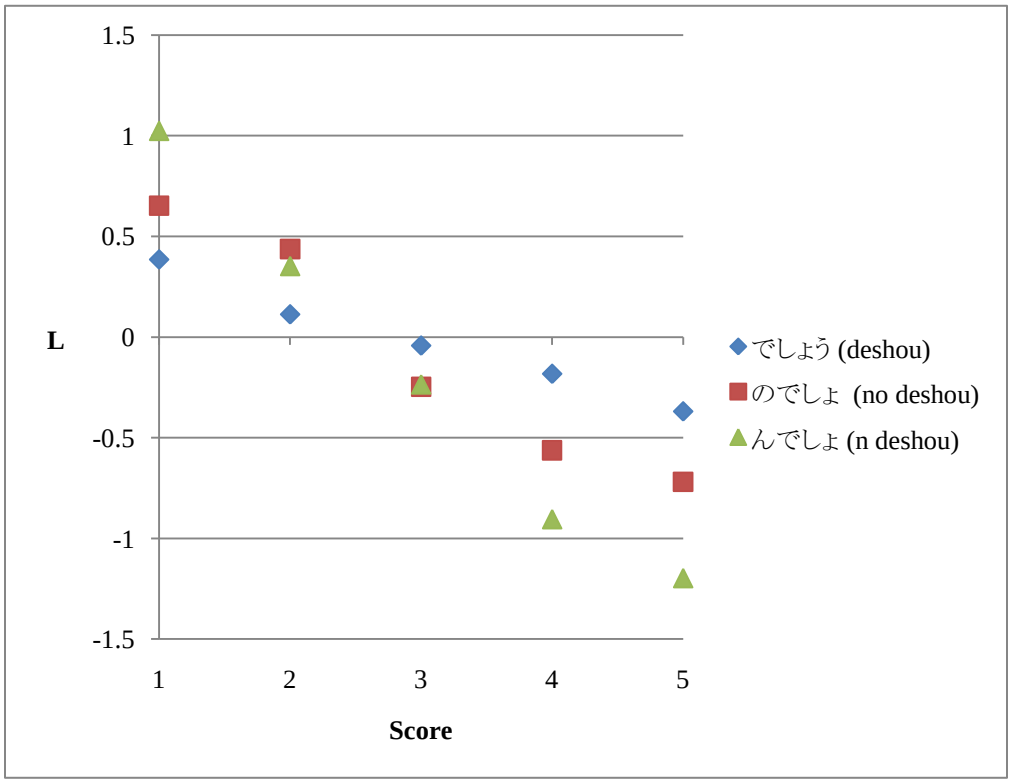


Figure 4: Deshou Usage

While deshou and its variants generally skew toward negative contexts, the versions with the n or no particle have steeper slopes, indicating that this particle tends to encode negative sentiment.

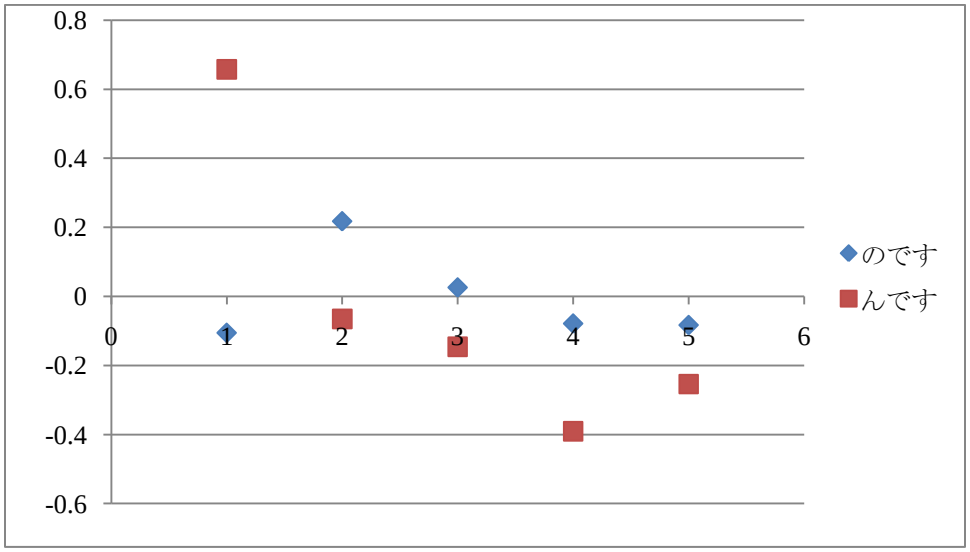


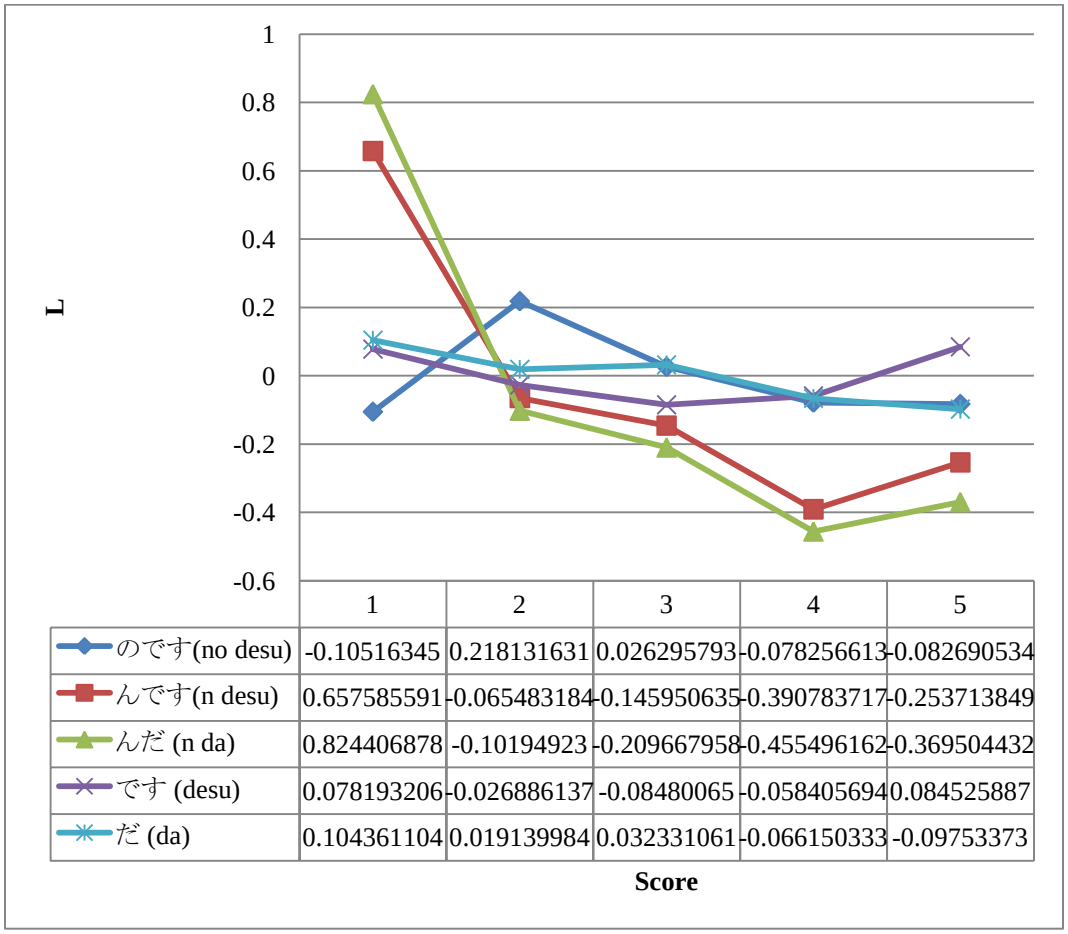
Figure 5: Explanatory Copula Bifurcation (no desu, n desu)

The particle-copula combinations no desu and n desu exhibitmarkedly different behavior in the one-star category especially.

Surprisingly, as shown in *Figure 6*, the explicit formality of the explanatory copula does not affect its representation within the respective rating classes, *except* in the case of *no desu*, which behaves more like its non-explanatory counterparts: both are relatively stable among ratings 2-5 (when compared to the highly-dominating one-star review class), exhibiting a comparatively slight but significant decrease in usage for rating class 4. This is in contrast to other forms of politeness, which, as will be shown later,

do indeed vary across ratings. We can see from both Figure 4 and Figure 5 that, while both *deshou* and the *n da* explanatory copula are dominant in one-star reviews, the more assertive *n da* is confined to this class, but the weaker *deshou*, which generally indicates less certitude, more gracefully slopes downward. Similarly, both explanatory versions of *deshou* have steeper negative slopes than *deshou* itself. This leads me to claim that *no* and *n* are primarily used in negative sentimental contexts. This is in sharp contrast to the non-explanatory versions of both of these copulas -- *desu* and *da* -- which are relatively stable across all ratings, indicative of their relative sentimental neutrality. Likewise, the explanatory *no deshou* exhibits a polarization not found in the non-explanatory (but otherwise equivalent) *deshou*. I therefore make the following claim: In general, the explanatory *no* (or *n*) is used primarily in contexts expressing markedly negative sentiment, except when combined with *desu*, exhibiting a polarization between strongly negative sentiment (rating class 1) and all other classes; when used in conjunction with *deshou*, it is likewise used primarily in cases of negative sentiment, making its descent in usage across rating classes much steeper than that of *deshou* alone. I further postulate that the less marked polarization of *no deshou* when

compared to *no da* may be explained by the fact that, on its own, *deshou* exhibits linear behavior from ratings 1-5. But why is this?



	1	2	3	4	5
のです(no desu)	-0.10516345	0.218131631	0.026295793	-0.078256613	-0.082690534
んです(n desu)	0.657585591	-0.065483184	-0.145950635	-0.390783717	-0.253713849
んだ (n da)	0.824406878	-0.10194923	-0.209667958	-0.455496162	-0.369504432
です (desu)	0.078193206	-0.026886137	-0.08480065	-0.058405694	0.084525887
だ (da)	0.104361104	0.019139984	0.032331061	-0.066150333	-0.09753373

Figure 6: Explanatory Copula vs. Naked Copula (no desu, n desu, n da, desu, da)

While desu and da are comparatively stable across the review classes, n desu and n da exhibit an identifiable pattern. The explanatory copula no desu is the exception in this case, as its behavior is relatively stable but somewhat peculiar.

McGloin argues, providing several examples, that "[g]iven certain observable evidence, the speaker makes various assumptions concerning what is going on. In this case, *n[o] deshou* is used."[20] If the uttered statement is based upon prior knowledge, based upon "observable evidence" regarding the state of the subject in question, *n deshou* is appropriate. To use McGloin's example, if one sees a baby crying, *onaka ga itai* ***n deshou*** would be appropriate, the use of *n deshou* affording the meaning, "I suppose he's hungry." If there were a lack of observable evidence, *deshou* (without *n*) would be used, he claims. He further notes that, "…even when the speaker only assumes and does not directly know the fact, he can utter…" sentences with *n deshou*. McGloin believes that "*n deshou* is more subjective, while *deshou* is based on more objective information," noting that a weather forecaster would not use *n deshou,* since he or she would sound uncertain, while someone who observes stormy weather would.

If this is the case, however, then the qualitative interpretation of this data is not clear. All usages of *deshou* have L-scores inversely proportional to review scores, with a steepening slope as the formality decreases. If we adopt Takatsu's interpretation of *no da*, it is possible to explain both the descent of *deshou* and the

polarization of *n* usage based on (3) above, ascribing it to hedging. Further research will be required to confirm this.

As stated above, the conventional assumption is that there is no difference in meaning (or in acceptable usage) between the explanatory *no* and *n* particles. While we have confirmed this in the cases of *n da* and *n desu* (Figure 5), the case of *no desu* (Figure 6) in particular indicates a substantial difference in usage from *n desu*, to which it is theoretically equivalent. But it apparently is not equivalent in actual statistical usage at all: it in fact all but overlaps with *da* for scores 3-5. It is somewhat remarkable that the usage of *no desu* mirrors that of *da* in any case, as one is typically viewed as especially "soft" and the other especially "hard," often described as "masculine" in certain contexts. The relative stability of *desu* is expected, given its neutrality, though the slightly curved shape requires further research to explain.

We have not yet considered *desu ne* due to its peculiar behavior, shown in Figure 7. We have here some quite unexpected results.

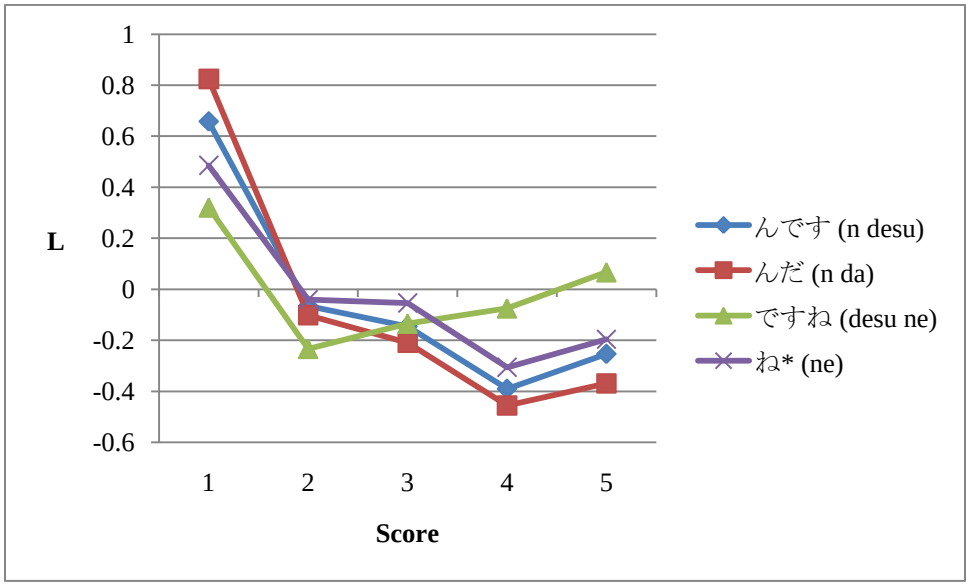


Figure 7: *(n desu, ne, desu ne, n da)*

**includes instances of desu ne*
The token desu ne notwithstanding, ne exhibits a similar pattern to that of n da and n desu. This appears to be because it is performing a similar implicature-embedding function.

Though one could certainly argue that this behavior is in some sense bipolar, *desu ne* is mostly represented in the one-star class. Its behavior, in addition, is roughly linear outside of the one-star class. It would seem, then, that the usage of *desu ne*, which is usually described as being anticipatory of agreement, either implied or explicit, is mostly confined to extreme cases, but especially to extremely negative cases. But things are not so simple: Most astonishing is that the unigram particle *ne* (the informal equivalent

of *desu ne*), which, in this corpus, would include instances of *desu ne*, follows the pattern of the explanatory copulas, rather than of *desu ne*. The particle *ne* does, in the same manner as the others, polarize the one-star and other review classes, supporting the notion, suggested earlier, that the writer is seeking mutual understanding when being strongly negative (as per postulation (3) above). It would be interesting to study the co-occurrence of *n* and *desu ne*、but that is not possible with bigrams.

But what kind of agreement, if any, is being sought? The following appeared in the title of a one-star review: *itadakenai* ***desu ne***. This translates to, "This is unacceptable, isn't it?" according to most conventional explanations of *desu ne*. But this sounds very odd in English. In English, it would be very strange to start a paragraph, let alone entitle a review, with "This is bad, isn't it?" prior to any explanation of why the other party should agree. Rather, a sentence ending in "isn't it?","right?", or even "you know?" would usually *follow* a statement that one makes with the expectation of the other party's agreement. In English, if one *were* to do so, it would be understood as adding a strong emphasis to the statement. "This is bad, *isn't* it?" would be more similar to, "This is awful." It is *so* bad that the other party's agreement is a given. This is an often

caustic, slightly forceful solicitation of agreement. Cook [22], noting that studies on *ne* are "rather scarce," proposes that "…*ne* signals an affective common ground between the speaker and the addressee." It is "…not limited to propositional content," in contrast to *deshou*. Jorden[17] , as Cook also notes, argues that *ne* with a falling intonation is not a question marker at all, but rather a way of expressing emphasis. Cook continues, saying, "*Ne* constitutes various speech functions and speech acts which call for the cooperation of the addressee. For example, *ne* is often used to get another's attention." Cook later argues that *ne* is instrumental "…in mitigating face threatening acts (FTA)," extending the meaning of *ne* to "positive politeness strategies," which are "…strategies that minimize the potential damage of a positive face (i.e. a desire to be appreciated) caused by an FTA; Cook links this to what he calls "Japanese disposition for avoiding confrontation." Most relevantly, Cook argues that "[t]he speaker often uses *ne* when s/he has to convey negative information or information that s/he assumes that the addressee will not like to hear," which fits our empirical data. However, Itani[23] disagrees with the notion that *ne* itself inherently functions in this way:

> Does the particle *ne* then affect the speaker's propositional attitude? The answer seems to be 'No'… [N]e cannot be associated with any specific level of commitment. *Ne* can be appended to utterances in which sentential attitudinal adverbs such as TABUN (= probably) and ZETTANI (= for sure) are used and it can be appended to auxiliary verbs such as DAROO (=will/may be) and NICHIGANI (= must be)…[These do] not further convey weakened and strengthened speaker commitment respectively. Contrary to Brown & Levinston (1987), *ne* itself is not a hedge which communicates the speaker's limited commitment. *Ne* has some other function than having to do with propositional content or attitude.

She, moreover, notes that the use of *ne* may, in fact, make a statement seem *more* threatening, as in his example, "You've broken a glass *ne*." Regarding this, she argues that "[h]ere *ne* communicates…the speaker's desire to establish common ground, and it has the effect of urging the hearer to admit that the hearer has broken the glass. So 'claiming common ground' is not always a

politeness strategy. It depends on what the speaker wants to establish as common ground."

Itani furthermore addresses the claim that *ne* might be used simply as an "intonation carrier," dismissing it due to the observation that "…it is not the use of *ne* but intonation put on this particle that makes [the exclamative example sentences] exclamative." But of course, Itani is referring to spoken Japanese, in which intonation is possible. In written Japanese – even Japanese written as though it were spoken – absolute intonation is an impossibility. Thus it is possible that *ne* communicates a mutually understood intonation in these instances: that is, it is possible that, based on context, the reader will read *ne* in a certain way understood as intonation, adding more information to the statement.

Itani concludes that, "It seems that *ne* makes a contribution to higher-level representations whether they are higher-level implicatures or higher-level explicatures," likening it to "please" in English in this regard.

If this is this case, as Itani and others note, *ne* is used to "establish common ground," and, as Itani argues, it embeds implicatures and explicatures, then we begin to understand why *ne* usage so closely mirrors the explanatory *n da* and *n desu* in our data.

In many cases, *ne* is actually performing a similar, if not identical, function. Rather than simply asking for agreement, or even merely emphasizing – both of which may be done with *ne* in certain contexts, as shown above – *ne* is, *in general*, responsible for implicature and explicature embedding, performing the function of establishing mutually-understood information. Both *ne* and *n desu* indicate what one might call an unspoken "sublayer" to the discourse: inferred understanding existing below the surface or between the lines. Our empirical data supports these clearly parallel (but independently analyzed) notions of *ne* and *n desu* usage, since they appear to not only typically encode negative sentiment, but also follow a predictable pattern. This indicates that these kinds of implicature and explicature embeddings are more commonly used in cases of negative sentiment. This still leaves unresolved the issue of why *desu ne* behaves differently from *ne*, but I suspect that this is due to politeness.

3.5 THE SENTENCE-FINAL PARTICLE YO

Like *ne*, the absolute function of *yo* (one-star unigram 17) has been the subject of some controversy. Often, in Japanese textbooks,

it is said that it roughly translates to “I say” or simply that it implies that the speaker is giving new information to the hearer, though Matsui[24] finds this to be inadequate for all cases. She, like Itani,[23] analyzes *yo* from a relevance-theoretic framework, arguing that *yo* “…overtly encodes a guarantee of relevance, and gives the hearer additional encouragement to pursue the relevance of the utterance,” adding that it is analogous to certain types of repetition, as in “This is a cold, cold place.”

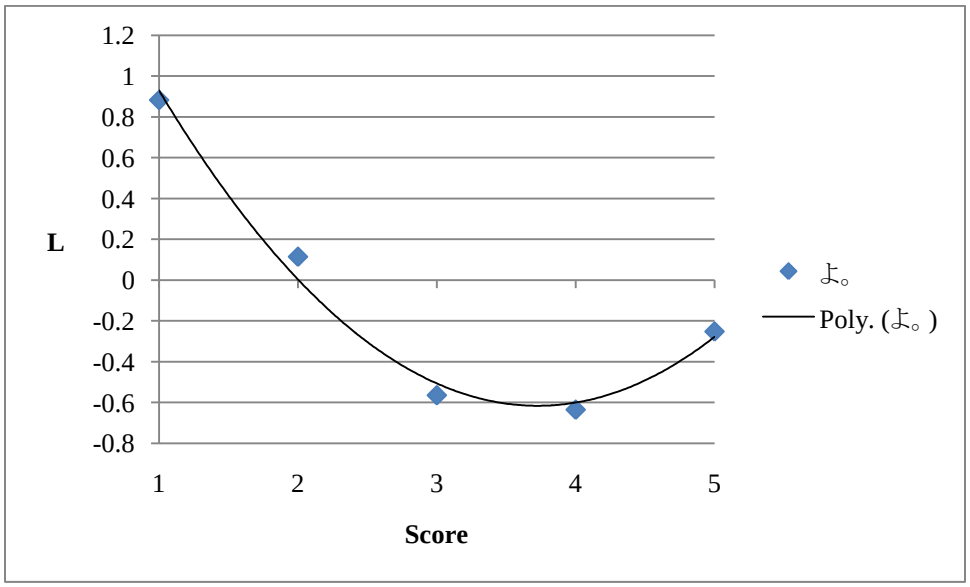


Figure 8: *Yo* Usage

Yo is used primarily in negative contexts.

This could certainly be construed as emphasis, and, indeed, *yo* often does carry a feeling of strong emphasis. Matsui claims, "In all cases, YO indicates that the contextual effects achieved by the utterance are greater than the hearer would have expected for the same utterance without the particle." Furthermore, this account "…suggests that an overt, additional, guarantee [yo] of relevance presupposes the speaker's judgment that the hearer needs such a guarantee as an extra encouragement…" for the hearer to figure out the nature of the relevance. This, taken in the context of our curved shape, makes sense. Davis[25] argues similarly, arguing that the usage of *yo* explicitly encodes what is implicit in languages such as English: namely, that "[w]ith assertions, *yo* is used to indicate that the asserted content is sufficient, given the common ground to make some action optimal for the addressee. On the other hand, with imperatives, *yo* indicates that the pre-update common ground is sufficient to make the action encoded by the imperative optimal, relative to some contextually specified ordering, for the addressee," where "optimal" means "optimal to the speaker." It is not clear, in the context of reviews, any way in which optimality to the speaker is connected to *yo* usage, but Davis also argues that *yo*, in many

instances, creates an association with the immediate context of the situation.

Very frequently in negative reviews, the reviewers are, in essence, warning against purchasing a product due to some unfortunate experience; likewise, positive reviews are very frequently urging (or reassuring) a potential buyer that a purchase is warranted. Consider the following from the title of a five-star review: **いい作品ですよ.**[23] This not only communicates that this is a good product, but, in accordance with Matsui's model, *assures* the potential buyer that, yes, this *is*, in fact, a good product. Also, consider the following, also from a five-star review: **なんでこんなに叩かれてるか分かりませんが、普通に楽しめましたよ**.[24]

And again,

私は今回も楽しかったですよ.[25]

In both of these examples, the reviewer is assuring the reader that they *did*, in fact, enjoy the product. In the former, we have a compound sentence, in which the second half of the compound is

---

[23] Ii sakuhin desu yo. "This is a good product."

[24] Nande konna ni tatakreteru ka wakarimasen ga, futsuu tanoshimemashita yo. "I don't understand why you're all denigrating [flaming], but I generally enjoyed it."

[25] Watashi wa konkaimo tanoshikatta desu yo. "For me, this time was enjoyable, too."

strengthened by *yo*, guaranteeing the relevance as juxtaposed to the others' negative opinions.

The following is taken from a one-star review:

あんなアクションじみたゲームじゃないんだよ。[26]

This example contains both *n da* and *yo* in a complex particle compound, and contains the negation form of the verb, covering much of the spectrum of the negative polarity terms we have addressed. The addition of *yo* garners attention and invites the reader to unearth the way in which this is relevant.

In Figure 8, we see that *yo* fits a second degree function quite well, though the *L* score for five-stars is still below 0.

3.6 Opinion Transference with Personal Pronouns

The pronoun *jibun*[27], often simply translated as "oneself," occurs in both our unigram and biagram five-star lists: unigram 19, and its

---

[26] Anna akushon gijita geemu ja nai n day o. "This is not a game with that kind of action."

[27] 自分, jibun, self

possessive counterpart as bigram 12. Additionally, 私[28] is unigram 24. The occurrence of personal pronouns in this rating class is relevant in and of itself, but also relevant is that the usage *jibun*, in particular, may have some significant cognitive implications. Indeed, Japanese contains a host of personal pronouns (even as they are often neither required nor used in sentences). To name a few, we have, *watashi*, *ware*, *ore*, *atashi*, *boku*, and the reflexive pronoun *jibun*, all of which mean some version of "I," though Japanese has no exact equivalent of "I."[26] This is merely a small sample of the dozens of possibilities. In English, there are only "I" and "me," which are of course used according to grammatical constraints (at least in theory), and lack an absolute distinction between the inner and relational selves. Japanese, by contrast, has a slew of pronouns – if, indeed, they are pronouns and not *roles* -- which depict how one views his or her socio-relational status at the time of utterance. That *jibun* appears more often in the higher review classes, likely indicating a heightened exposure of personal feelings, is relevant, particularly in light of multilingual and multicultural work done by Su et. al., comparing blog trends across languages [27]. They note that, "In many of [their] comparisons, Japanese bloggers were

[28]私, watashi. "I"

exceptions." Of relevance to us in the present context is the fact that, compared to bloggers from other regions, Japanese bloggers were much more likely to conceal their identities, "…even with the use of aliases," while at the same time exposing their feelings.

As with most of the structures we have addressed, there are a number of theories which might explain their usage in this context. Ono et. al. [28], working from recorded conversations, note, for example, that *watashi* specifically may serve as an emotive function. (Interestingly, in his data, the more feminine equivalent, *atashi* was used more often, though this does not occur in significant numbers in our review data, perhaps due to of the extremely feminine tone it presents, which may seem inappropriate in a review setting.)

It is well-noted [26, 29-30] that *jibun* refers to one's *true* self – one's internal self, whereas the other pronouns, such as *watashi*, refer to the various masks that one wears over the naked self, *jibun*. Using *jibun* in situations normally reserved for the public self sounds peculiar, and has the effect of rendering the speaker exposed.[29]. Kuwayama[30], in addition, constructs a model of the Japanese self, wherein the self is defined in relation to others, with *jibun* at the center, and Kunishige[15] asserts that *jibun* indicates the presence of empathy.

The prominence of *jibun no*, the possessive form of *jibun*, indicates that *jibun* is likely being used in conjunction with *watashi* in many cases. As shown in Figure 9, for the most part, they are highly correlated. The data suggest that the private self appears to be much more willingly exposed – one's inner attitudes, emotions, or abilities – in positive contexts. This is explicable by the psychological distance generally attributed to negative situations. This fits the general trend of the data, which suggest that users are, at least at some point, more psychologically distant in negative contexts. Furthermore, Suzuki[31] has argued that *-tte* (one-star unigram 28) and *nante* (one-star unigram 10) are markers of psychological distance. Suzuki does note that psychological distance can be associated with positive emotions, such as in the case of admiration for someone which makes him or her seem unapproachable. However, our data suggest that, based on the assumption that *nante* and *-tte* are markers of psychological distance, it is generally negative.

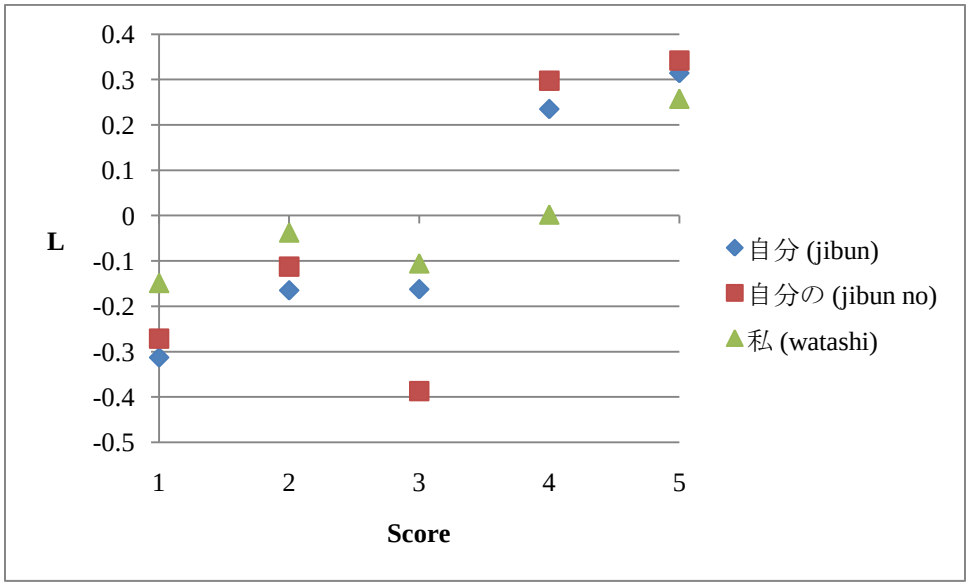


Figure 9: Personal Pronouns (jibun, jibun no, watashi)

Personal pronounces are favored in positive contexts in general. The usage of jibun (one's [internal] self) indicates an explicit expression of one's internal state.

3.7 INTERROGATIVES

In our one-star list, there are several indicators that questions are popular among disgruntled writers. While certainly not all Japanese questions end in *ka*, many do, and almost all do in polite speech. The following tokens indicate that a question is being asked: *ka*?, *desu ka*, *no ka*, and the question mark itself.

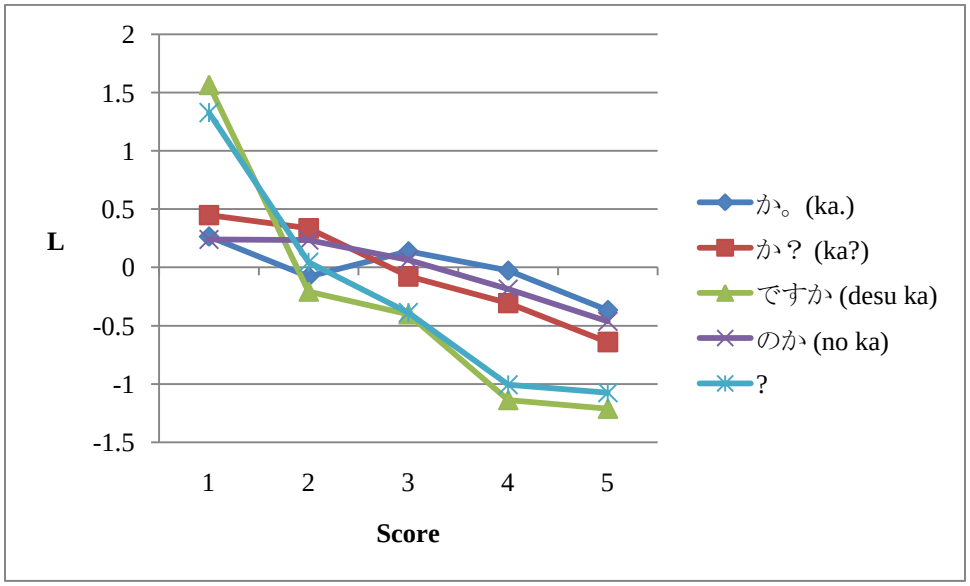


Figure 10: Question Markers (ka., ka?, desu ka, no ka, and "?")

Users tend to ask questions in negative contexts, but some question markers are more stable than others. The most formal, desu ka, and the question mark exhibit the most marked skewing toward negativity.

Without a doubt, interrogative usage is heavily skewed toward highly negative situations. We have, as well, two distinct patterns of usage: on one hand, we have *ka*。and *no ka*, with comparatively modest curves; on the other, we have *desu ka* and *ka*？, and the question mark itself, which do not. It is not difficult to imagine a situation in which an irritated person might ask many (perhaps rhetorical) questions. The following is an excerpt from a two-star

Amazon review: ディレクターの降板が響いているのでしょうか？[29] In this example, an irritated reviewer is questioning why the product is so bad. Essentially, the reviewer is saying, "I guess the director's departure is why this is so bad," , but it is phrased as a rhetorical question, adding *deshou* to indicate that he or she is heavily leaning toward to the affirmative conclusion. Consider, as well, the following two-star excerpt, which lacks a question mark: 。素材はいいのに調理は下手だった、といった感じでしょうか。[30] Again, the combination *of deshou ka is* used, and this rhetorical question appears to be directed toward the reader of the review. In fact, nearly every example that I examined in the negative category that used *ka* used the combination of *deshou* and *ka*。 This does appear to be a rhetorical device used when asking rhetorical questions, perhaps slightly sardonic, much in the same way that *ne*, analyzed earlier, was shown to be used; but in this case, the fact that it is a question is somewhat artificial. It is as though, in English, one were to say, "I guess that the director's leaving affected the quality?" The speaker believes it to be true, but it is not a *real* question; it is a statement

---

[29] Direkutaa no koubun ga hibiiteiru no deshou ka? "The director's departure had an effect, I suppose?

[30] Sazai wa ii no ni chouri wa heta data, to itta kanji deshou ka. "Didn't [you] get the impression that, despite the fact that the pieces are all there, the preparation wasn't good?"

framed as a question, which leaves open the possibility, though not the expectation, that the assertion is incorrect. In general, we would expect rhetorical questions in the present context, as actually receiving an answer is most unlikely. It is worth asking, then, how question markers compare to *ne* and *deshou*. This is shown in Figure 11.

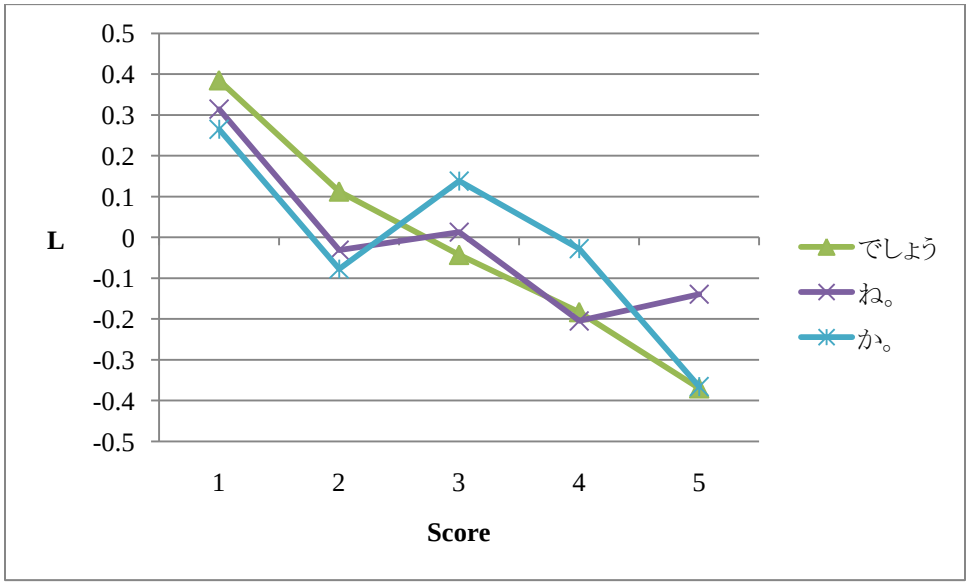


Figure 11: *deshou, ne*, and *ka.*

The correlations here are not especially striking. In general, it appears as though *ka* questions are pseudo-rhetorical in nature. Based on my examination of actual reviews, it appears that the reviewers tend to ask *why*, but the questions do not appear to be

directed toward anyone, in contrast to *ne* usage, which indicates that the reviewer is addressing the reader: whereas *ne* is, by its very nature, relational, *ka* is not.

3.8 A Few Words Regarding Politeness

As described in Appendix A, Japanese has explicitly polite verb forms, which are completely absent in English. Five-star unigram 10, *o*, indicates that this is important. The *o* honorific token, not to be confused with the direct object marker [*ɯ*]*o,* is attached to the beginnings of some words to make them especially humble. For verbs, the *masu* (past tense, *mashita*) ending is the more polite, humble form. (At times, they are combined). Is humbleness correlated to sentiment? Figure 12 clearly shows that it is in this domain.

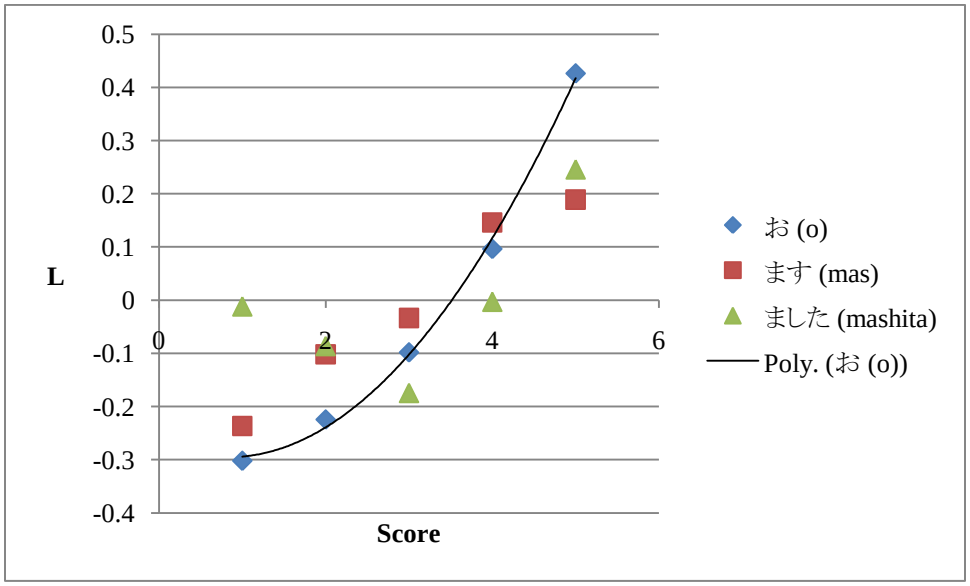


Figure 12: Politeness Indicators (*o*, *masu*, *mashita*)

The honorific prefix, o, shows quickly-increasing usage as the review scores increase; the verb endings masu and mashita (past tense) are also skewed toward positive contexts, though not quite as markedly so.

The *o* and *masu* tokens are heavily used in positive contexts. The past tense *mashita* is as well, but the cause of the V-like shape is not entirely clear. As *deshita*, the past tense of *desu*, shows, there is a strong tendency to use formal past tense forms in negative contexts, and a milder tendency to use past tense forms in general, which may explain the behavior of *mashita.*

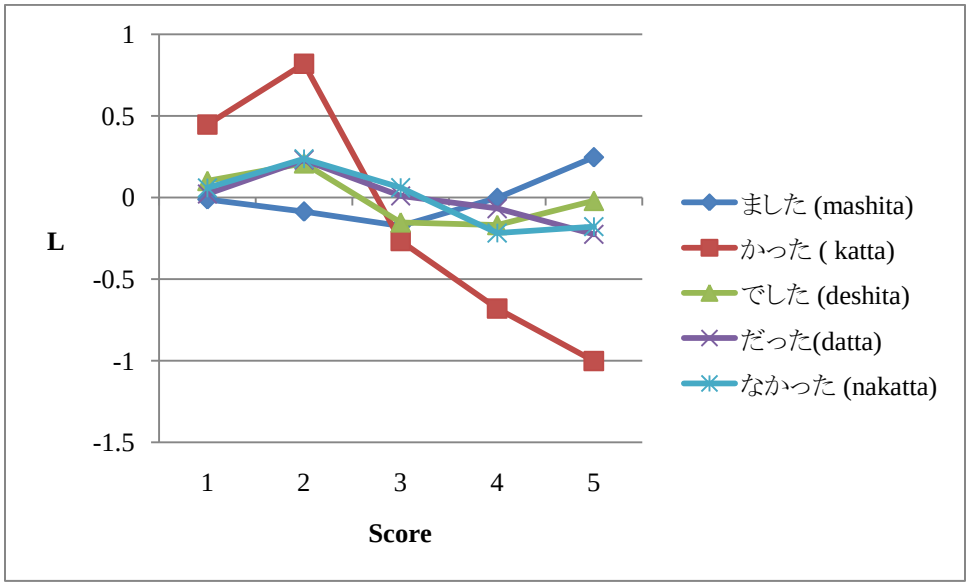


Figure 13: Past Tense Forms (*mashita, katta, deshita, data, nakatta*)

Here we see that informal past tense verb forms (katta) are skewed toward negative contexts. Interestingly, the negative version, nakatta, is relatively stable, as is the formal equivalent of katta, mashita.

3.9 SECTION 3 CONCLUSION

I have analyzed the notions of tense, politeness, questions, negation, personal pronouns, intensifiers, the particles *no, n, ne,* and *yo*, and their usage in conjunction with copulas. There is substantial evidence that users in negative contexts describe what is missing or absent, as shown by the high usage of

negation and words of restriction. That politeness is more prominent in higher review classes and personal pronouns are more prominent indicates not only an increase in etiquette, but more explicit openness. In negative contexts, the usage of indirectness, characterized by *n*, *no*, and questions as a way of communication are in contrast to this. The very different usage profiles of seemingly similar intensifiers such as *totemo* and *hontou ni*, in addition, make evident that the usage distinctions are subtle and perhaps not conscious. By analyzing the high-level trends, as opposed to domain-specific words, we see the divergent linguistic behavior in these two contexts of sentiment.

4 SELECTED TOKENS IN NEUTRAL REVIEWS

Having considered some of the more interesting tokens in the polar cases of one-star reviews, we will now consider some selected tokens from the three-star category. As before, I shall present lists of the tokens in question.

Table 8: Three-Star Tokens

Rank	Bigram	L	Meaning	Translit.
1	。ただ	0.506913	however	. Tada
2	だが	0.392304	but	da ga
3	としては	0.38232	"In X's capacity as a..."; "In X's role as..."	toshite wa
4	かな	0.349256	I wonder	ka na
5	思う。	0.326834	to think	omou
6	というこ	0.32145	That is to say...	to iu koto
7	なので	0.317984	because	na no de
8	かと	0.292042		ka to
9	ないか	0.262028	[indicates question/negative form]	nai ka
10	になる	0.253694	become	ni naru
11	ので、	0.251035	because	no de
12	と思う	0.246825	I think	to omou
13	人は	0.238895	The People are /person is	hito wa
14	ので	0.228081	because	no de
15	ことは	0.213601	[indicates fact or verb nomnalization]	koto wa
16	ば、	0.184727	[indicates conditional statement]	ba,
17	ある。	0.173841	[indicates something exists]	aru
18	のは	0.157603	[verb nominalizer/topic marker]	no wa
19	方が	0.157474	[explicit comparison/recommendation	hou ga
20	かもしれ	0.153984	I wonder	ka mo shire
21	思います	0.152106	think (polite)	omoiumasu
22	ような	0.151715	seems/looks/as if	you na
23	のが	0.145819	verb nominalizaer/subject marker	no ga
24	か、	0.143243		ka,
25	的に	0.140804		teki ni
26	か。	0.137867	question marker	ka
27	はない	0.137006	negative verb form	wa nai
28	には	0.132938	combination particle	ni wa
29	だろう	0.127577	informal variant of deshou/uncertainty	darou
30	し、	0.124416	indicates multiple reasons (because)	shi,

Unigram	L	Meaning	Translit.
ちょっと	0.503936	a bit/slightly	chotto
収録	0.460574	printed/recorded/taped	shouroku
残念	0.382535	[expression used in bad situations]	zannen
部分	0.359278	portion	bubun
もっと	0.355425	more	motto
あまり	0.350706	not very	amari
ただ	0.305675	however/just	tada
思う	0.256665	to think	omou
いう	0.234479	say/called	iu
どう	0.229228		dou
性	0.222213		shou
かも	0.221344		ka mo
として	0.219044	As a...	toshite
内容	0.211674	inside/interior	naibu
という	0.208395	[indicaes a title/name]	to iu
良い	0.207039	[intensifier]	ii
感	0.203219	feeling	kan
ので	0.202271	because	no de
なる	0.197835	become	naru
気	0.191915	sign of/ touch of	ke
ところ	0.189139	a place (physical or temporal)	tokoro
評価	0.182234	estimation, appriasal	hyouka
言う	0.179041	say/called	iu
あっ	0.172965	past tense existence]	a[tta]
など	0.167577	etc.	nado
ため	0.144638		tame
けど	0.143264	but	kedo
しれ	0.143016		shire
今回	0.141219	This time	konkai
か	0.140793	question marker or "or"	kai

Immediately apparent is the precipitous dropoff of *L* for the three-star category. Despite this, there are some interesting tokens which are relatively unique to this class and carry with them some relevant

implications. In this section, I shall address contrastive conjunctions and their prominence in the 3-star reviews, including distinctions between them; explicit uncertainty; explicit explanations; and explicit comparisons. Furthermore, I shall argue that they are indicative of measured thinking, or, at the very least, an attempt to portray oneself as exhibiting this.

4.1 Two Sides of a Coin: Contrastive Conjunctions

In section 1.3, we addressed various words for "just" or "only" in Japanese and analyzed their behavior in polar cases. The word *tada* (bigram 1, unigram 7) often has a meaning similar to "just," both in the temporal sense, as in "I just arrived home," and in the restrictive sense, as in, "I just wanted to help." There are actually a number of words with this pronunciation in Japanese, but the fact that bigram 7 begins sentences indicates that it probably means "however" or "nevertheless." An examination of the reviews reveals that this does indeed appear to be the most common usage, though, at times, a reviewer does say, for example, "This is just horrible." We will assume, for the purpose of this analysis, though, that it means "however." In

this case, both bigram 1, which means “however,” and bigram two, which means “but,” indicate opposition of two propositions. The more abrupt *shikashi* is also heavily favored in three-star reviews, though it did not meet our .06% threshold for the list. So, then, we have at least three versions of “but” in our three-star list. In general, we can say that *da ga* is the most casual and least abrupt of the three, while *tada* and *shikashi* both tend to begin sentences. The word *shikashi* sounds particularly dramatic. While *da ga* is sometimes used to for dramatic contrast at the start of a sentence, *shikashi* requires this. We will also consider here *kedo*, *desu ga*, *demo*, and *no ni*. Because they behave most similarly, we shall first consider *shikashi*, *da ga* (and its formal counterpart, *desu ga*), and 。*ta da*, shown in Figure 14.

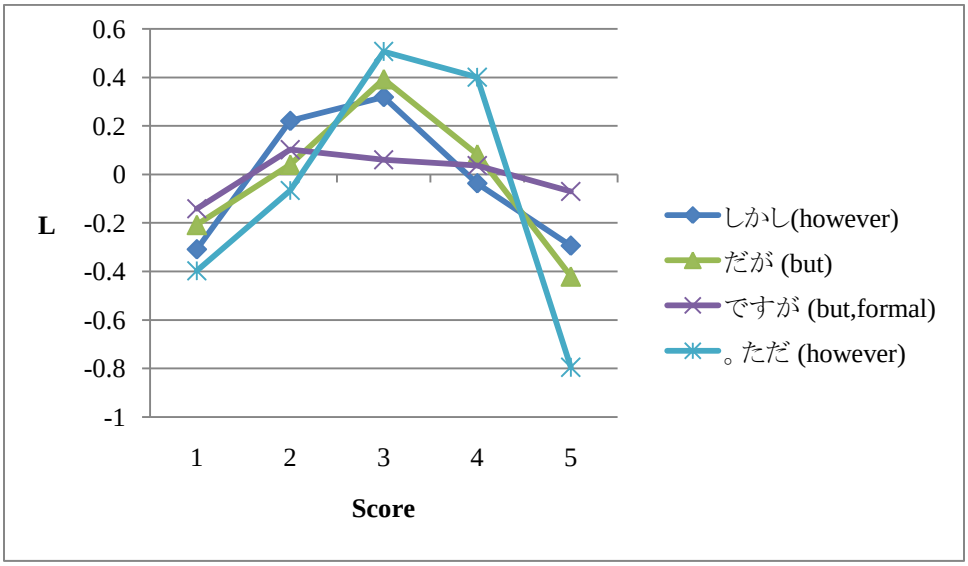


Figure 14: *shikashi (however), da ga (but), desu ga (but,polite), tada (however)*

The behavior of these contrastive terms indicates opposing propositions are prominent in neutral contexts.

While they are all likely to occur in neutral contexts, they all behave somewhat differently. The abrupt *tada* makes the most precipitous dropoff from four- to five-star reviews; aside from that, the stability from three- to four-stars is noteworthy, probably indicative of a user's explaining why he or she is not giving a full score. Five-star reviewers presumably have little need for an abrupt, attention-getting contrastive device. Both *shikashi* and *da ga* exhibit more predictable declines from the peak, whereas *desu ga* shows

relatively little variance overall. When weighing pros and cons, *desu ga* appears to be the most neutral option of the three. We see again that politeness levels do indeed carry over to sentiment.

In Figure 15, we see a quite different behavior: The conjunction *kedo*, a generic and extremely common word for "but," is used in negative contexts. Often, in Japanese speech, *kedo* effectively ends a sentence, and the hearer is left to piece together the implication. (This is also true of *da ga* and its variants). This is often done in situations when it is thought to be rude to finish the thought. The token *demo* may be either a conjugation or the start of a sentence, often translated as "even though," or "but," depending on the context. Its counterpart, *te mo*, means "even though." We can differentiate the two by ensuring that *demo* starts a sentence (meaning "but"). The conjunctions and *te mo* do, in fact, behave very differently. The chiefly negative *no ni* has a meaning analogous to "despite." The difference between "despite" and "even though" (or even "even though" and "though") is by no means obvious, even in English, but this is how these tokens are commonly translated.

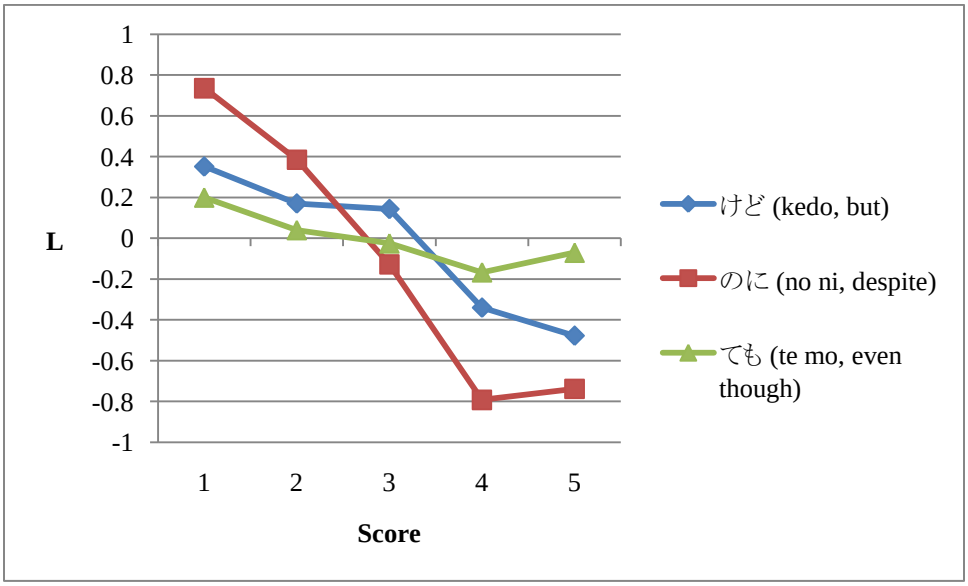


Figure 15: kedo (but), no ni (despite), te mo (though)

These contrastives, especially te mo and no ni, likely emphasize the second second half of the contrast, which tends to be the negative aspect.

Upon further inspection, we find that, usually, the particular version of "but" chosen is less crucial to the distribution than whether or not the chosen one begins a sentence (Figure 17). When this is taken into account, *demo* and *shikashi*, in fact, exhibit quite similar behavior, especially from scores 3-5. While *kedo* is not shown due to its sparsity at the start of sentences, it exhibits similar behavior when it does occur. We also see that *nagara*, which

translates to “while,” both in the contrastive and temporal senses, is inversely proportional to *no ni*, though the fact that one of them would be used in chiefly negative contexts is certainly not obvious.

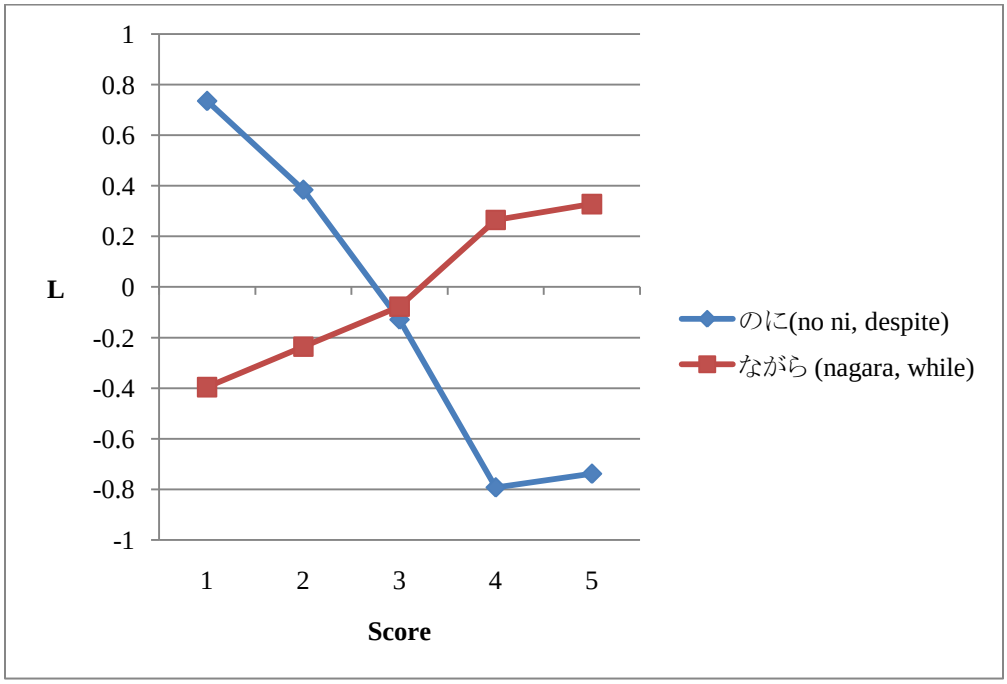


Figure 16: “Despite” vs. “While” (*no ni, nagara)*

Despite their similar nature in English, “despite” and “while” have different distributions in Japanese.

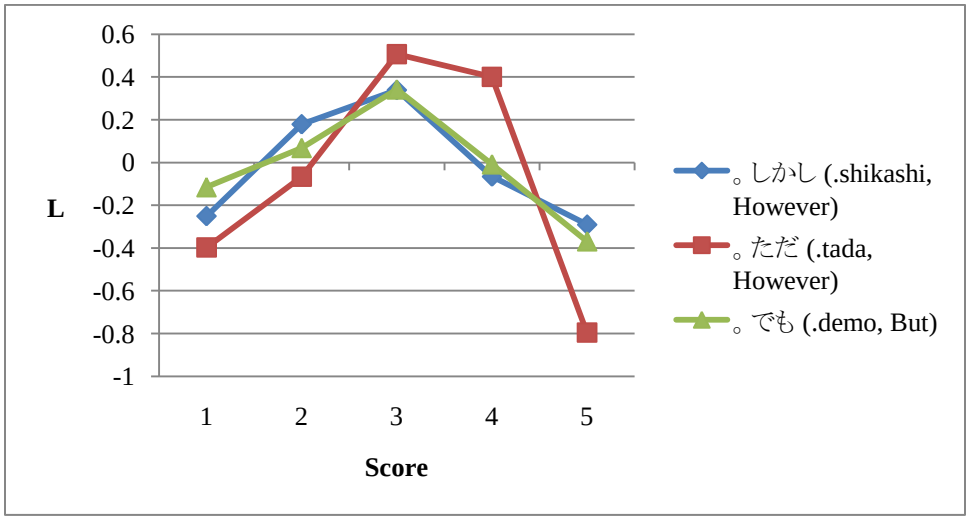


Figure 17: Sentence-starting contrastives shikashi, 。tada, 。demo

These sentence-starting conjunctions are used most frequently in the 3-star category.

We have confirmation, then, of several statistically variant ways of expression opposition in various contexts. In contexts in which writers are attempting to balance positive and negative viewpoints, it would seem, abrupt, sentence-initial opposition words are employed. In negative contexts, clause-final conjunctions are preferred, as is *no ni*. This is explicable by the notion that *no ni* emphasizes the second half of the conjunction. For future automatic sentiment analysis

tasks, weighting of various constituent parts of a sentence based on the particular conjunction used may prove fruitful.

4.2 EXPLICIT SUBJECTIVITY AND UNCERTAINTY

Having analyzed *deshou* in Section 1 as a device for making less-than-certain statements, we turn our attention now to *ka na* (bigram4), *to omou* (bigram 12), and *ka mo shire[nai]* (bigram 20) usage. The usage of *ka na* and *ka mo shirenai* both connote a sense of uncertainty and are often translated as "I wonder"; the sentence-final *to omou* may also connote uncertainty, but is translated as "I think."

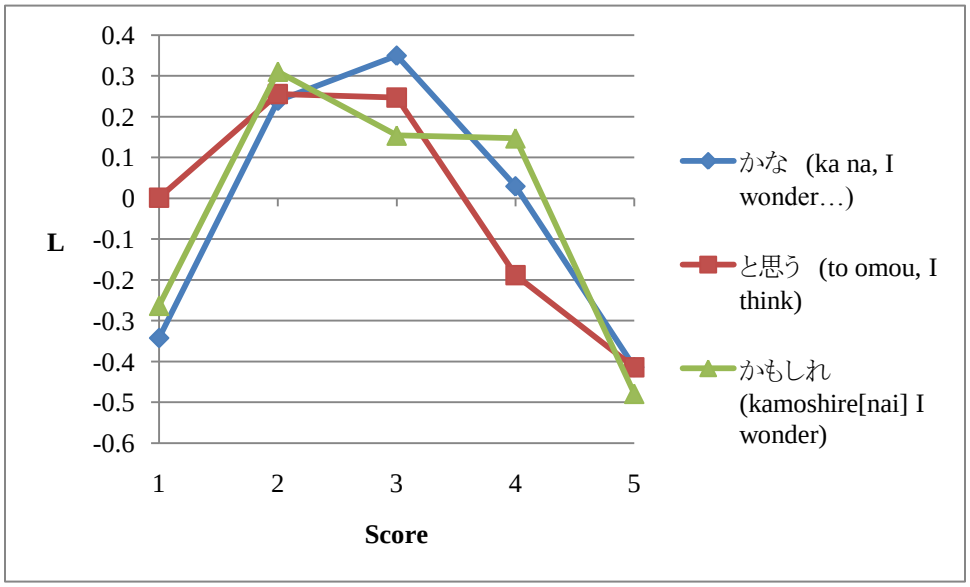


Figure 18: *ka na (I wonder), to omou (I think), kamoshire[nai] (I wonder)*

The explicit uncertainty expressed has a decidedly less certain tone than that shown by deshou, the usage of which skews toward negative contexts.

We can see in Figure 18 that they all behave similarly: they are very unlikely to occur in the extreme cases, but they are by no means excluded from the three intermediate ratings. From these data, we may glean a picture of the usual ranges for each of these subjective terms. We see that *ka mo shirenai* is the broadest, spanning rating classes 2-4; we see that *ka na* and *to omou* occur mostly in the range of 2-3; and, as shown earlier, we see that *deshou* is principally used in negative situations.

When taken with the our data from Section 2.1 regarding rhetorical questions, we see that, while users appear more likely to ask explicit questions in very negative contexts than in neutral ones, they appear to be more inclined to express their opinions with reservations in neutral ones. It is certainly true that they express their opinions *indirectly* with rhetorical questions in the more negative contexts, often with greater rhetorical poignancy. The use of *ka na* or *ka mo shirenai* creates quite a weak tone. Either could reasonably be considered rhetorical questions in many instances.

However, unlike with *ka* and more direct rhetorical questions, *ka na* and *ka mo shirenai* sound much more noncommittal, despite the fact that they appear to indicate that the speaker has a predisposition to believe what he or she is wondering. The use of *to omou* indicates less certitude than simple fact-stating, generally: just as with "I think" in English, it encodes the notion that what is being stated is one's opinion and not objective fact. The usage of *deshou* is often compared to that of *ka na* and *ka mo shirenai*. The confirmation of statistically significant difference in sentimental usage is significant. This would seem to fit with what appears to be the naming of various properties of things and their corresponding comparisons. Users in neutral settings are unwilling to make forceful, direct statements or use the kind of irony inherent to the lower classes. In addition, the existence of 感[31] (unigram 18), and 評価[32] (unigram 23) indicate a degree of overt subjectivity, perhaps ironically to appear to be more objective and not overly emotional.

4.3 PROPERTIES AND EXPLICIT EXPLANATION

---

[31] 感, kan. Feeling, impression

[32] 評価, hyouka. Estimation, assessment, appraisal

There is evidence that those who write thee-star reviews are more prone to name specific properties of the item being discussed, give reasons for their opinions, do explicit comparisons, and avoid extreme language.

Consider *toshite wa* (bigram 3), *koto wa* (bigram 18), *no wa* (bigram 23), and *no ga*, (bigram 23). All of these particle combinations are necessarily involved in describing a property or an action. The use of *koto wa*,*no wa*, and *no ga* in particular indicate the nominalization of a verb phrase and declaring it to be the topic or the subject of conversation. Thus, the users who employ these are describing specific actions and presenting facts (or weakly asserted opinions, as we showed earlier) about them. The use of X *toshite wa* Y explicitly describes a fact, Y, about something in its capacity as X. Consider the following example taken from a at three-star review:

お手軽な機器としては重宝します。[33]

Naturally, if one names capacities in which something is good , such a person may name capacities in which it is bad, and vice versa. The use of *wa*, as in *no wa*, is used in comparisons more often than *ga*, sometimes in a manner similar to “one the one

---

[33] O tegaru na kiki toshite wa jouhou shimasu. “As an [in its capacity as a] easy-to-use machine, it is priceless.”

hand…;on the other hand…" or "while"; and the usage of *no wa* is considered to be generic; so, it is surprising to see that *no wa* is biased toward the lower three scores in the spectrum. (See Figure 19.) The appearance of *hou ga* furthermore, indicates an explicit comparison or a restrained recommendation[34]. We may differentiate between them, since *no hou* is used for comparisons and *–ta hou* is used for implicit recommendations. We see in Figure 20, however, that they exhibit similar behavior -- in fact, converging at rating 3 – and are more likely in class 1 than class 3. The use of *hou* is not the only evidence of comparison we see, however. In the table of unigrams, we have *chotto* ("a bit," unigram 1), *motto* ("more," unigram 5), *bubun* ("portion," unigram 4), and *amari*[35] (unigram 6).

The use of deductive and elaborative particles and words is also quite pronounced. In particular, we have *no de* ("because," bigram 11), its counterpart *–na no de* (bigram 7), *tame* (unigram 26), and, relatedly, *to iu koto* [*wa*] (bigram 6).

---

[34] In reality, *~ta hou ga ii* is more often translated as, "It is better that you [do something]."

[35] Depending on the context, *amari* may either be an intensifier or mean "not very."

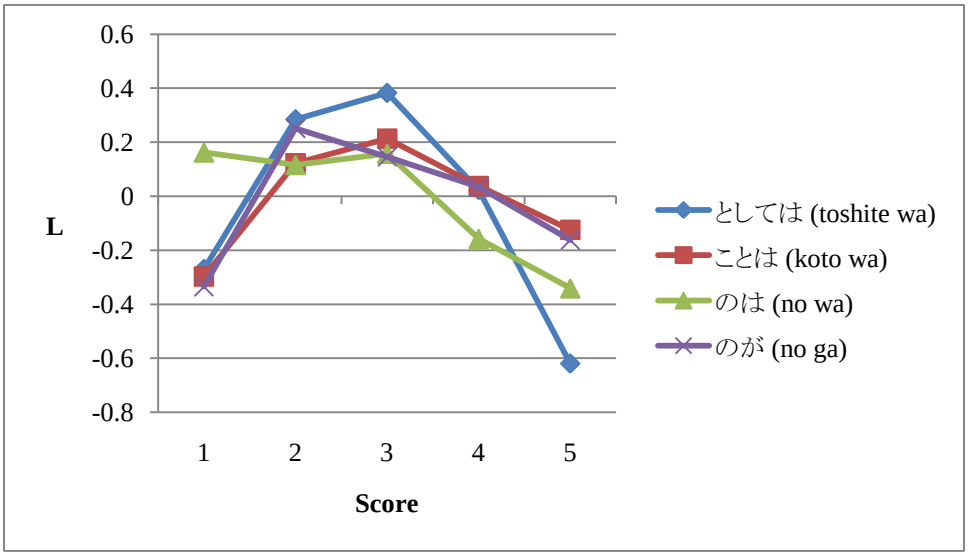


Figure 19: *toshite wa, koto wa, no wa, no ga*

The prominence of verb nominalizers indicates that something is being said about an action: an action is being further expounded upon. The usage of toshite wa, which refers to the capacity of function of something, indicates that a specific property of something is being expounded.

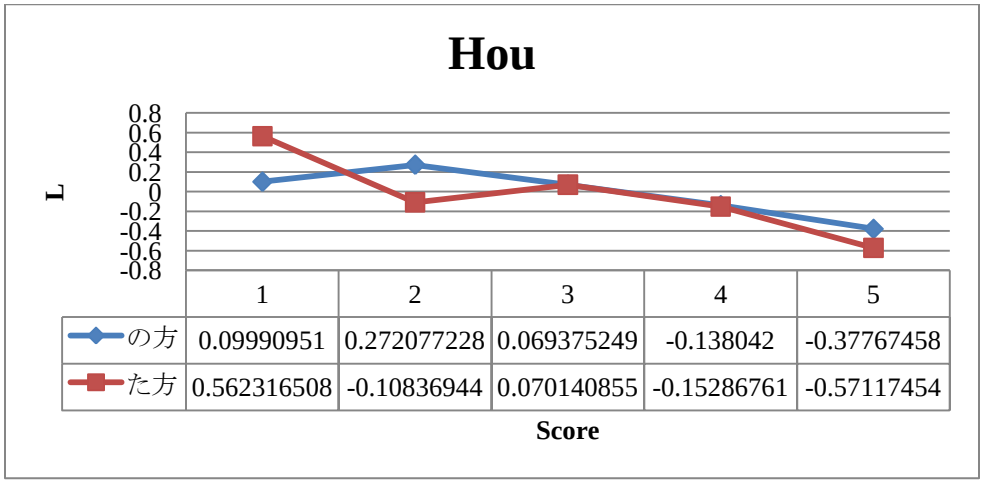


Figure 20: *no hou, ta hou*

The use of *no de* necessitates that the writer explain a reason for a proposition. In Section 2.1, we saw that conjunctions indicate further explication, here *no de* is even more explicitly so. In a similar manner, the presence of *to iu koto wa* ("That is to say") betrays elaboration, as well – a restatement of a proposition in an alternate way. We further include *-ba*、(bigram 16), the result of a particular conjugation that creates one of many possible types of conditional statements. Figure 21 makes plain that, unlike some of the tokens we have addressed, these are very clearly biased towards the middle, and negatively correlated in strongly negative reviews.

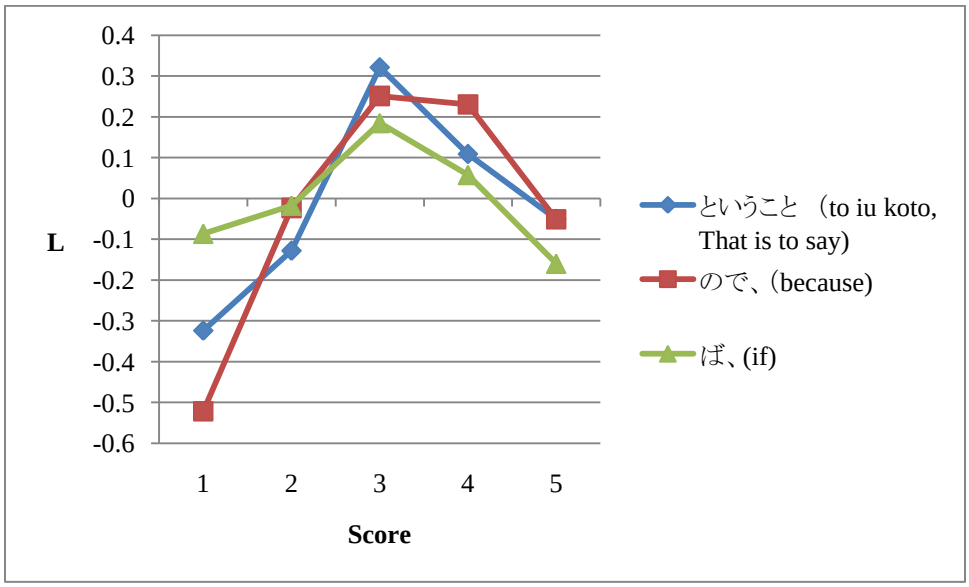


Figure 21: to iu koto wa, no de, -ba

These deductive tokens of explanation and reasoning indicate that the users in neutral categories are likely providing reasoned explanations for their positions.

What do these have in common? These are all indicative of *reasoning* – in particular, reasoned argument. While the extremes are dominated by assertions and innuendo, as we saw in Section 1, we see that the middle is dominated by deductive reasoning and hedging.

4.4 SECTION 4 CONCLUSION

In this section, I have shown the usage of terms such as *chotto*, which are expressly limiting, as evidence of hedging, in addition to tone-softening tokens such as *ka mo shire[nai]*. The explicit uncertainty avoids the language of absolutes. This, I have shown, is further supported by the usage of contrastive words, which necessarily qualify a proposition. In addition, the usage of *no de* (because) and -*ba* (if) indicate reasoned, deductive language. We see, then, that the language is measured, avoiding extremes, encoding less sentiment than disposition.

5 COMPARISON TO TRENDS IN ENGLISH

I now proceed to address the following question: How does language usage in English differ from that in Japanese in a similar context? As is perhaps obvious, but I note here again, not every explicit construct in Japanese maps to English, and even fewer map to English bigrams. Indeed, not even every Japanese *noun* maps to a corresponding English noun and vice versa. We will, however, consider some reasonable mappings, and, when appropriate, consider some concept mappings and observe the trends.

Here, I present the polar tokens, just as in Section 1, with the addition of English "translations" for the Japanese tokens. I also present the English unigram and bigram tokens for these categories and the Japanese "translations." In general, especially due to the abstract nature of many of the tokens, there is no canonical translation, and I make no claims regarding the appropriateness of such translations here, except that, together, they paint a picture of high-level similarities and differences.

Compounding this issue of translation is the fact that both Japanese and English have multiple ways of expressing various

ideas. The purpose of this section is to present a high –level overview and gain a general idea of what the top tokens for each language represent, and how this may or may not map to the other language in question. The translations I have chosen might rightly be called as arbitrary for the reasons I have outlined, but they do show interesting trends and will be the impetus for further analysis. Often, I chose not to attempt a translation either because the term was too vague without context or because there was no meaningful one- or two-word translation. In other instances, the most obvious translation simply did not occur in the alternate language's corpus.

In the tables that follow, there are cross-references to similar translated terms in the charts from Section 1. "U" refers to unigrams, and "B" refers to bigrams. This is followed by a 1, 3, or 5, to indicate the rating class of the corresponding chart, and then the entry number. For example: U:1-6 refers to sixth entry in the one-star unigram chart.

5.1 English and Japanese Polar Clauses

For the most part, for the purposes of this study, the English bigrams did not provide significantly more information than the

English unigrams. Often, the only difference was the presence of a definite or indefinite article. As a result, for the purposes of this analysis, I only include those English bigrams which are distinct from the unigrams in some meaningful way in Figure 22. We can see significant overlap, however, between the English terms and the Japanese ones.

Due to the inherent difficulty of translating Japanese tokens and particles into English – especially English bigrams – the Japanese-English lists are less substantial, but all of those selected are positively correlated in both languages.

In

Table 9, which lists the top 1-star English unigrams, we see the English words “nothing,” “don’t,” “anything,” “not,” “only,” and “never,” all of which may denote the absence of something. As shown before, this also is a defining characteristic of the Japanese

reviews. In addition to this, both English-speaking and Japanese reviewers have a tendency to ask questions in negative contexts. We also see that, while, generally, those terms which have Japanese counterparts are positively correlated, the correlations are not especially striking.

Table 9: English Unigrams-Japanese 1 Star Comparison

Rank	Unigram	Eng. L	JP Equiv. L	Cross-Ref.	JP Equiv.
1	waste	1.776072			
2	money	0.806162	1.467574242		金
3	nothing	0.730342	0.751057582		何も
4	reviews	0.649432	0.412405133		レビュー
5	bad	0.529448	0.696849002		悪い
6	?	0.508434	0.336996882	U:1-6,B:1-2,3,17	？
7	believe	0.475648	1.377934029		信じる
8	buy	0.460018	0.646588386	U:1-22	買う
9	anything	0.45325	-0.013510455		何か
10	no	0.443048		B:1-16,18,20	ー
11	instead	0.409025	0.117060628		でなく
12	mr	0.3691	0.064314614		さん
13	any	0.354972			ー
14	don't	0.338748		B:1-16,18,20	
15	actually	0.330467	0.367064934	B:5-14	本当に
16	real	0.306643	0.117060628		本当
17	ever	0.302085			ー
18	should	0.291371	0.5623165077	B:1-24	た方
19	pages	0.272422	-0.255104607		ページ
20	!	0.269921	0.0252622448	U:5-1, U:1-21	！
21	they	0.267821			ー
22	only	0.254522	0.0252622448	U:1-20	だけ
23	never	0.252543			ー
24	want	0.252487	0.714923517		ほしい
25	your	0.249271			
26	man	0.245537	-0.321307711		男
27	need	0.234584			ー
28	not	0.233842		See 10.	ー
29	if	0.233153	0.233153492	B:3-16	ば、
30	or	0.221702			

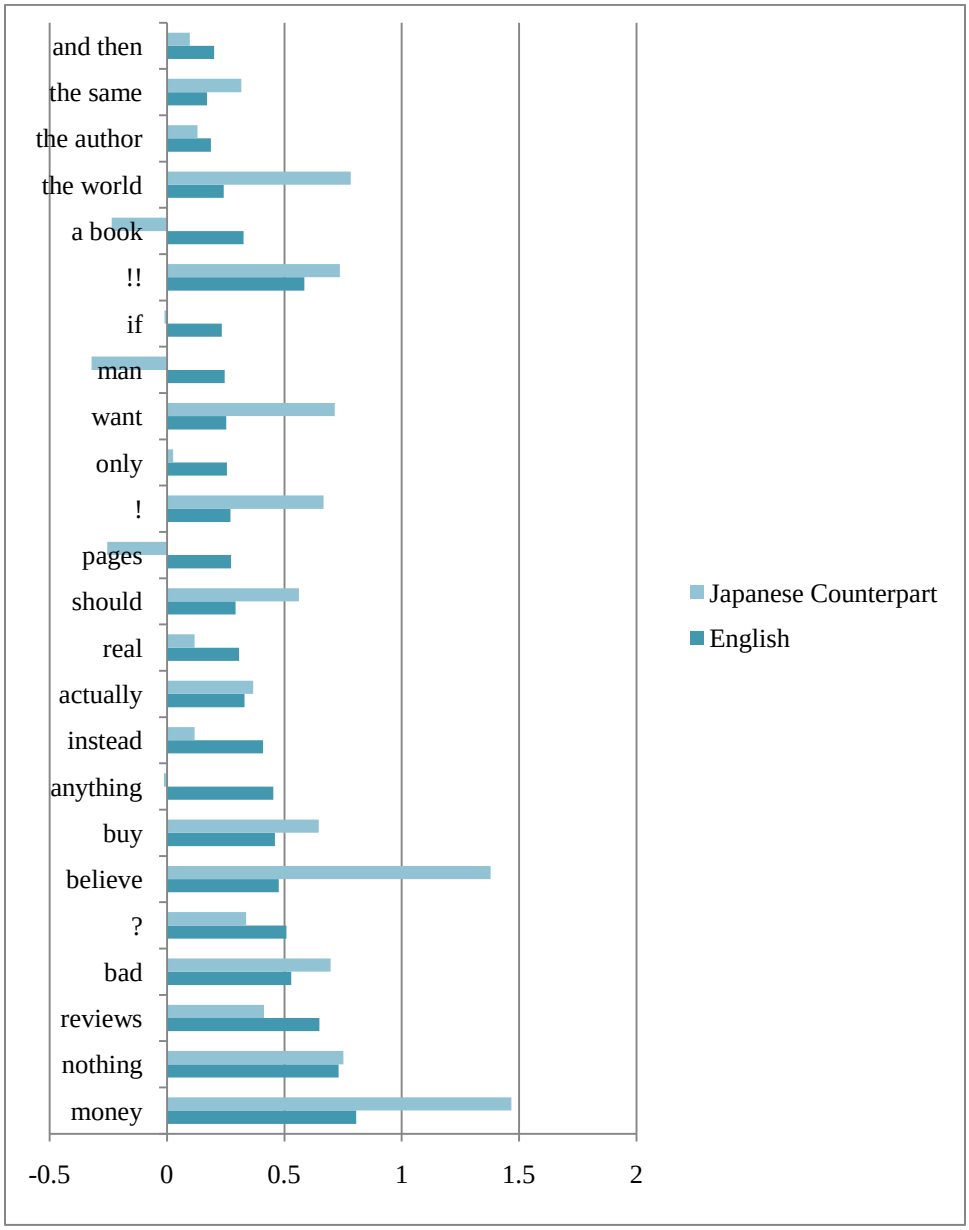


Figure 22: English to Japanese 1
Star Unigrams and Bigrams

Rank	Token	JP L	EN Equiv.	Cross-ref.	
1	章の	1.724394336	0.060042		chapter
2	か?	1.596696475	0.508434	U:1-6	?
3	ですか	1.565261098	See 2.	See 2.	See 2.
4	曲を	0.935426698	0.277685		music
5	よ。	0.882879407	-		
6	…。	0.877746691	-		
7	んだ	0.824406878	0.042215		is
8	なん	0.691584855	-		
9	んです	0.657585591	-		See 7.
10	のでしょ	0.652266681	0.197838		I guess
11	の曲	0.650630842	-		See 4.
12	じゃない	0.646069777	0.260635		isn't
13	だから	0.581479851	0.378314		Therefore
14	言って	0.538954959	0.146598		say
15	んでし	0.513857113			was
16	ないです	0.491990138		See 2.	is not
17	か？	0.447832876		See 2.	See 2.
18	ない。	0.42783282		See 12.	See 12.
19	ん。	0.425502061			
20	ません	0.405419389		See 12,16	See 12,16
21	うか	0.400114921			
22	でしょう	0.384815293			See 10.
23	うと	0.351392029			
24	方が	0.339733787			-
25	ですね	0.320686507	0.839786		isn't it
26	ね。	0.314235181			See 25.
27	をし	0.307620366			
28	から、	0.259070253	0.179748		because
29	購入し	0.25581963	0.460018		buy
30	のか	0.240350936		See 2.	See 2.

Table 10: Japanese Bigrams-

English 1-Star Comparison

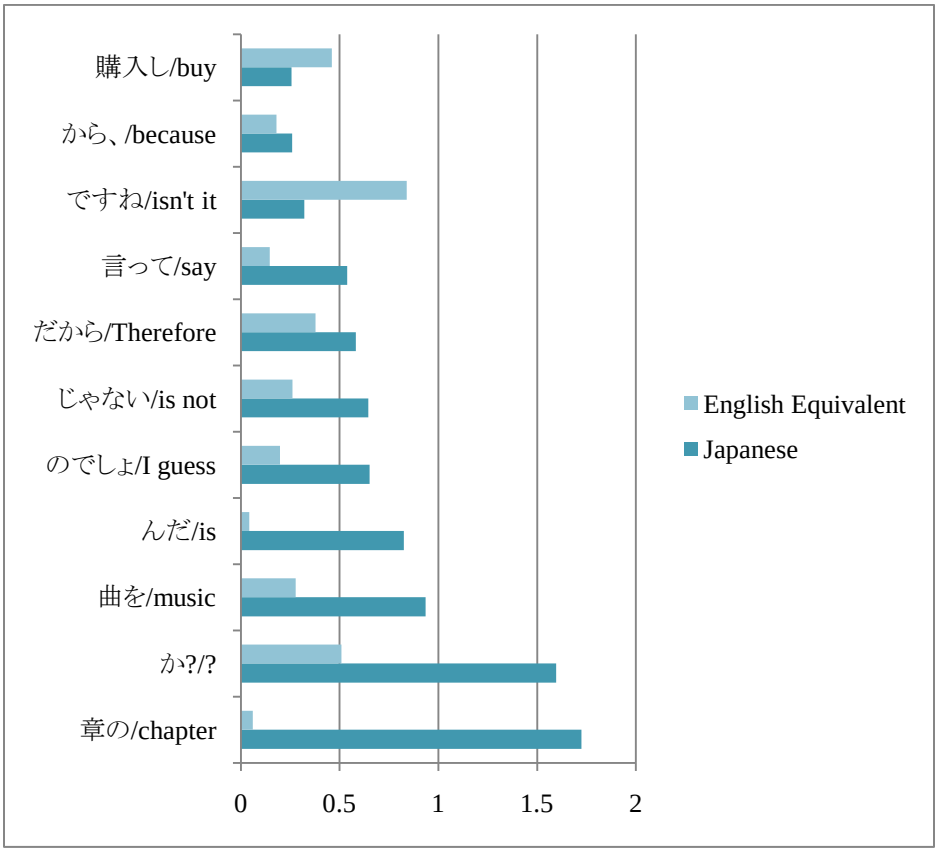


Figure 23:1-Star Japanese Bigrams to English

Tokens for five-star reviews have more obvious overlaps. The English tokens for this rating class contain a number which speak to a broad spectrum: “everyone,” “anyone,” “each”, and “both,” in direct contrast to the one-star tokens which are exclusionary to a similar

degree. More generally, there is, I believe, significance to the abundance of tokens which are *sweeping*, in that they are extreme: in addition to the ones just mentioned, we have “ever,” “must,” “great,” “each,” and “love,” There are also references to personal pronouns, i.e., “I’ve,” “my,” and “our,” “us,” and “we.” More obviously domain-specific tokens, such as “highly,” “recommend,” and “easy” also occur. Of the tokens in our list, aside from the personal pronouns, only “understand,” “help,” “change,” “years,” and “life” are not of this type.

Rank	Unigram	Eng. L	JP Equiv. L	Cross-ref	JP Equiv.
1	highly	1.1129744	0.6059451		とても
2	!	0.8123988	0.4838005	U:5-1,B:5-1	
3	easy	0.7641969	1.1202253		やさしい
4	recommend	0.6967391	0.7201213		進め
5	best	0.607807	0.4020629		一番
6	everyone	0.5993185			
7	i've	0.5906359	0.2602002		私は
8	life	0.5738668	0.5346527		生活
9	years	0.5516759	0.2422324		年
10	now	0.5341858	0.3430705	U:5-15	今
11	anyone	0.5285848	0.3604342		誰か
12	ever	0.528042			
13	change	0.5239279	0.2185044	U:5-30	なり
14	must	0.5238545			
15	our	0.5110769	0.5656978		私たち
16	great	0.5086737	0.9253639		すばらし
17	day	0.4959286	0.710534		日
18	my	0.4936759	0.3420821	B:5-12	自分の
19	love	0.4728278	0.1287248		大好き
20	every	0.4704668	0.429673		すべての
21	put	0.4661862			
22	read	0.424647	0.4435067		読み
23	each	0.3913958	0.4171623		それぞれ
24	understand	0.3723253	0.1961919		分かり
25	help	0.3472531			
26	both	0.3280449			
27	will	0.3131939			
28	us	0.3092554			See 15.
29	we	0.2777361			See 15.
30	old	0.2714594	-0.090549		古い

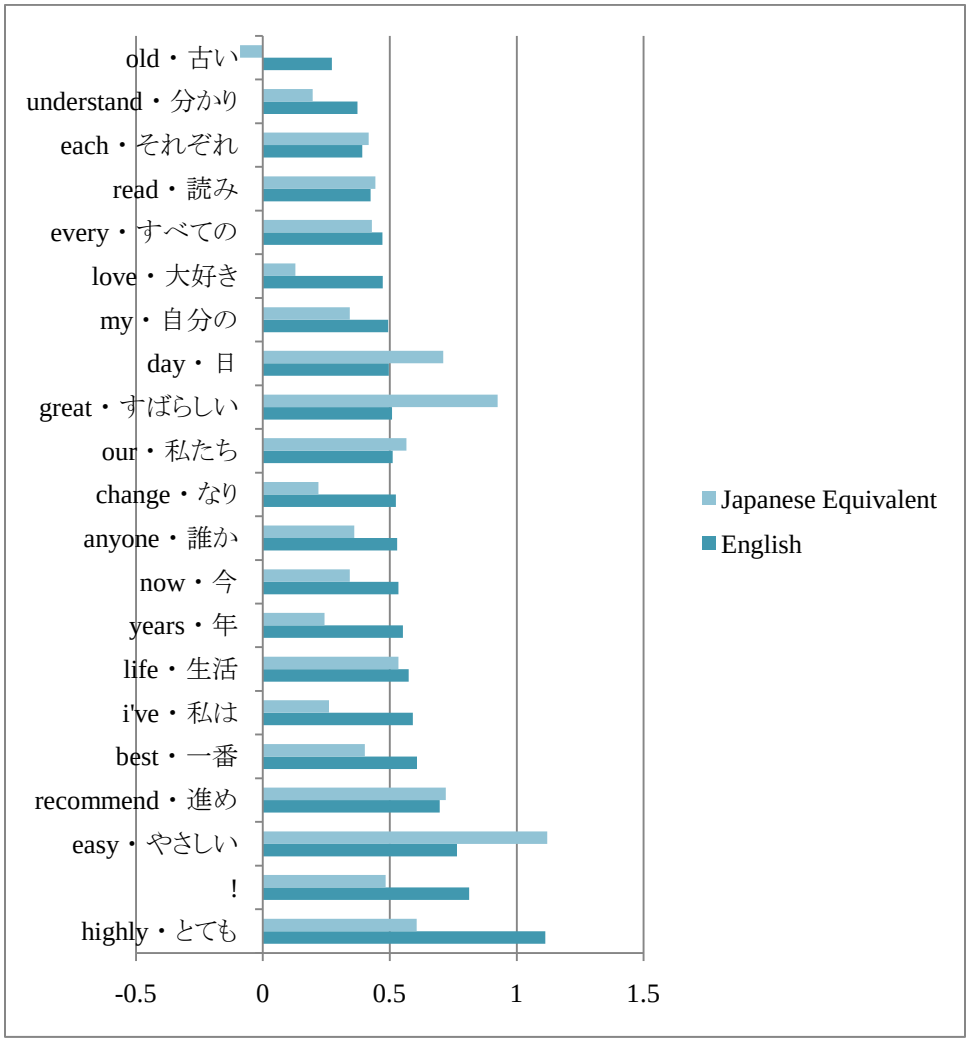


Figure 24: 5 Star English Unigrams to Japanese

5.2 Neutral English and Japanese Characteristics

In the English three-star list, just as in the Japanese list, we see an abundance of contrastive words: “however,” “but,” “though,” and

possibly “rather.” We see ranked highly words such as “part,” “some,” “little,” and “bit,” which correspond nicely to the top Japanese word, *chotto* (3 star unigram 1). “Part” may also map more specifically to *bubun* (3 star unigram 4). In addition, we have “seems,” which reasosnably maps to *you na* (3-star bigram 22). Furthermore, we have “more,” denoting an explicit comparison, which, in Japanese, is represented by *hou ga* (bigram 19), analyzed more thoroughly in Section 1, and which exactly maps to *motto* (unigram 5). Pang et. al.[6] refer to the so-called “thwarted expectations” phenomenon, occurring when “…the author sets up a deliberate contrast to earlier discussion,” which make initially make the appear to be positive. Pang et. al. were using polar English reviews only, however, and our data suggest that this phenomenon may be more so biased toward “neutral” reviews, or at least “not positive” reviews, rather than strictly negative ones. We see in Figure 25 that “however” follows essentially the same pattern as the Japanese equivalents.

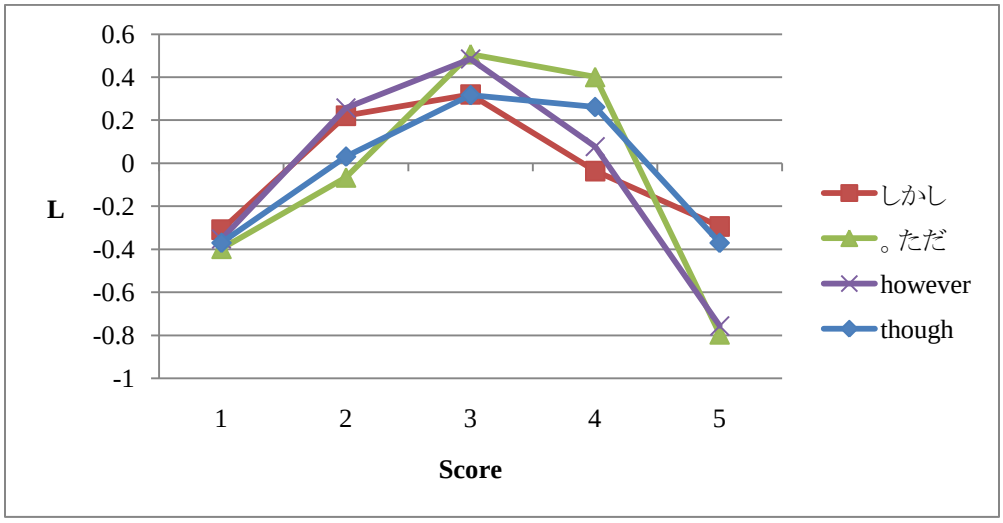


Figure 25: "However" words (*shikashi, .tada, however, though)*

We see that shikashi and tada follow pattern similar these English counterparts.

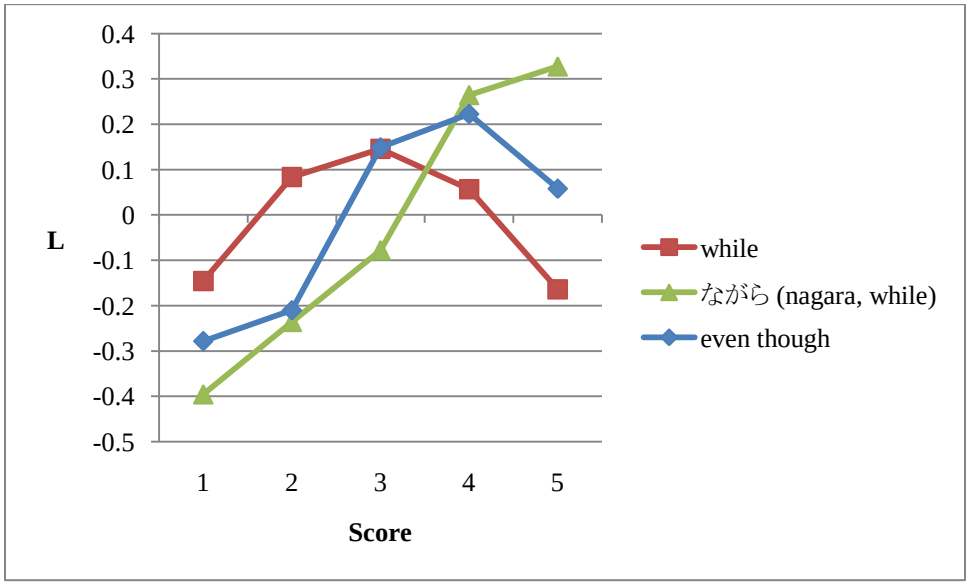


Figure 26: "While" words (while, nagara, even though)

"While" in English is used mostly in the middle ratings, whereas the Japanese equivalent, despite its similar denotation, appears to be used in positive contexts. While "even though" more closely mirrors nagara in Japanese, the correlation is not especially striking.

In Figure 26, we see that the English word "while" and the Japanese word, which carries both the temporal both its temporal and contrastive meaning, have very different behaviors. While the English "while" is symmetric, peaking in neutral usage, the Japanese *nagara* is used in much more likely to be used in positive contexts, perhaps being used in the capacity of expressing thwarted expectations espoused by Pang et. al.[6].

Rank	Unigram	Eng. L	JP Trans. L	Corss-ref.	JP Trans.
1	bit	0.499216	0.503936	U:3-1,3	ちょっと
2	however	0.484949	0.506913	B:3-1,2;U:1	ただ
3	interestin	0.4351	0.36937		面白い
4	seems	0.349665			ような
5	rather	0.345616			
6	lot	0.332592	0.018067		たくさん
7	though	0.316338	0.002393	de mo	
8	end	0.300615	-0.0035		終わり
9	novel	0.287711	0.320585		
10	character	0.277157			
11	too	0.265172			
12	characters	0.257624	0.352507		キャラク
13	some	0.252075	0.207521	See 1.	
14	but	0.25165	0.398303		だが
15	point	0.247261			
16	good	0.231606	0.045068		いい
17	writing	0.228192			
18	did	0.217508			
19	more	0.213055	0.157474	B:1-24	もっと
20	times	0.207749	-0.49718	U:5-6	度
21	didn't	0.20697	0.059653		なかった
22	little	0.202354		See 1.	
23	here	0.194172	-0.2761		ここ
24	her	0.178841			
25	part	0.176247	0.359338	U:3-4	部分
26	much	0.171841			
27	still	0.170402	0.017507		もう
28	really	0.166969	-0.4936	U:5-14	本当に
29	found	0.165536			

Figure 27: 3-Star English Unigrams to Japanese

Rank	Unigram	JP L	EN Trans L	Cross-ref.	EN Trans.
1	ちょっと	0.503936		U:3-1	a bit
2	収録	0.460574			
3	残念	0.382535	0.545536		unfortuna
4	部分	0.359278	0.525278	U:3-25	portion
5	もっと	0.355425	0.213055	U:3-19	more
6	あまり	0.350706	0.547752		not very
7	ただ	0.305675	0.484949	U:3-25	however
8	思う	0.256665	0.141307		think
9	いう	0.234479			
10	どう	0.229228			
11	性	0.222213			
12	かも	0.221344			
13	として	0.219044			
14	内容	0.211674	0.08667		contents
15	という	0.208395			
16	良い	0.207039			
17	感	0.203219	-0.10058		music
18	ので	0.202271	0.022559		because
19	なる	0.197835	0.008695		become
20	気	0.191915			
21	ところ	0.189139			
22	評価	0.182234			
23	言う	0.179041	0.081729		say
24	あっ	0.172965			
25	など	0.167577	0.100265		etc
26	ため	0.144638			
27	けど	0.143264		U:3-7,14	but
28	しれ	0.143016			
29	今回	0.141219	0.393176		this time
30	か	0.140793			

Figure 28: 3 Star Japanese Unigrams to English

Rank	Bigram	JP L	EN Trans.	Cross-ref	EN Trans.	
1	。ただ	0.506913	0.484949		however	
2	だが	0.392304	0.25165		but	
3	としては	0.38232				
4	かな	0.349256	0.222639		I wonder	
5	思う。	0.326834	0.141307		think	
6	というこ	0.32145				
7	なので	0.317984	0.022559		because	
8	かと	0.292042				
9	ないか	0.262028				
10	になる	0.253694	0.008695			
11	ので、	0.251035			See 7.	
12	と思う	0.246825			I think	
13	人は	0.238895	-0.26694		People are	
14	ので	0.228081			See 7.	
15	ことは	0.213601				
16	ば、	0.184727	0.055531		if	
17	ある。	0.173841				
18	のは	0.157603				
19	方が	0.157474				
20	かもしれ	0.153984			See 4.	
21	思います	0.152106			See 12.	
22	ような	0.151715				
23	のが	0.145819				
24	か、	0.143243				
25	的に	0.140804				
26	か。	0.137867				
27	はない	0.137006	-0.03129		is not	
28	には	0.132938				
29	だろう	0.127577	0.236158		probably	
30	し、	0.124416				

Figure 29: 3-Star Japanese

Bigram Tokens to English

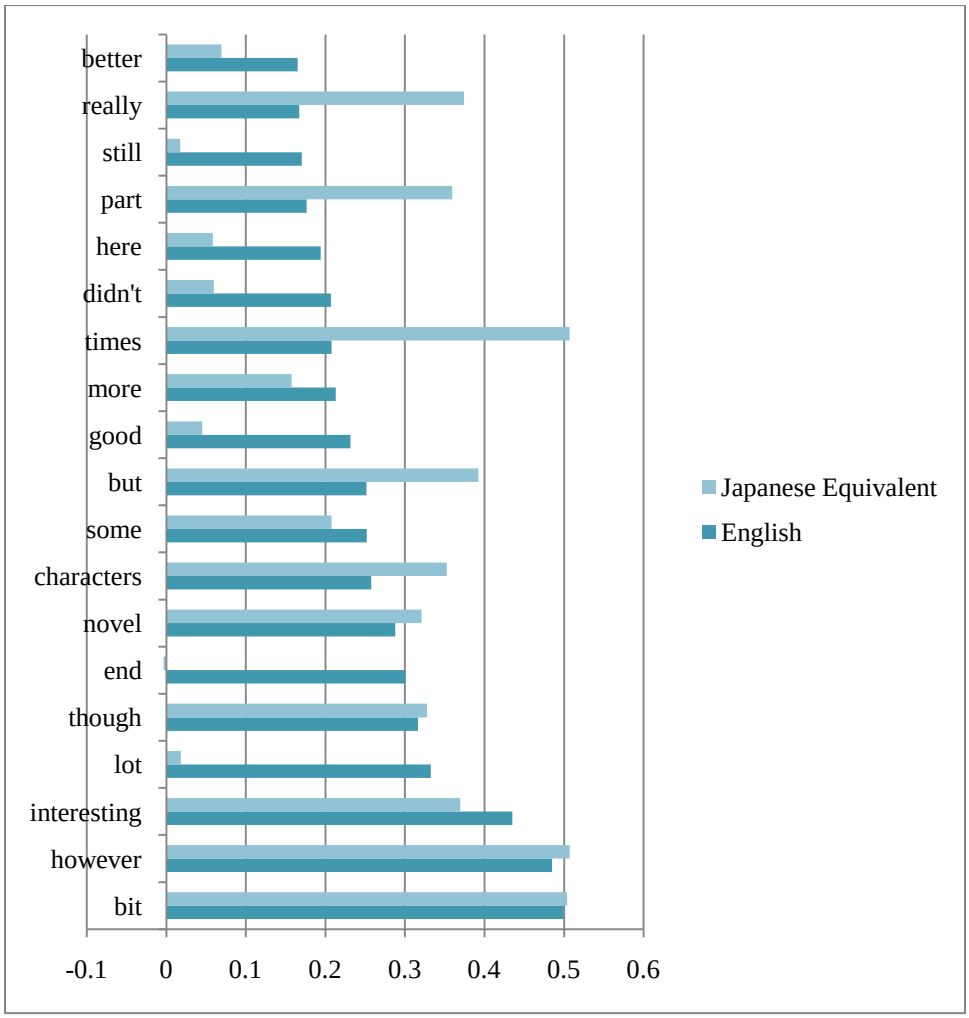


Figure 30: 3-Star English

Tokens to Japanese

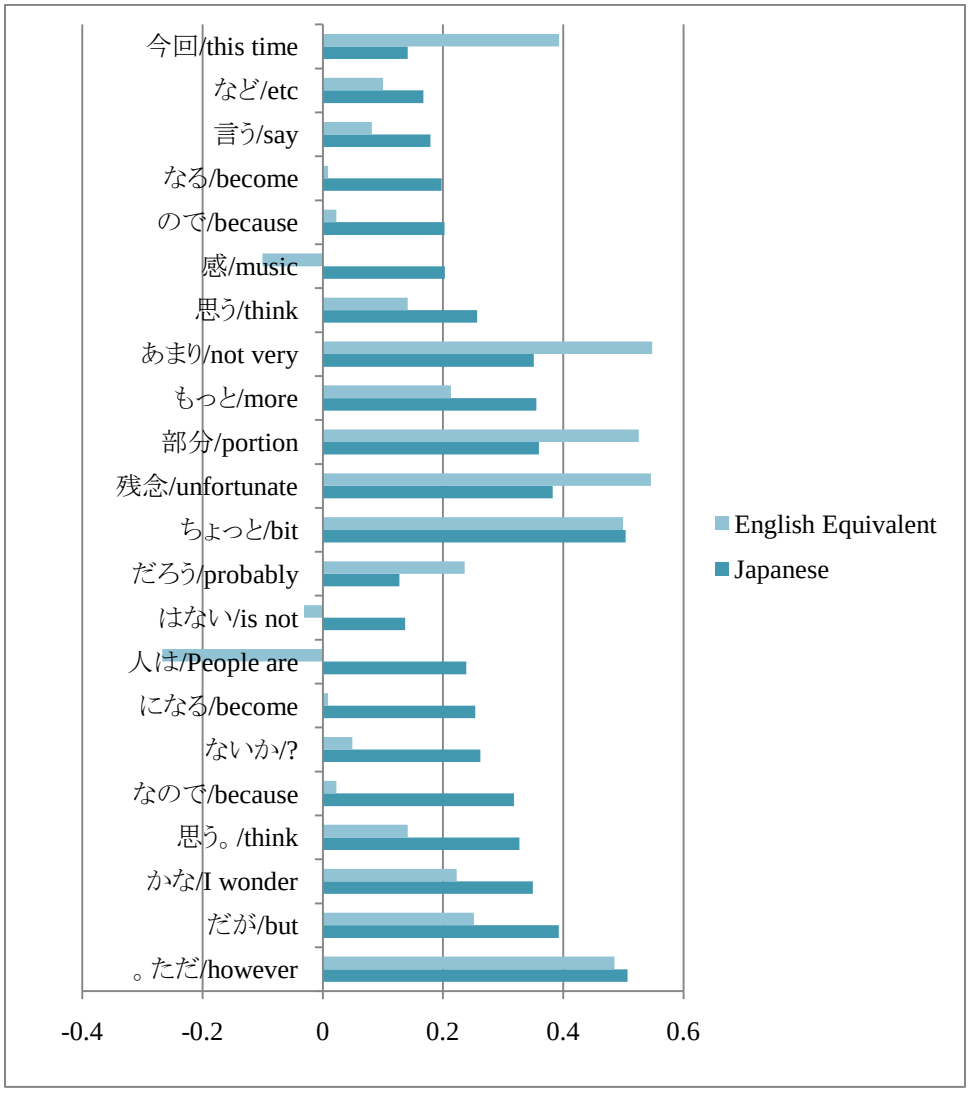


Figure 31: 3 Star Japanese Tokens to English

5.3 ENGLISH-JAPANESE COMPARISON SUMMARY

I have shown in this section that both English and Japanese share a propensity for tokens of negation in negative contexts, scope-limiting words and contrastive words in neutral contexts, as well as personal pronouns and intense words in positive contexts. While not all of the Japanese terms have English equivalents, those that have equivalents tend to be at least positively correlated with a specific class in both English in Japanese. We may conclude, then, that Japanese-speakers and English –speakers do indeed share some similarities in expression in the realm of sentiment.

6 A CLASSIFICATION EXPERIMENT

While the focus of this thesis is not the explicit tweaking of classification algorithms *per se*, I do believe it helpful to include some experiments with some standard algorithms. There does not appear to be a single published paper on the subject of Japanese sentiment analysis using the usual statistical algorithms on native Japanese text, without the aid of translation or deep syntactic parsing.

I manually collected 1,241 Japanese reviews from Amazon.com from various genres. To perform these experiments, I created a program to convert the text to *romaji* in order to circumvent character encoding issues. This has two side effects: it removes the problem of alternate spellings, and it increases the potential number of homophones with equivalent spellings. However, as most homophones in Japanese are of Chinese origin – words which are in general used in more formal speech – I do not believe that this will be a problem with experiments of this small scale, in this domain. The program also interfaces with MeCab in order to tokenize the text, thus providing, for our purposes, an interim solution to the problem of word segmentation. All experiments were performed with 10-fold cross-validation, with 80% of the data as training data. To simplify matters and to increase the amount of training data, reviews were divided into positive and negative polarity, where positive reviews have scores of greater than 3 and negative reviews have scores of less than three. Three-star reviews were excluded. I attempted two experiments: in the first, I simply used the Romanized tokens as they appeared in the raw text; in the second, when available, verb and adjectival forms were automatically reduced to

their dictionary forms, in order to remove differentiating morphologies. This did not make a substantial difference in the performance and appears to have slightly decreased performance. The maximum entropy classifier in the MALLET Machine Learning Toolkit was used.

Table 11: Classification Results

	Avg. F1(neg)	Avg. F1(pos)	Avg. Accuracy
Original Tokens	.751	.872	.832
Neutral Form	.737	.865	.826

While more rigorous testing and corpus selection is necessary to adequately assess these techniques, this shows that standard statistical machine learning algorithms work quite well with Japanese text for sentiment analysis without much added sophistication. These results fall short of Kanayama's, which were done on a different dataset.

7 CONCLUSION

I have shown that there are generalized usage trends of specific linguistic features in Japanese for this domain. Since the tokens analyzed here are general linguistic features, as opposed to domain-specific tokens, it is likely that they will generalize well to other domains. As many of the linguistic features are unique to Japanese, some insight has been gained into the nature of these features of the language, many of which would be completely lost in a translation to English. Many of these structures, such as particles and specific pronouns, may be reasonably interpreted to betray the psychological state or intent of the writer: the nuances of the language provide encode properties of the state of the writer of which he or she may not even be conscious.

In Section 1, I argued, counterintuitively, that *ne* and *no* perform similar functions of implicature and explicature embedding, which may explain their similar usage in the review classes. That *ne* would be used primarily in negative contexts is not at all obvious, showing that intuition regarding which terms are negative and which are positive can be insufficient. Some trends, such as the usage of negation in negative contexts and the usage of conjunctions in

neutral contexts, consists of tokens which dominate the tokens for certain classes. Ultimately, it became apparent that many of the most important tokens could not be translated into English, lending credence to the notion that native language analysis will prove fruitful for sentiment analysis.

English and Japanese, for all of their differences, do share some usage patterns, suggesting some universality of expression in certain contexts and the potential for language-independent analysis of sentiment based upon certain linguistic categories, such as pronoun usage, "extreme" word usage, or conjunction usage. While the languages themselves are quite different, what is being expressed, at a high level, is, at least sometimes, consistent across the languages. Both languages, I showed, use conjunctions and hedging words heavily in 3-star contexts; both use personal pronouns in positive contexts; and both use words of negation in negative contexts. That these features can be are correlated to specific sentiments is clear, and the potential exists for exploiting these cues for automated analysis.

APPENDIX A: UNEDITED JAPANESE TABLES

Unigram	1	2	3	4	5
,	-1.1733	-1.14132	0.657668	0.392116	0.33349
-	-0.09447	-1.24913	0.616314	0.042824	0.145921
点	-0.78652	-0.14236	0.514184	0.455444	-0.41229
ちょっと	-0.58941	-0.04441	0.503936	0.30832	-0.44021
収録	-0.06278	-0.69562	0.460574	0.120833	-0.03053
残念	0.260318	0.519288	0.382535	-0.46342	-1.45548
部分	-0.69765	0.266201	0.359278	0.212749	-0.40054
もっと	-0.03137	0.321098	0.355425	-0.23823	-0.59524
あまり	-0.62642	0.491215	0.350706	-0.06994	-0.42057
ただ	-0.11863	0.157909	0.305675	0.188935	-0.74061
思う	-0.03786	0.229921	0.256665	-0.14036	-0.39659
人は	-0.04778	-0.33463	0.240161	0.03138	0.058463
いう	-0.13119	0.106522	0.234479	0.050159	-0.31517
どう	-0.04433	0.119089	0.229228	-0.02305	-0.33484
性	-0.38246	0.195755	0.222213	0.039214	-0.15143
かも	-0.33619	0.275777	0.221344	0.154255	-0.45346
として	-0.04526	0.131483	0.219044	-0.05709	-0.29518
内容	-0.01472	0.127496	0.211674	-0.05053	-0.32416
という	-0.28877	0.018666	0.208395	0.125228	-0.10969
良い	-0.27396	0.069362	0.207039	0.062555	-0.10612
感	-0.61088	0.123809	0.203219	0.196314	-0.04954
ので	-0.4451	0.011966	0.202271	0.23361	-0.09205
なる	-0.42317	-0.01611	0.197835	0.112906	0.059258
気	-0.05374	0.07226	0.191915	0.075956	-0.33472
ところ	-0.39072	-0.06465	0.189139	0.248366	-0.0586
評価	0.406176	0.432177	0.182234	-0.55028	-0.89963
言う	0.209927	-0.15152	0.179041	-0.11683	-0.16228
あっ	-0.23251	0.158459	0.172965	0.001283	-0.13853
など	-0.35775	-0.14104	0.167577	0.218769	0.045334

Table 12: Unmodified Japanese 1-Star Unigrams, .06% Cutoff

Unigram	1	2	3	4	5
,	-1.1733	-1.14132	0.657668	0.392116	0.33349
-	-0.09447	-1.24913	0.616314	0.042824	0.145921
点	-0.78652	-0.14236	0.514184	0.455444	-0.41229
ちょっと	-0.58941	-0.04441	0.503936	0.30832	-0.44021
収録	-0.06278	-0.69562	0.460574	0.120833	-0.03053
残念	0.260318	0.519288	0.382535	-0.46342	-1.45548
部分	-0.69765	0.266201	0.359278	0.212749	-0.40054
もっと	-0.03137	0.321098	0.355425	-0.23823	-0.59524
あまり	-0.62642	0.491215	0.350706	-0.06994	-0.42057
ただ	-0.11863	0.157909	0.305675	0.188935	-0.74061
思う	-0.03786	0.229921	0.256665	-0.14036	-0.39659
3	-0.04778	-0.33463	0.240161	0.03138	0.058463
いう	-0.13119	0.106522	0.234479	0.050159	-0.31517
どう	-0.04433	0.119089	0.229228	-0.02305	-0.33484
性	-0.38246	0.195755	0.222213	0.039214	-0.15143
かも	-0.33619	0.275777	0.221344	0.154255	-0.45346
として	-0.04526	0.131483	0.219044	-0.05709	-0.29518
内容	-0.01472	0.127496	0.211674	-0.05053	-0.32416
という	-0.28877	0.018666	0.208395	0.125228	-0.10969
良い	-0.27396	0.069362	0.207039	0.062555	-0.10612
感	-0.61088	0.123809	0.203219	0.196314	-0.04954
ので	-0.4451	0.011966	0.202271	0.23361	-0.09205
なる	-0.42317	-0.01611	0.197835	0.112906	0.059258
気	-0.05374	0.07226	0.191915	0.075956	-0.33472
ところ	-0.39072	-0.06465	0.189139	0.248366	-0.0586
評価	0.406176	0.432177	0.182234	-0.55028	-0.89963
言う	0.209927	-0.15152	0.179041	-0.11683	-0.16228
あっ	-0.23251	0.158459	0.172965	0.001283	-0.13853
など	-0.35775	-0.14104	0.167577	0.218769	0.045334
DVD	0.170917	-0.07627	0.146918	-0.2491	-0.02813

Table 13: Unmodified Japanese 3-Star Unigrams, .06% Cutoff

Unigram	1	2	3	4	5
□	-1.13046	-2.54531	0.069304	0.436127	1.046677
！	0.025262	-0.57814	-0.50361	-0.1637	0.839447
心	-0.38889	-0.29355	-0.21802	-0.08071	0.737261
そして	-0.32712	-0.29203	-0.45234	0.085959	0.716928
□□	-0.39614	#NAME?	-0.59145	0.964644	0.715484
とても	-0.63121	-0.37048	-0.03253	0.157185	0.605945
くれ	-0.04271	-0.69408	-0.28662	0.207998	0.55008
度	-0.07918	-0.04256	-0.49718	-0.03943	0.507049
!	0.667302	-0.51124	-0.61954	-0.4723	0.4838
聴い	-0.01024	0.005802	-0.31579	-0.27807	0.477925
読み	-0.90352	-0.01509	-0.18651	0.331367	0.443507
とき	-0.35258	-0.2904	-0.16556	0.227376	0.438497
お	-0.30217	-0.2245	-0.09823	0.096415	0.426534
読ん	-0.6725	-0.12103	-0.02442	0.210737	0.409481
本	-0.23664	-0.33892	0.003884	0.085016	0.388759
中	-0.18815	-0.42901	-0.1077	0.224017	0.375586
本当に (h	0.367065	0.063555	-0.4936	-0.55561	0.374454
今	0.203959	-0.36161	-0.27411	-0.02	0.343071
,	-1.1733	-1.14132	0.657668	0.392116	0.33349
ながら	-0.39576	-0.23612	-0.07781	0.264499	0.3276
時	0.030414	-0.37593	-0.1473	0.092898	0.318536
自分	-0.3132	-0.16507	-0.16256	0.235106	0.314465
いく	-0.70187	-0.63752	0.152148	0.508721	0.308293
この	-0.0131	-0.10403	-0.14464	-0.04204	0.271619
できる	-0.68978	-0.28217	0.012303	0.454447	0.263004
い	-0.17836	-0.09781	-0.14703	0.119954	0.259855
私	-0.14831	-0.0371	-0.10517	0.002492	0.257847
年	0.315923	-0.064	-0.38053	-0.21932	0.242232
まし	-0.00861	-0.08177	-0.16976	-0.00653	0.238907
また	-0.19602	-0.02657	-0.02386	-0.00717	0.226394

Table 14: Unmodified Japanese 5-Star Unigrams, .06% Cutoff

Bigrams	1	2	3	4	5
のEXILE	2.365218	-0.6193	-0.90146	-2.24176	-1.88794
章の	1.724394	-0.10272	-0.55703	-1.53382	-1.44323
か?	1.596696	0.035093	-0.74058	-1.27737	-1.19649
ですか	1.565261	-0.20786	-0.4023	-1.13847	-1.21288
曲を	0.935427	0.103857	-0.33278	-0.82433	-0.40495
よ。	0.882879	0.114868	-0.56414	-0.63491	-0.25163
…。	0.877747	0.101869	0.035108	-0.51236	-1.1741
んだ	0.824407	-0.10195	-0.20967	-0.4555	-0.3695
なん	0.691585	-0.01325	-0.09039	-0.43363	-0.39817
んです	0.657586	-0.06548	-0.14595	-0.39078	-0.25371
のでしょ	0.652267	0.437352	-0.24753	-0.56273	-0.71858
の曲	0.650631	0.057279	-0.05582	-0.72238	-0.21895
じゃない	0.64607	0.167333	-0.10161	-0.48014	-0.50818
だから	0.58148	0.162981	-0.04428	-0.514	-0.42431
言って	0.538955	0.088258	-0.09268	-0.41478	-0.28532
んでし	0.513857	0.401748	-0.09447	-0.60985	-0.5294
ないです	0.49199	0.346519	-0.12872	-0.31881	-0.65842
か？	0.447833	0.334866	-0.07674	-0.30698	-0.63982
ない。	0.427833	0.501417	-0.05343	-0.49709	-0.73486
ん。	0.425502	0.214716	-0.07629	-0.19412	-0.53334
ません	0.405419	0.218597	-0.06437	-0.23344	-0.47228
うか	0.400115	0.119298	0.030125	-0.20593	-0.47436
でしょう	0.384815	0.112299	-0.04298	-0.18318	-0.36987
うと	0.351392	-0.1845	-0.02358	-0.08691	-0.10951
方が	0.339734	0.246206	0.157474	-0.25844	-0.70854
ですね	0.320687	-0.23334	-0.13472	-0.07484	0.066777
ね。	0.314235	-0.03114	0.012671	-0.2048	-0.13917
をし	0.30762	-0.05409	-0.07981	-0.14754	-0.06509
・・	0.280513	0.247256	0.118807	-0.31112	-0.4767
か。	0.264836	-0.07759	0.137867	-0.02768	-0.36563

Table 15: Unmodified 1-Star Japanese Bigrams, .06% cutoff

Bigram	1	2	3	4	5
。ただ	-0.39758	-0.06608	0.506913	0.400864	-0.79529
だが	-0.20794	0.039956	0.392304	0.082712	-0.42027
としては	-0.27003	0.284342	0.38232	0.022805	-0.62008
かな (ka	-0.34212	0.238635	0.349256	0.029267	-0.41373
思う。	-0.29797	0.266803	0.326834	-0.0613	-0.35161
というこ	-0.32365	-0.12788	0.32145	0.109088	-0.05014
なので	-0.60909	-0.01719	0.317984	0.301438	-0.16614
かと	-0.31091	0.291021	0.292042	-0.03802	-0.35035
ないか	-0.06839	0.236208	0.262028	-0.15852	-0.3547
になる	-0.40037	-0.06353	0.253694	0.127179	0.010478
ので、 (	-0.52178	-0.02328	0.251035	0.230293	-0.05132
と思う	0.001396	0.255491	0.246825	-0.18839	-0.41416
人は	0.024169	-0.10198	0.238895	-0.09741	-0.08932
ので	-0.2725	0.261351	0.228081	-0.02284	-0.27478
ことは	-0.29766	0.121598	0.213601	0.03739	-0.12436
ば、(if)	-0.08634	-0.01751	0.184727	0.057639	-0.15969
ある。	-0.02946	-0.16957	0.173841	-3.85E-04	0.007646
のは	0.162601	0.116321	0.157603	-0.1571	-0.3405
方が	0.339734	0.246206	0.157474	-0.25844	-0.70854
かもしれ	-0.2631	0.31058	0.153984	0.147248	-0.47879
思います	-0.20897	-0.1594	0.152106	0.153431	0.028001
ような	-0.09759	0.183516	0.151715	-0.09227	-0.17715
のが	-0.33464	0.250868	0.145819	0.033896	-0.16253
か、	0.028772	0.020333	0.143243	0.042273	-0.26208
的に	-0.34283	0.051566	0.140804	0.226374	-0.13849
か。	0.264836	-0.07759	0.137867	-0.02768	-0.36563
はない	-0.10254	0.238821	0.137006	-0.03893	-0.28475
には	-0.12676	0.163543	0.132938	0.049271	-0.25756
だろう	0.137485	0.221277	0.127577	-0.191	-0.37232
し、	-0.22156	0.177433	0.124416	0.03258	-0.14848

Table 16:Unmodified Japanese 3-Star Bigraam, .06% Cutoff

Bigram	1	2	3	4	5
！！	0.127095	-0.79144	-0.77529	-0.35273	1.08168
この本	-0.31693	-0.36328	-0.195	0.01934	0.651441
本を	-0.42275	-0.36399	-0.00977	0.045515	0.564636
てくれ	0.004122	-0.66936	-0.3256	0.179714	0.551501
の中	-0.32337	-0.55405	-0.11947	0.225993	0.543773
います	-0.42225	-0.32094	-0.32696	0.318663	0.524992
読んで	-0.74612	-0.13922	-0.09047	0.2115	0.507646
ことが	-0.58593	-0.26463	-0.10876	0.309313	0.439968
いまし	0.103168	-0.21599	-0.29746	-0.10596	0.417583
になり	-0.26416	-0.11005	-0.13901	0.056117	0.382356
にも	-0.13185	-0.20228	-0.11904	0.048652	0.346303
自分の	-0.2712	-0.11263	-0.38725	0.297659	0.342082
。この	-0.04971	-0.15326	-0.1415	-0.027	0.325743
見て	-0.11497	-0.07859	-0.15996	0.006728	0.305384
てい	-0.28242	-0.11782	-0.11795	0.171579	0.281972
いて	-0.63539	-0.20893	0.046072	0.331965	0.279866
思って	-0.19145	0.031947	-0.12428	-0.02073	0.267131
私は	-0.13589	-0.03948	-0.11487	-2.07E-04	0.2602
ます。	-0.31585	-0.15353	-0.05079	0.194114	0.257522
ました	-0.01168	-0.08622	-0.17437	-0.00297	0.245771
てみ	-0.24328	-0.01301	-0.09814	0.078273	0.236942
。	-0.59159	0.136657	-0.08374	0.173797	0.230849
、この	0.072861	-0.12338	-0.17863	-0.03284	0.230108
。「	0.157987	-0.33339	-0.09364	-0.01702	0.227477
なった	-0.08159	-0.25154	0.029118	0.053424	0.214837
いる。	-0.4325	-0.06439	0.02919	0.180486	0.208059
てき	0.05534	0.05582	-0.23895	-0.11205	0.20468
」の	-0.39964	0.104895	-0.00236	0.033788	0.200609
ことを	-0.01141	-0.03109	-0.20535	0.030957	0.192587
と思っ	-0.08257	-0.09562	-0.09156	0.064486	0.186303

Table 17: Unmodified Japanese 5-Star Bigrams, .06% Cutoff

APPENDIX B: A BRIEF OVERVIEW OF BASIC JAPANESE

The following sections are intended as a brief, high-level overview of some of the very basics of the Japanese language. The intent is that it give any potential readers who are not familiar with Japanese at least a sense of the language and some of the constructs which are addressed in this thesis.

Writing Systems

Modern Japanese consists of essentially four alphabets. The first is the alphabet, borrowed from the west, which is often interspersed with native Japanese characters. These characters are widely used, though, aside from individual letters, the typical Japanese person would not necessarily be able to pronounce them well. Japanese also consists of thousands of complex, angular characters, called *kanji*. These characters were originally introduced to the Japanese by the Chinese. As such, *kanji* function in Japanese much in the same way that Greek and Latin roots function in English. They encode meaning as well as pronunciation. However, due to various historical factors, *kanji* often have multiple pronunciations

depending on the context. Further complicating matters is the fact that Chinese is tonal and Japanese is not. As a result, *kanji* pronunciations necessarily have no tones. This is compounded by the fact that Japanese already has a limited number of phonemes. The result is an unusually large number of homophones. Due to this, many complex *kanji* words are only used in writing, as the meaning encoded by the character allows such characters to be understood.

Japanese consists of two phonetic writing systems, as well, which function in much the same way that the alphabet does. The two systems of writing, *hiragana* and *katakana*, have a one-to-one mapping in pronunciation, but they do not encode meaning. Theoretically, all Japanese could be written in *hiragana* or *katakana*, but this is not done. *Hiragana* is a cursive script, and it is the standard for phonetic writing. It typically is used for words which do not have *kanji* (or which have difficult *kanji*), by schoolchildren who have not yet mastered *kanji*, by foreigners who do not have time to learn *kanji*, and, most importantly, for the morphology of verbs and adjectives in Japanese. *Katakana* is used for loan words – of which there are plenty in Japanese, perhaps more than in most languages – and for emphasis and onomatopoeia.

TRANSLITERATION

As mentioned, Japanese has a limited number of phonemes. In fact, pronunciation adjustments notwithstanding, every Japanese sound has a corresponding English sound, though the reverse is certainly not true. Thus, transliteration of Japanese to English is trivial when the pronunciation is known. However, the chief difficulty in reading Japanese is that one must know *kanji*. Since *kanji* do not have consistent pronunciations, one must simply memorize them in order to read Japanese.

From the standpoint of a computer, Japanese presents a unique set of problems, due to its multiple syllabaries, lack of space delineation, inconsistent pronunciation, and multiple possible spellings. Consider the following: ほう, ホウ, 方 are all pronounced *hou* – the first in *hiragana*, the second in *katakana*, and the third with a *kanji* and this is just a small sample of *hou* words. There are dozens of kanji with this pronunciation.

PRONUNCIATION

It is not difficult to pronounce Japanese and be understood compared with many other languages. As mentioned earlier, there are no tones. Japanese is, in fact, often described as particularly flat. It consists of five vowels: *a, i, u, e, o*, pronounced, roughly "ah-ee-oo-eh-oh," as in Italian or Spanish In addition, there is *n*, which is pronounced exactly as *n* in English, although perhaps slightly more verbalized in some contexts. Every other sound in Japanese consists of a consonant attached to one of the vowels: Thus, we have *ka, ki, ku, ke ko*; *na ni, nu, ne, no*; etc. There are also dipthongs, yielding sounds such as *kya*, *kyo*, etc. Often, English speakers mistakenly pronounce sounds such as *kyo* as "kee-o." For example, "Kyoto" is often pronounced, "kee-o-to" by English-speakers, but *y* is pronounced as merely a movement of the mouth and not as a long *e.* In Japanese, *i* is pronounced as a long e, and *e* is pronounced as short e. Moreover, in Japanese, there are no isolated consonants, with the exception of *n*.

The following examples illustrate correct pronunciation.

sake (rice wine) - "sah-keh"

saki (ahead) – "sah-kee"

ima (now) – "ee-mah"

Japanese has a few other quirks generally unfamiliar to English speakers: First, Japanese has long vowels. That is to say, Japanese has vowels which may be pronounced twice as long as other vowels, and they are distinct from their short counterparts. The long vowel *aa* ("aah") differs from *a* in that *aa* is pronounced twice as long. Additionally, the long vowel *ou* is pronounced as an elongated long *o*. That is, it is not "o-u" but "ohh." Japanese also has pauses. For example, the sound *ita* differs from *itta.* While *ita* is pronounced exactly as it appears, *itta* is pronounced *it-ta*, with a pause between the two syllables. (In general, *-tta* indicates past tense). This may be done with almost any consonant. The word *isshou* is pronounced *ish-shohh,* and the loan word *toppu* ("top") is pronounced *top-pu*. There are not actually two constants; there is only a rather rigid pause before completing it. Recall that *pu* is a single unit. There is no isolated *p* sound in Japanese. The sound *ppu* sounds like a very forceful *pu* – so forceful that there is a pause before it.

Finally, one aspect of Japanese that often confuses English speakers is the tendency of Japanese speakers tounconsciously drop vowels in conversation (but not in spelling, as this would be impossible). For example, *sukoshi* is often pronounced (to English

ears) *skosh.* The sound *masu* is usually pronounced *mas*, and *desu* is typically pronounced *des.*

BRIEF GRAMMAR OVERVIEW

Japanese is a subject-object-verb (SOV) language. Additionally, in Japanese there is a rather important distinction between the subject of a sentence and the topic, but that is beyond the scope of this overview. Parts of the sentence are identified by *particles*, undulations which are not considered to be words, but which are nevertheless necessary in conversation. Particles are almost always postpositions, and some end sentences.

GRAMMAR : EQUIVALENCE

In Japanese, an optional ending, *desu*, is often added to the end of a sentence. This is a copula that denotes equivalence. Its informal counterpart is *da*. The particle *wa* marks the topic of the sentence,.

Watashi = "I"

Ex: *watashi wa jon desu.* "I am John." (formal)

I [topic marker] jon [equivalency marker]

The informal equivalent of *desu* is *da*. Neither is absolutely necessary.

In general, the Japanese avoid pronouns, and especially personal pronouns. Rather than *watashi wa jon desu*, one is more likely to say *jon desu*. Subjects and topics are often not explicitly stated, but are instead gleaned from context.

POLITENESS LEVELS

Japanese has several level of politeness which depend on one's standing in respect to those around him. The most obvious way that this manifests itself is in the form of verb endings and copulas. Verbs typically end in *u* or *ru,* e.g., *aru* ("to exist"). However, a more humble way of saying that something exists is ar**imasu**. This is often called the *masu* ending. Any verb can be used politely by converting it to a *masu* form. In addition, *desu* may be appended to the past tense and plan negative forms of verbs to make them more formal.

At times, honorific prefixes are appended to nouns or verbs to indicate an extra level of formality of humbleness. The two of note are *o* and *go*. For example, *sake* means "rice wine," but it is often pronounced as *osake,* which is more formal.

ADJECTIVES

There are two kinds of adjectives in Japanese: *na* adjectives and *i-adjectives. Na* adjectives end in *na, and i*-adjectives end in *i*. Consider, *benri na* and *omoshiro**i***

TENSES

In general, -*tta* indicates informal past tense. For example, *omoshiroi* means "interesting," whereas *omoshiro**katta*** means "was interesting." The *desu* copula may be appended to either to make them more formal. Past tense in the *masu* form is indicated by a *shita* ending. Thus, *omoimasu* ("think") becomes *omoima**shita*** ("thought"). The past tense of *desu* is *de**shita***. There is no explicit future tense in Japanese. The present tense is used to refer to the future.

NEGATION

In general, *nai* (present tense) or *na**katta*** (past tense) indicates a negation. *Omosiroi* ("interesting") becomes *omoshiroku* ***nai*** ("not interesting"). Likewise, *jon desu* ("I am John.") becomes *jon ja* ***nai***. ("I'm not John.") In formal speech, *masu* becomes ma***sen*** to indicate negation. Thus, *omoimasu* would become *omoima**sen***. This may be combined with the past tense to become *omoimasen* ***deshita***.

QUESTIONS

In formal Japanese speech, questions often end in *ka*. Whereas *jon desu*, means "[I am/He is]" John, *jon desu ka* means "[Is he/Are you] John?" This may be phrased in a negative form, as well.

BIBLIOGRAPHY

1. Sifry, D., *The State of the Live Web, April 2007. Technorati, Inc*. 2007, April.

2. Ahmed Abbasi, H.C., Arab Salem, *Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums.* ACM, 2007. **26**(3).

3. Hiroshi Kanayama, T.N., *Fully automatic lexicon expansion for domain-oriented sentiment analysis*, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*. 2006. p. 355–363.

4. Christopher Potts, F.S., *Exclamatives and heightened emotion: Extracting pragmatic generalizations from large corpora*. 2008: Ms., UMass Amherst.

5. Noah Constant, C.D., Christopher Potts, Florian Schwarz, *The pragmatics of expressive content: Evidence from large corpora.* Sprache und Datenverarbeitung, 2008.

6. Pang, B., L. Lee, and S. Vaithyanathan. *Thumbs up?: sentiment classification using machine learning techniques*. 2002: Association for Computational Linguistics Morristown, NJ, USA.

7. Turney, P. *Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews*. 2002.

8. Mihalcea, R., C. Banea, and J. Wiebe. *Learning multilingual subjective language via cross-lingual projections*. 2007.

9. Bautin, M., L. Vijayarenu, and S. Skiena. *International sentiment analysis for news and blogs*. 2008.

10. Hiroshi Kanayama, T.N., Hideo Watanabe, *Deeper Sentiment Analysis Using Machine Translation Technology*, in *Proceedings of the 20th International Conference on Computational Lingustics*. 2004, Association for Computational Linguistics: Morrison, NJ. p. 494-500.

11. Christopher Potts, S.K., *The performative nature of Japanese honorifics*, in *Proceedings of Semantics and Linguistic Theory 14*, K.W.a.R.B. Young, Editor. 2004, Cornell University Linguistics Department: CLC Publications.

12. TSUJIMURA, N., *Degree words and scalar structure in Japanese.* Lingua, 2001. **111**(1): p. 29-52.

13. Nakanishi, K., *Even, Only, and Negative Polarity in Japanese*, in *Proceedings of SALT XVI*, J.H. Gibson, Editor. 2006, CLC Publicaitons: Cornell University, Ithaca, NY. p. 138-155.

14. Takatsu, T., *A Unified Semantic Analysis of the NO DA Construction in Japanese.* The Journal of the Association of Teachers of Japanese, 1991. **25**(2): p. 167-176.

15. Kunishige, T., *On How Speaker's and Participant's Viewpoints Function in Japanese.*

16. Bolinger, D., *Degree words*. 1972: Mouton.

17. Jorden, E. and M. Noda, *Japanese: The spoken language*. 1988: Yale University Press.

18. Alfonso, A. and J. Daigaku, *Japanese language patterns: a structural approach*. 1966: Sophia University LL Center of Applied Linguistics.

19. Kuno, S., *The structure of the Japanese language*. 1973: MIT press.

20. McGloin, N., *Some observations concerning no desu expressions.* The Journal of the Association of Teachers of Japanese, 1980. **15**(2): p. 117-149.

21. Halliday, M. and R. Hasan, *Cohesion in english*. 1976: Longman London.

22. Cook, H., *The sentence-final particle ne as a tool for cooperation in Japanese conversation.* Japanese/Korean Linguistics, 1990. **1**: p. 29-44.

23. Itani, R., *Japanese Sentence-Final Particle NE: A Relevance-Theoretic Approach'.* Working Papers in Linguistics, 1992. **4**: p. 215-237.

24. Matsui, T., *Linguistic encoding of the guarantee of relevance: Japanese sentence-final particle yo.* Pragmatic markers and propositional attitude, 2000: p. 145–172.

25. Davis, C., *Contexts, Decisions, and the Japanese Particle yo.*

26. Hasegawa, Y. and Y. Hirose, *What the Japanese language tells us about the alleged Japanese relational self.* Australian Journal of Linguistics, 2005. **25**(2): p. 219-251.

27. Su, N., et al. *A Bosom Buddy Afar Brings a Distant Land Near: Are Bloggers a Global Community?* 2005: Springer.

28. Ono, T. and S. Thompson, *Japanese (w) atashi/ore/boku I: Theyre not just pronouns.* Cognitive Linguistics, 2003. **14**(4): p. 321-347.

29. Suzuki, S., *Emotive communication in Japanese: An introduction.* Emotive Communication in Japanese, 2006: p. 1-13.

30. Kuwayama, T., *The reference other orientation.* Japanese sense of self, 1992: p. 121-151.

31. Suzuki, S., *Tte and nante: Markers of psychological distance in Japanese conversation.* Journal of pragmatics, 1998. **29**(4): p. 429-462.