Cluster Analysis of Role Languages in Visual Novel Game *AIR*

Ruixuan Tu
ruixuan.tu@wisc.edu
University of Wisconsin–Madison
7 December 2024

Contents

1 Introduction

As defined by Teshigawara and Kinsui (2011), role language “yakuwarigo” is defined on individual characters in fictional media, which are “sets of spoken language features (e.g. vocabulary and grammar) and phonetic characteristics (e.g. intonation and accent patterns), associated with particular character types.” However, this definition is conceptual and lacks direct quantitative evidence. As investigated by Ma (2019), there is a similar concept of keyword “特徴語” defined on a group of speakers as significant (CoS >2) or minor (CoS <0.5) words. According to NINJAL (1997), Coefficient of Specialization (CoS) of Word W in Cluster C “特化係数” is defined as: $\mathrm{CoS}(W, C) = \frac{\text{freq of } W \text{ in cluster } C}{\text{number of words in cluster } C} / \frac{\text{freq of } W \text{ in whole}}{\text{number of words in whole}}$ (proportion in cluster compared with the proportion in whole; >2 means at least as twice frequent in cluster as the occurrences in whole).

This work further investigates the research question that whether keyword implies yakuwarigo (is a keyword also a yakuwarigo?) in a specific visual novel game *AIR*. *AIR* (Figure 1, [official homepage]) is an adventure game developed by Key/Visual Arts and released in 2000, known for its emotional storytelling. There are three volumes with different settings and main characters: Dream (Yukito and Misuzu), Summer (Kanna, Ryuuya, and Uraha), and Air (Misuzu without Yukito), and the script is mainly written by Jun Maeda and Takashi Ishikawa along with four scenario assistants.

2 Methods

We only select speakers with over 100 lines (Figure 2) of dialogue (not considering thoughts and narration) for statistical value; otherwise, the analysis is on insignificant characteristics. We then split and classify parts of speech of the tokens by morpheme analyzer MeCab through fugashi (McCann 2020).

We focus on the parts of speech that usually correlated to yakuwarigo: Pron (代名), Aux Verb

Figure 1: *AIR* Gameplay: conversation in dialogue box [src]

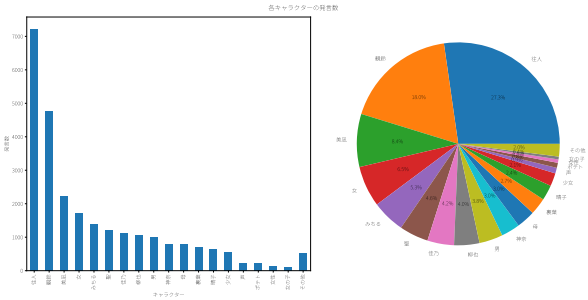

Figure 2: Dialogue Sizes: we only use characters with over 100 lines, excluding 2% of lines.

	声	往人	みちる	神奈	裏葉
えいっ/感動詞/一般	14.9	0.5	9.0	9.4	6.5
た/助動詞/*	551.4	451.1	355.0	416.5	348.8
に/助詞/格助詞	313.0	355.2	750.4	410.3	668.2
は/助詞/係助詞	283.2	502.2	413.4	858.1	700.8
はっ/感動詞/一般	14.9	1.9	0.0	0.0	3.3

Table 1: A Part of Adjusted Frequency Matrix

(助動), Case Part (格助), Adv Part (副助), Conj Part (接続), Final Part (終助), Noun Part (準体), and Interj (感動). We then calculate the adjusted frequency for word W and speaker S as $T(S,W) = 10000 \times \frac{\text{freq of } W \text{ in speaker } S}{\text{number of words in speaker } S}$ (frequency of W per 10000 words for speaker S). Combining all speakers and words, we formulate the adjusted frequency matrix as demonstrated in Table 1.

Take the adjusted frequency matrix, we input every column $T(S) = (T(S,W_1),\ldots,T(S,W_n))$ as a vector of frequency of words for speaker S into the clustering algorithm. For example, for Yukito we have the vector $T(\text{往人}) = (0.5, 451.1, 355.2, 502.2, 1.9, \ldots)$ as in the matrix.

We use agglomerative hierarchical clustering with Ward's method (Ward Jr. 1963) and Euclidean distance to cluster speaker vectors $T(S)$. Technically, we want to split the clusters that minimizes variance from cluster centeroids. For clusters A and B, the Ward distance is $d(A,B) = \frac{|A|\cdot|B|}{|A\cup B|}\|\mu_A - \mu_B\|_2^2$, which is also the increase of sum of squares ($SS_A = \sum_{x_i\in A}(x_i-\mu_A)^2$) when we merge clusters A and B, which could be linked to the minimization of variance $s_A^2 = \frac{SS_A}{|A|-1}$. For two vectors $\mathbf{x}$ and $\mathbf{y}$, the Euclidean distance is $d(\mathbf{x},\mathbf{y})^2 = \|\mathbf{x}-\mathbf{y}\|_2^2 = \sum_{i=1}^n (x_i-y_i)^2$.

From the clustering algorithm, we obtain the dendrogram (Figure 3) and the cluster assignments for each speaker. We then calculate the CoS for each word in each cluster and spot the significant words (CoS >2) as yakuwarigo candidates, as well as the minor words (CoS <0.5). Since

there might be a lot candidates, we only list at most five ranked by high frequency in the tables. The full tables are Table 11 and Table 12 in the Appendix.

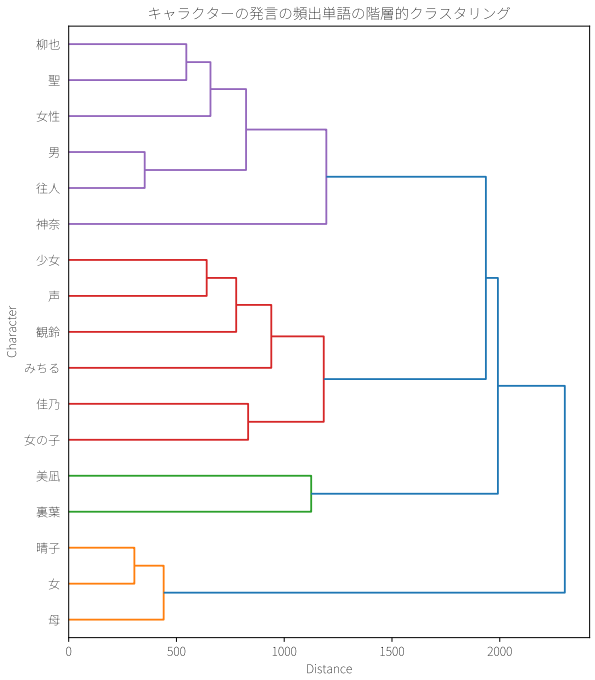

Figure 3: Dendrogram of Hierarchical Clustering

3 Analysis

We have four clusters in the dendrogram (Figure 3), splitting speakers as below. We then analyze a few significant words in each cluster to confirm the classification and discover potential yakuwarigo, on randomly extracted lines as tables.

Cluster 0: Ryuuya 柳也, Hijiri 聖, Josei 女性, Otoko 男, Yukito 往人, and Kanna 神奈

Cluster 1: Shoujo 少女, Koe (speech) 声, Misuzu 観鈴, Michiru みちる, Kano 佳乃, and Onna-no-ko (girl) 女の子

Cluster 2: Minagi 美凪 and Uraha 裏葉

Cluster 3: Haruko 晴子, Onna (woman) 女, and Haha (mother) 母

By unreliable Internet sources [src1], [src2], and [src3], Misuzu is written by Jun Maeda; Kano is written by Takashi Ishikawa; Minagi is written by Jun Maeda or Takashi Ishikawa (first half), Kai Unryuji and Toya Okano (second half); the whole Summer Volume (Kanna, Ryuuya, and Uraha) is written by Yuuichi Suzumoto; and the whole AIR Volume is written by Jun Maeda.

3.1 Non-female Language

In Cluster 0, we have all male characters (Ryuuya, Otoko, and hero Yukito) and mostly typical male languages: 俺, おまえ, あいつ, だろ(う), てろ, ぞ, and ぜ. However, we notice there are words that are not typical male languages such as よ (Pron) and うむ (Interj). From their identities, we see that they can use male language with additional words.

We found that 余 is only used by Kanna: lived 1000 years ago, goddess in a shrine, last winged beings 翼人, calm speaking; met Ryuuya and Uraha to escape and meet her mother. 余 is a first-person pronoun since Heian period, approximately 1000 years ago, for male language and arrogant speech. The usage of 余 by Kanna emphasizes the time setting and the high hierarchy of the character (Table 3).

うむ is mainly used by Hijiri and Kanna as majority (Table 2). Hijiri is the elder sister (not a student) of Kano, running Kirishima clinic, and carrying a lancet. うむ always appears at the beginning of the sentence, indicating agreement, understanding, and consideration. This usage aligns with the stereotype of a doctor who thinks carefully as Hijiri, and the distinct calm identity of Kanna.

speaker	count
神奈	18
聖	17
往人	5
女性	1

Table 2: Cluster 0: count of lines with Interj うむ per speaker in cluster

3.2 Casual Female Language

Identifiable characters in Cluster 1 are heroine Misuzu, Michiru, and Kano. All characters in this cluster are female.

By Internet source [src], Misuzu is a second-year high school student; Michiru is a elementary school student; Kano is a first-year high school student. They all use usual schoolgirl conversation such as よぉ, うん, ねー, and ちゃう. However, there are in-group preferences even with some shared occurrences.

According to Table 6, we see that ちゃ and うん are shared among all characters with no clear preference, since Misuzu has many more total lines than others (Figure 2). On the other hand, あたし, よぉ, and ねぇ are favored by Kano and わたし, ねー, and ね (also slightly favored by Michiru) are favored by Misuzu. As in the first example in Table 5, Kano says “あたしお金持ってるから、オゴりだよぉ” (I have money, so I will treat you), which uses both preferred words. This sentence also uses script variation of オゴり to show casuality, similar to the use of katakana role language in playful situations as Example (3) described by Dahlberg-Dodd (2022). We can also see preference of ね and ねー from the remaining examples in Table 5.

From the unreliable Internet sources mentioned at the beginning of this section, we see this cluster is by the main writers without assistants, while Cluster 2 is written by mainly assistants. This difference might explain the split of common female language into two clusters through distinct groups of writers.

3.3 Formal and Polite Language

In Cluster 2, we only have female characters Minagi and Uraha. They use formal and polite language (specifically, desu/masu form) in common. However, they are not speaking in the same way despite the same form: Uraha excessively uses Keigo to emphasize politeness, and Minagi mostly uses desu/masu form to show formality.

We can clearly identify this difference by comparing sentences in Table 7. “神奈さまの御為なら、この命ささげようとも惜しくはございません” (I will not regret giving my life for the sake of Kanna-sama, by Uraha), and “実は私…こういうものを持ってます” (Actually, I have this kind of thing, by Minagi) and “あまりに楽しいので…思わずこれを進呈” (It’s so fun that I can’t help but present this, by Minagi).

Minagi is a second-year high school student and the leader of astronomy club with a calm/mild speech style. Uraha is the female court lady of Kanna, also the ancestor of Misuzu, leading the development of the story. The difference in speech style is due to the different roles and settings of the characters. Minagi is speaking with her friends in ordinary school setting, but her usage of Keigo emphasizes her reserved style with reluctance in speech by simple distancing (Obana 2019). Uraha is speaking with the goddess, giving difference in hierarchy of Keigo. Usage of Keigo towards gods is also described as the origin of Keigo (Obana 2019), justifying the excessive use.

3.4 Dialectal Language

In Cluster 3, we only have female characters Haruko, Onna (woman), Haha (mother). Haruko is the aunt of Misuzu and the single guardian of her after the death of her mother, addicted to

alcohol. Haha is the mother gave birth to Misuzu. They all use dialectal language from the same region.

From the examples in Table 8 and significant words in Table 11, we see that they mainly use Osaka dialect: せや, なんや, ねん, へん. However, there are still words like あんた that are not typically used by female Osaka dialect speakers. In this way, the authors might intentionally build a modified Osaka dialect to show the outgoing and humorous characteristics as in the stereotype of Osaka people without offending true Osaka residents. This dialectal role language is similar to the adaption of Osaka dialect for the funny Maebashi father in the movie *Like Father, Like Son* (SturtzSreetharan 2017). Another reason for the authors to adapt Osaka dialect should be the location of the office in Osaka (Figure 4), so it is easy for them to understand and produce Osaka dialect.

Figure 4: Visual Arts Co., Ltd., the Osaka producer of the game *AIR*, was located at 大阪府大阪市北区本庄西2-12-16 VA第一ビル before 2024, and migrated to 大阪府大阪市浪速区難波中2丁目10番70号パークスタワー17階 in June 2024. Photograph by Ruixuan Tu on 15 June 2023.

4 Conclusion

Through our analysis of the visual novel game *AIR*, most keywords “特徴語” from our method could be recognized as “yakuwarigo” that represents characteristics of specific individuals or groups, but might not the reverse side (not all “yakuwarigo” are keywords that could be found). From our method, we have observed non-female language, casual female language, formal and polite female language, and dialectal language as clusters. We also found that different groups of script authors might affect extracted keywords.

Limitations

In this work, we only analyze the significant group (CoS >2), but we missed the minor group (CoS <0.5), which might be also treated as “negative yakuwarigo” that is the yakuwarigo of all other groups. Furthermore, neither our method exhaustively discover all yakuwarigo, and we did not carefully analyze all possible keywords.

Since we use computational methods to analyze the text into morphemes, there might be misclassification/missplit of morphemes and their parts of speech, and the issue that the morpheme-level granularity might also hinder us to discover longer phrases as yakuwarigo. For example, the phrase ふんだ is wrongly split into [ふん (Pron)][だ] instead of [ふんだ (Interj)] as in Table 9. Moreover, even we extracts に by Michiru, there is no clue about the official character phrases for her like わぷっ and にょわ. We also did not find Misuzu’ s がお as her character phrase, which might also be caused by missplit of morphemes.

We planned to analyze other media (e.g., anime) and a series of work (e.g., all three seasons of *Yuru Camp*), which might show the variance of yakuwarigo over years of production. However,

we found the difficulty in anime speech diarization (speaker recognition given subtitle line and corresponding audio clip), so it is difficult to extract the speaker information not existing in closed captions subtitle lines from the audio clips. We then decided not to analyze anime in this work. There is a related work (Sato 2023) analyzing yakuwarigo in *5-toubun no Hanayome* anime series, but they manually anootated the speakers on the subtitle lines, which is not feasible for us.

Acknowledgements

This paper serves as the final project report (Language Use Analysis Project) of ASIAN 358 (Japanese Sociolinguistics) at University of Wisconsin–Madison in Fall 2024. I would specially thank the instructor Prof. Junko Mori for her guidance on sociolinguistics analysis and my friend Mike Qi for his help in data collection from the game *AIR*.

References

Dahlberg-Dodd, Hannah E. 2022. "Katakana and the Mediatized Other: Script Variation in Fantastical Narratives." *Japanese Studies* 42 (1): 61–79. https://doi.org/10.1080/10371397.2022.2027749.

Ma, Tzuhsuan. 2019. "計量的アプローチによる役割語の分類と抽出の試み." 計量国語学 32 (2): 103–16. https://doi.org/10.24701/mathling.32.2_103.

McCann, Paul. 2020. "Fugashi, a Tool for Tokenizing Japanese in Python." In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, edited by Eunjeong L. Park, Masato Hagiwara, Dmitrijs Milajevs, Nelson F. Liu, Geeticka Chauhan, and Liling Tan, 44–51. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.nlposs-1.7.

NINJAL. 1997. "テレビ放送の語彙調査 2 語彙表." 国立国語研究所. https://doi.org/10.15084/00001283.

Obana, Yasuko. 2019. "Politeness." In *Routledge Handbook of Japanese Sociolinguistics*. Routledge.

Sato, Manaka. 2023. "キャラクターの性格を特徴づける言葉づかい—アニメ『五等分の花嫁』の分析—." In. 東京外国語大学大学院国際日本学研究院. https://tufs.repo.nii.ac.jp/records/5692.

SturtzSreetharan, Cindi. 2017. "Language and Masculinity: The Role of Osaka Dialect in Contemporary Ideals of Fatherhood." *Gender and Language* 11 (4): 552–74. https://doi.org/10.1558/genl.31609.

Teshigawara, Mihoko, and Satoshi Kinsui. 2011. "Modern Japanese 'Role Language' (Yakuwarigo): Fictionalised Orality in Japanese Literature and Popular Culture." *Sociolinguistic Studies* 5 (1): 37–58. https://doi.org/10.1558/sols.v5i1.37.

Ward Jr., Joe H. 1963. "Hierarchical Grouping to Optimize an Objective Function." *Journal of the American Statistical Association* 58 (301): 236–44. https://doi.org/10.1080/01621459.1963.10500845.

5 Appendix: Tables

5.1 Individual Clusters

line	speaker	text
SEEN0700-113	神奈	なにゆえに、余は閉じ込められねばならぬのだ
SEEN0700-1643	神奈	そのような者に、余は護られとうない
SEEN0701-1261	神奈	…余はおぬしが今、この上なく無礼なことを考えておるような気がするぞ
SEEN0700-376	神奈	それに、余は神の使いなどではない
SEEN0700-322	神奈	余は余であって、まろでもわらわでもないぞ

Table 3: Cluster 0: Usages of Pron 余

line	speaker	text
SEEN0701-1367	神奈	うむ。わかった
SEEN0702-1315	神奈	うむ。母上がそう申すのなら…
SEEN0240-974	聖	うむ。ダシのとり方も絶妙、薬味も厳選素材を使っているからな
SEEN0502-445	往人	うむ。水筒の中身はスポーツドリンクでもＯＫだ
SEEN0700-1736	神奈	うむ…

Table 4: Cluster 0: Usages of Interj うむ

line	speaker	text
SEEN0295-299	佳乃	あたしお金持ってるから、オゴりだよぉ
SEEN0307-295	みちる	うんっ。けっこーなことだ。にゃはは
SEEN0203-613	観鈴	暑いねー
SEEN0200-552	少女	ねっ、ポテト
SEEN0200-692	観鈴	社会人〜
SEEN0317-812	みちる	じゃーね。みちるはここでおわかれでーーーす
SEEN0203-736	観鈴	痛かったよね。でも、こうしたら痛いの飛んでくからね
SEEN0180-136	少女	うんっ…
SEEN0250-1291	佳乃	えええぇーっ。行こうよぉ、きっと楽しいよぉ
SEEN0502-748	佳乃	えっと…
SEEN0183-837	観鈴	そうそう。わたしもそう思う。お母さん、惚れちゃうよね
SEEN0220-956	観鈴	うん。一度だけ飲んだことあるの
SEEN0220-344	少女	キミは？
SEEN0240-988	佳乃	ふあっ。やっぱり最高にサワヤカでおいしいね〜
SEEN0230-101	観鈴	なんか、ほんとにオタマジャクシみたいになっちゃったね

Table 5: Cluster 1: lines with significant words

speaker	わたし (代名)	あたし (代名)	ちゃっ (助動)	ちゃう (助動)	っと (格助)	し (接続)	に (接続)	ちゃ (接続)	ね (終助)	よぉ (終助)	ねぇ (終助)	ねー (終助)	うん (感動)
みちる	0	0	6	9	2	3	36	8	131	19	22	14	54
佳乃	0	61	18	18	1	5	0	11	75	166	84	14	42
声	3	0	5	1	0	1	0	0	8	3	3	2	1
女の子	0	1	0	0	0	0	0	1	11	5	5	1	7
少女	8	16	4	5	4	9	0	0	41	36	14	9	10
観鈴	381	0	67	35	71	110	0	10	550	1	9	92	358

Table 6: Cluster 1: count of lines with significant words

line	speaker	text
SEEN0702-1116	裏葉	しかし**ながら**、神奈さまのお気持ちはよくわかり**ます**
SEEN0644-537	美凪	…星は…出ていま**せん**でした
SEEN0701-931	裏葉	神奈さま、**こちら**ならよく見え**ます**
SEEN0260-1269	美凪	…実は**私**…こういうものを持って**ます**
SEEN0611-66	美凪	…あまりに楽しいの**で**…思わずこれを進呈
SEEN0609-21	美凪	…着替えとかは持って**ますし**…
SEEN0703-364	裏葉	生き残らねば…ならないの**ですか**…
SEEN0702-1610	裏葉	いささか不作法**ですが**…
SEEN0614-431	美凪	…**私**はみちるのお陰**で**、**私**でいれた…
SEEN0260-1043	美凪	…わかりま**した**
SEEN0260-1232	美凪	…**はい**…**えっと**
SEEN0250-794	美凪	…どうかしま**したか**
SEEN0700-1072	裏葉	神奈さまの御為なら、この命ささげようとも惜しくはございま**せん**
SEEN0277-222	美凪	…**あ**
SEEN0701-485	裏葉	**はい**、ただいま

Table 7: Cluster 2: lines with significant words

line	speaker	text
SEEN0430-660	母	あんたが一晩中寝んと、トランプしとる夢や…
SEEN0233-666	女	カラスにさえかまってもらわれへん…
SEEN0180-1662	晴子	当然や。あの子、誘えるかいな
SEEN0233-742	女	ヒヨコの親はニワトリや。恐竜なんかにならへんのに
SEEN0410-104	女	せやろ？
SEEN0230-1306	晴子	あんた、出ていく前に、今着てる服返しや
SEEN0180-1641	晴子	あの子な、なんでかひよこを恐竜の子供や、思い込んでたねん
SEEN0430-776	母	応援されてるっ…なんやわからんけど、飲んでみよっ
SEEN0200-884	晴子	ちゃんと立っとるな。よし、無事や
SEEN0213-155	女	あほっ、あんたは後ろやっ
SEEN0420-957	女	重いで。あんたが抱いとった頃から、ずいぶん時間経ったからな
SEEN0200-1464	晴子	ほなまぁ、頑張って稼ぎや
SEEN0180-1796	晴子	なんや、弱いんかいな
SEEN0440-73	母	それにな、その保育所に、あほな子がおるねん
SEEN0241-370	晴子	それもなー、誰かと遊んでたんやなくて、一人でいて泣き出したっていうねん

Table 8: Cluster 3: lines with significant words

line	speaker	text
SEEN0230-965	みちる	ふんだ、もうあんたの相手なんかしてやんないっ
SEEN0230-999	みちる	ふんだっ
SEEN0230-1042	みちる	ふんだっ、あんたなんか泣いちゃえっ
SEEN0230-1111	みちる	ふんだっ。あんたなんか、そこで泣いてろっ
SEEN0307-159	みちる	ふんだっ。そんなこというなら、お金のかせぎかた、おしえてやんない
SEEN0317-85	みちる	ふんだっ。やっぱり＊Ｂ＊Ａなんかに、はなすんじゃなかった
SEEN0317-323	みちる	ふ、ふんだっ。食べたかったら、かってに食べればいいよ

Table 9: Cluster 1: Wrong Classified Usages of Pron ふん

line	speaker	text
SEEN0613-26	みちる	んに…
SEEN0602-22	みちる	んに…でも、まだお星様でてないもん
SEEN0317-463	みちる	んに…そだね。まだ、あしたがあるね
SEEN0613-18	みちる	んに…おはよう…
SEEN0607-85	みちる	んに…やっぱりおかしいねぇ
SEEN0317-412	みちる	んに…＊Ｂ＊Ａにざんねんなおしらせがあります
SEEN0307-298	みちる	んにっ!?
SEEN0250-611	みちる	んにっ!?
SEEN0307-332	みちる	んに…
SEEN0612-60	みちる	んに…

Table 10: Cluster 1: Wrongly Classified Usages of Conj Part に

5.2 Significant and Minor Words

CN	Pron (代名)	Aux Verb (助動)	Case Part (格助)	Adv Part (副助)	Conj Part (接続)	Final Part (終助)	Noun Part (準体)	Interj (感動)
0	俺 (2.24, 551) おまえ (2.23, 469) 君 (2.18, 111) 余 (2.19, 95) あいつ (2.13, 57)	だろ (2.18, 378) だろう (2.14, 147) てろ (2.17, 33) だろっ (2.24, 19)		ぞ (2.2, 161) ツ (2.24, 7)		ぞ (2.17, 477) ぜ (2.16, 55)		ああ (2.07, 501) いや (2.1, 202) おい (2.19, 45) うむ (2.24, 42) ぐはっ (2.24, 28)
1	わたし (3.31, 402) あたし (3.3, 78) うぬ (3.34, 57) キミ (3.34, 12) ふん (3.34, 7)	ちゃっ (3.06, 100) ちゃう (2.64, 68) ね (2.61, 64) だぁ (3.09, 38) たー (3.08, 36)	っと (2.13, 78) 〜 (2.54, 32) でー (2.0, 9) じゃあ (2.43, 8) にぃ (3.34, 5)	ばっかり (2.34, 7)	し (2.15, 135) に (3.34, 36) ちゃ (2.38, 30) じゃあ (3.34, 10) からぁ (3.34, 6)	ね (2.71, 848) よぉ (3.24, 234) の (2.98, 185) ねぇ (2.97, 137) ねー (3.26, 133)		うん (2.83, 485) あ (2.07, 128) うーん (3.09, 98) えっと (2.31, 92) うんっ (3.34, 88)
2	私 (5.92, 149) わたくし (9.56, 39) こちら (6.32, 15) どちら (4.86, 6)	です (7.58, 410) ます (8.4, 381) で (2.13, 220) まし (7.65, 167) ませ (8.51, 126)	へ (2.37, 9)	など (4.21, 18) ばかり (2.04, 6) さえ (2.39, 5)	ながら (2.4, 13) とも (6.02, 8)			はい (7.05, 184) あ (2.08, 31) えっと (2.3, 29) ありがとう (6.27, 25) いえ (9.3, 15)
3	あんた (4.78, 312) なん (2.13, 157) うち (6.3, 26) かん (6.3, 20) わらわ (3.5, 10)	や (6.05, 589) たら (2.23, 217) やろ (6.14, 160) へん (6.25, 139) せ (3.49, 101)		や (5.92, 603) なんか (2.54, 42) て (3.45, 17) かー (4.72, 6) なん (6.3, 5)	で (2.24, 126) たって (3.43, 12) ど (3.33, 9) さかい (6.3, 6)	わ (3.16, 186) ねん (5.84, 116) や (5.55, 82) い (4.35, 78) なー (3.65, 77)		ん (3.89, 94) ほな (6.3, 62) ええ (4.97, 45) よっしゃ (6.3, 20) よし (2.38, 20)

Table 11: Coefficient of Specialization (CoS) >2: significant words, at most 5 for each cell ranked by frequency. Our focus of analysis. Number in parentheses after every word are its CoS and frequency in the cluster.

CN	Pron (代名)	Aux Verb (助動)	Case Part (格助)	Adv Part (副助)	Conj Part (接続)	Final Part (終助)	Noun Part (準体)	Interj (感動)
0	あんた (0.45, 83)	ます (0.2, 43) です (0.1, 25) まし (0.16, 16) や (0.05, 13) ませ (0.16, 11)	っと (0.13, 7)	なんか (0.5, 23) や (0.06, 17) たり (0.41, 15) て (0.43, 6)	し (0.49, 46) ちゃ (0.43, 8)	ね (0.05, 25) い (0.42, 21) なぁ (0.21, 19) ねぇ (0.25, 17) わ (0.07, 11)		はい (0.3, 37) うん (0.12, 31) あ (0.33, 30) あ (0.4, 28) えっ (0.46, 12)
1	何 (0.47, 33) あんた (0.13, 16) こいつ (0.41, 9)	ます (0.33, 47) なら (0.38, 28) せ (0.24, 13) だろ (0.09, 11) や (0.05, 10)	を (0.45, 244)	や (0.06, 11)	が (0.35, 30)	な (0.45, 337) い (0.38, 13) ぞ (0.08, 12) なあ (0.45, 7) や (0.18, 5)		ええ (0.35, 6) ありがとう (0.4, 5) いや (0.08, 5) ああ (0.03, 5)
2	なん (0.29, 13) どこ (0.44, 7)	だ (0.18, 68) たら (0.34, 20) てる (0.21, 16) じゃ (0.24, 11) だっ (0.21, 7)		って (0.29, 19) や (0.16, 10)		よ (0.46, 62) の (0.25, 5)		ああ (0.14, 7) ん (0.42, 6)
3	何 (0.48, 18)	ない (0.16, 34) だ (0.05, 31) です (0.15, 14) ませ (0.44, 11) ます (0.09, 7)	を (0.37, 107)		ば (0.4, 27)	よ (0.21, 46) ね (0.14, 24) の (0.27, 9)	の (0.45, 121)	ああ (0.35, 30) いや (0.23, 8) うん (0.08, 7)

Table 12: Coefficient of Specialization (CoS) <0.5: minor words, at most 5 for each cell ranked by frequency. Number in parentheses after every word are its CoS and frequency in the cluster.