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

ASIAN 358 Japanese Sociolinguistics:
Final Presentation (2024/12/5)

Ruixuan Tu (ruixuan.tu@wisc.edu)

Role Language “Yakuwarigo” and “特徴語”

- **“Yakuwarigo” (defined on individual character):** sets of spoken language features (e.g. vocabulary and grammar) and phonetic characteristics (e.g. intonation and accent patterns), associated with particular character types (Kinsui 2003 as cited in Teshigawara & Kinsui 2012)
- **Keyword “特徴語” (defined on group): significant** (CoS >2) and minor (CoS <0.5) words (Ma 2019)
- **C**oefficient **o**f **S**pecialization of **W**ord in **C**luster “特化係数”:

$$\mathrm{CoS}(W,C)=\frac{\text{freq of } W \text{ in cluster } C}{\text{number of words in cluster } C}\Big/\frac{\text{freq of } W \text{ in whole}}{\text{number of words in whole}}$$

- **RQ:** Keyword → Yakuwarigo? **Example:** AIR

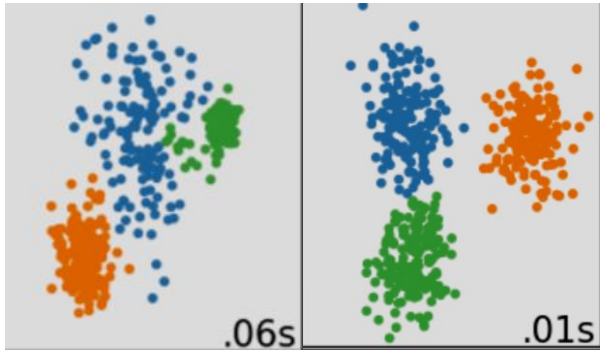
Cluster Analysis Technical Terms

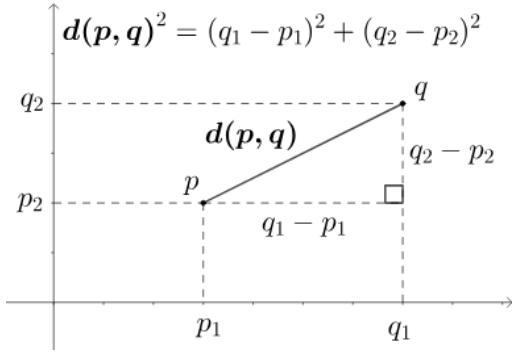
- **Distance between clusters:** Ward's method
  - Metric: increase of **S**um of **S**quares when we merge two clusters: $d(A,B) = \frac{|A|\cdot|B|}{|A \cup B|}\|\mu_A - \mu_B\|_2^2$
  - Idea: minimize variance from cluster centers
  - Connection:

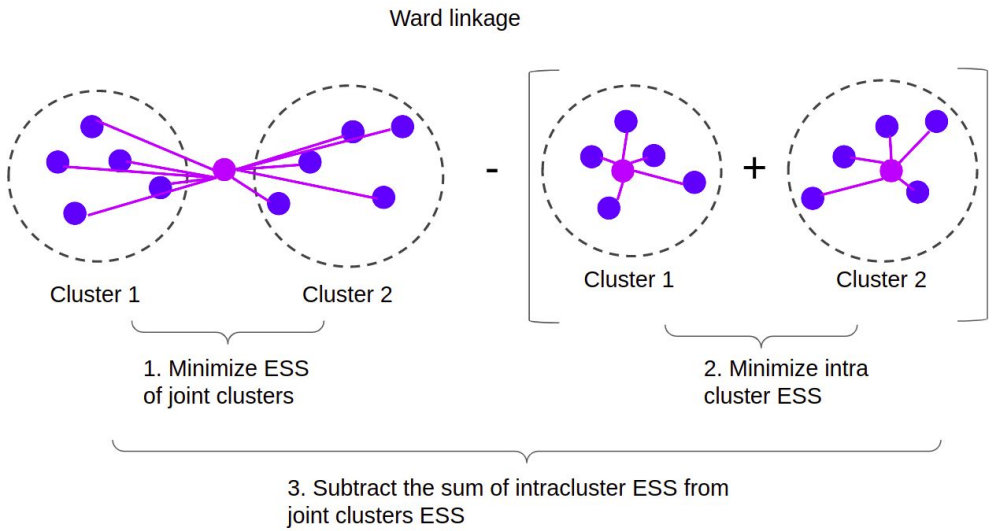
$SS = \sum(x_i - \bar{x})^2$; variance $s^2 = \frac{SS}{n-1}$

- **Distance between points:** Euclidean

$d(\mathbf{x},\mathbf{y})^2 = \|\mathbf{x}-\mathbf{y}\|_2^2 = \sum_{i=1}^{n}(x_i - y_i)^2$







AIR (2000): Visual Novel Game and Data

少年少女の恋愛劇に不可思議要素を絡めた**アドベンチャーゲーム**であり、シナリオが感動に特化した**泣きゲー**として支持を集めた。

Volumes: "Dream (Yukito & Misuzu)", "Summer (Kanna, 1000 years ago)", "Air (Misuzu only)"

Script: 麻枝准、イシカワタカシ

Scenario Assistant: 涼元悠一、雲龍寺魁、丘野塔也、藤井知貴

Line	Type	Speaker	Text
SEEN0180-885	dialogue	観鈴	「わ、すごい…掴んだ」
SEEN0180-887	dialogue	往人	「なんだ、セミか…」
SEEN0180-890	dialogue	観鈴	「わわわーっ！」
SEEN0180-891	dialogue	観鈴	「放さないでっ！」
SEEN0180-893	dialogue	観鈴	「わ、また掴んだ…」
SEEN0180-894	dialogue	往人	「…どうすればいいんだ、俺は」
SEEN0180-895	dialogue	観鈴	「外っ。はいはいっ」
SEEN0180-896	dialogue	往人	「……」
SEEN0180-898	dialogue	観鈴	「はい、放してー」
SEEN0180-901	dialogue	観鈴	「ふぅ…」
SEEN0180-902	dialogue	往人	「騒がしいやつだな」
SEEN0180-903	dialogue	観鈴	「騒がしいのはセミっ。わたしじゃないの」
SEEN0180-904	dialogue	観鈴	「換気扇から、たまに入ってくるの」
SEEN0180-905	dialogue	往人	「ふぅん…」
SEEN0180-906	dialogue	観鈴	「でも、すごく器用。飛んでるセミを掴めるなんて」
SEEN0180-907	dialogue	往人	「そうか？」

Morpheme Analyze “日本語を勉強する” →
“[日本][語][を][勉強][する]”

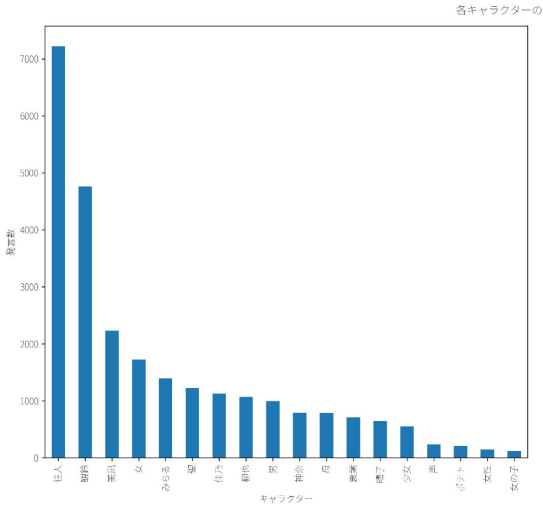
AIR: Preparation for Quantitative Analysis

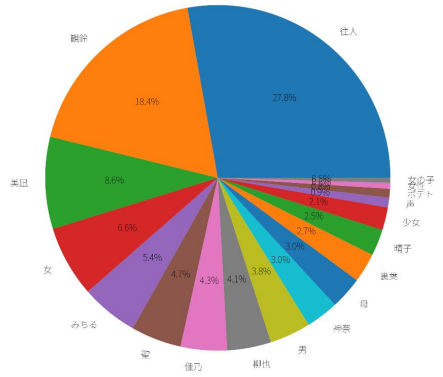
1. **Select** analyzable speakers with >100 lines
2. **Split** tokens by morpheme analyzer
3. Adjusted Frequency for **W**ord and **S**peaker $T(S, W)$:

$$10000 \times \frac{\text{freq of } W \text{ in speaker } S}{\text{number of words in speaker } S}$$

(frequency of W per 10000 words for speaker S)

4. Adjusted Frequency Table for (word, speaker) pairs





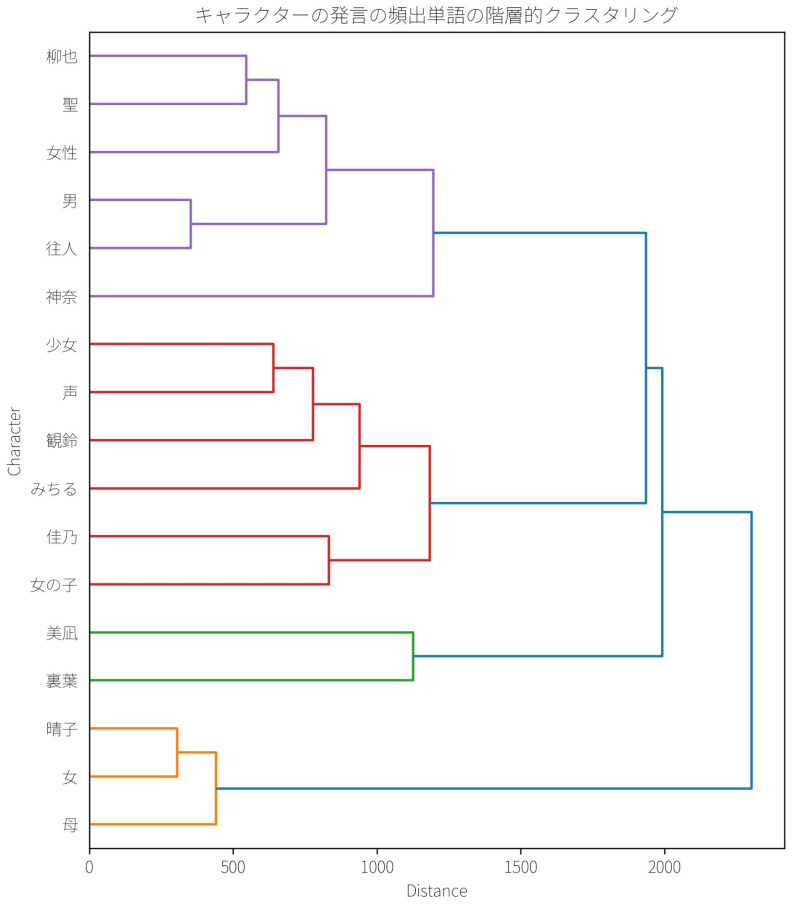
	声	往人	みちる	神奈	裏葉	少女	観鈴	女
えいっ/感動詞/	14.9	0.48	8.99	9.4	6.52	0	0	
た/助動詞/*	551.42	451.14	354.98	416.54	348.76	590.57	589.81	502
に/助詞/格助詞	312.97	355.18	750.39	410.27	668.19	397.5	474.24	445
は/助詞/係助詞	283.16	502.22	413.39	858.13	700.78	244.18	389.68	343
はっ/感動詞/—	14.9	1.91	0	0	3.26	0	16.91	
よし/感動詞/—	44.71	7.64	0	0	0	0	4.23	18
あ/感動詞/フィ	29.81	9.55	31.45	9.4	22.82	39.75	20.44	31
で/助動詞/*	59.61	84.02	78.63	357.03	563.89	107.89	141.64	137
も/助詞/係助詞	193.74	271.16	267.36	294.39	335.72	300.97	373.48	335
ない/助動詞/*	119.23	206.71	184.23	15.66	48.89	107.89	220.56	19
し/助詞/接続助	14.9	12.89	6.74	6.26	3.26	56.79	81.74	18

AIR: Cluster Analysis

- **Input:** every column of adjusted frequency table

$$T(S) = (T(S, W_1), \ldots, T(S, W_n))$$

- **Output:** clusters
  - Cluster 0: non-female language (往人)
  - Cluster 1: casual female language (観鈴)
  - Cluster 2: formal and polite language (美凪)
  - Cluster 3: dialectal language (晴子)



$$\mathrm{CoS}(W,C) = \frac{\text{freq of } W \text{ in cluster } C}{\text{number of words in cluster } C} \Big/ \frac{\text{freq of } W \text{ in whole}}{\text{number of words in whole}}$$

AIR: Extraction of Role Language

- **Calculate C**oefficient **o**f **S**pecialization, **select** significant and minor morphemes, **rank** by high frequency
- **Confirm** clusters:
  - Cluster 0: non-female language
  - Cluster 1: casual female language
  - Cluster 2: formal and polite language
  - Cluster 3: dialectal language

SEEN0230-1108 dialogue 往人 ぐはっ！やっぱりかっ！

往人 majority

聖、神奈 majority

神奈 only

してろ=していろ

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) ~~さん (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)

CoS >2 Significant **Focus**

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)

CoS >0.5 Minor

Cluster 0: 柳也、聖、女性、男、往人、神奈

AIR: Exceptions in Non-female Language

- **Pron「余」** by Kanna 神奈: Lived 1000 years ago, goddess in a shrine, last winged beings (翼人), calm speaking; met Ryuya 柳也 and Uraha 裏葉 to escape and meet her mother
- **Interj「うむ」** by Hijiri 聖、Kanna 神奈-majority Elder sister of Kano 佳乃, doctor at Kirishima Clinic, always carry a lancet
- →Non-typical female language

43795	SEEN0700-91	dialogue	神奈	なにゆえ、余が案内せねばならぬ
43799	SEEN0700-95	dialogue	神奈	なにゆえ、余が、おぬしを、案内せねば、ならぬ？
43817	SEEN0700-113	dialogue	神奈	なにゆえに、余は閉じ込められねばならぬのだ
43818	SEEN0700-114	dialogue	神奈	余はひとりでも生きてゆけるぞ…
43852	SEEN0700-148	dialogue	神奈	余はここで暮らしておる
…	…	…	…	…
48523	SEEN0702-1593	dialogue	神奈	余の夢だ…
48545	SEEN0702-1615	dialogue	神奈	余の最後の命である
48986	SEEN0703-290	dialogue	神奈	…余の命であるぞっ…
48991	SEEN0703-295	dialogue	神奈	なぜ…なぜにみな、余だけを…残して…
49375	SEEN0703-679	dialogue	神奈	余も手伝おうぞ

38513	SEEN0603-323	dialogue	聖	うむ。良い選択だ
39978	SEEN0608-313	dialogue	聖	うむ。悩みがあるのなら、この綺麗なお姉さんに話してごらん
40352	SEEN0609-28	dialogue	往人	うむ。それが何かは、道々教えてやる
43987	SEEN0700-283	dialogue	神奈	うむ。わかればよい
44110	SEEN0700-406	dialogue	神奈	うむ、大儀であった
44579	SEEN0700-876	dialogue	神奈	むーっ！むむむっ！むむむぐうむうむ……

Cluster 1: 少女、声、**観鈴**、**みちる**、**佳乃**、女の子



AIR: Casual Female Language

- **Pick** 15 lines randomly
- Usual schoolgirl conversation
- However, occurrences (よぉ, うん, ねー, ちゃう) are shared within the group – same author?

	わたし (3.31, 402)	ちゃっ (3.06, 100)	っと (2.13, 78)		し (2.15, 135)	ね (2.71, 848)	うん (2.83, 485)
	あたし (3.3, 78)	ちゃう (2.64, 68)	～ (2.54, 32)		に (3.34, 36)	よぉ (3.24, 234)	あ (2.07, 128)
1	うぬ (3.34, 57)	ね (2.61, 64)	でー (2.0, 9)	ばっかり (2.34, 7)	ちゃ (2.38, 30)	の (2.98, 185)	うーん (3.09, 98)
	キミ (3.34, 12)	だぁ (3.09, 38)	じゃあ (2.43, 8)		じゃあ (3.34, 10)	ねぇ (2.97, 137)	えっと (2.31, 92)
	ふん (3.34, 7)	たー (3.08, 36)	にぃ (3.34, 5)		からぁ (3.34, 6)	ねー (3.26, 133)	うんっ (3.34, 88)

speaker	text
佳乃	**あたし**お金持ってるから、オゴりだ**よぉ**
みちる	**うんっ**。けっこーなことだ。にゃはは
観鈴	暑い**ねー**
少女	**ねっ**、ポテト
観鈴	社会人～
みちる	じゃー**ね**。みちるはここでおわかれでーーーす
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**
少女	**うん**っ…
佳乃	えええーっ。行こう**よぉ**、きっと楽しい**よぉ**
佳乃	**えっと**…
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**
観鈴	**うん**。一度だけ飲んだことあるの
少女	**キミ**は？
佳乃	ふあっ。やっぱり最高にサワヤカでおいしい**ね**～
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**

Cluster 2: **美凪、裏葉**

AIR: Formal and Polite Language

Minagi 美凪: schoolgirl, astronomy club leader, calm/mild

Uraha 裏葉: female court lady of Kanna 神奈, ancestor of Misuzu 観鈴

In common: desu/masu form, use of Keigo

Uraha: excessive use of Keigo

speaker	text
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**
美凪	...星は...出てい**ません**でした
裏葉	神奈さま、**こちら**ならよく見え**ます**
美凪	...実は**私**...こういうものを持って**ます**
美凪	...あまりに楽しいの**で**...思わずこれを進呈
美凪	...着替えとかは持って**ますし**...
裏葉	生き残らねば...ならないの**ですか**...
裏葉	いささか不作法**ですが**..
美凪	...**私**はみちるのお陰**で**、**私**でいれた...
美凪	...わかり**ました**
美凪	...**はい**...**えっと**
美凪	...どうかし**ました**か
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**
美凪	...**あ**
裏葉	**はい**、ただいま

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)

Cluster 3: 晴子、女、母

AIR: Dialectal Language

Haruko 晴子: Misuzu 観鈴's aunt; single guardian of Misuzu 観鈴 after the death of her mother; addicted to alcohol

Mother 母: Misuzu 観鈴's mother

In common: Osaka dialect

「せや」「なんや」「ねん」「へん」– company location

Still dialect? 「あんた」「ほな」「よし」

speaker	text
母	**あんた**が一晩中寝んと、トランプしとる夢**や**...
女	カラスにさえかまってもらわれ**へん**...
晴子	当然**や**。あの子、誘えるか**い**な
女	ヒヨコの親はニワトリ**や**。恐竜**なんか**にならへんのに
女	**せやろ**？
晴子	**あんた**、出ていく前に、今着てる服返し**や**
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**
母	応援されてるっ...**なんや**わからんけど、飲ん**で**みよっ
晴子	ちゃんと立っとるな。**よし**、無事**や**
女	あほっ、**あんた**は後ろやっ
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
晴子	**ほな**まぁ、頑張って稼ぎ**や**
晴子	**なんや**、弱いんか**い**な
母	それにな、その保育所に、あほな子がおる**ねん**
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**

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)

Conclusion

- In *AIR*,
  - Keyword “特徴語” from our method could be recognized as “Yakuwarigo” that represents characteristics of specific individuals and groups;
  - Different script authors might affect extracted Keywords;
  - Observed non-female language, casual female language, formal and polite female language, and dialectal language as clusters.

Limitations

- Not analyzing the minor group (CoS <0.5);
- Misclassification of morphemes (or granularity issue);
  - ふんだ → [ふん (Pron)][だ] ❌
- Not exhaustive analysis of all possible role languages;
- Michiru みちる has phrases like「わぷっ」「にょわ」not recognized by our system (confuses morpheme analyzer);
- Only analyzed *AIR*, not other media (e.g., anime) and series
  - Difficulty in anime speech diarization – speaker recognition given subtitle line and corresponding audio clip

speaker	text
みちる	ふんだ、もうあんたの相手なんかしてやんないっ
みちる	ふんだっ
みちる	ふんだっ、あんたなんか泣いちゃえっ
みちる	ふんだっ。あんたなんか、そこで泣いてろっ
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
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
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

Credits

麻子軒. (2019). 計量的アプローチによる役割語の分類と抽出の試み テレビゲーム『ドラゴンクエスト 3』を例に. 計量国語学, 32(2), 103-116.

国立国語研究所, The National Language Research Institute, n.d., Vocabulary survey of television broadcasts 2 : Vocabulary lists: 国立国語研究所.

Teshigawara, M., & Kinsui, S. (2011). Modern Japanese 'Role Language' (Yakuwarigo): fictionalised orality in Japanese literature and popular culture. Sociolinguistic Studies, 5(1), 37.

McCann, P. (2020). fugashi, a tool for tokenizing Japanese in python. arXiv preprint arXiv:2010.06858.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. Journal of the American statistical association, 58(301), 236-244.

AIR Homepage: https://key.visualarts.gr.jp/product/air/

AIR Wikipedia: https://ja.wikipedia.org/wiki/AIR_(%E3%82%B2%E3%83%BC%E3%83%A0)

Q&A

Thank you