Linguistics D Conference

2022.11.16

November 15, 2022

Neural-Chain- and Database-Based Neural Network for Simulating First Language Acquisition Process by Infants

Tomoyuki Shsihdio[1], Yasuhiro Ono[2], Hirotaka James Okano[3], Itsuo Kumazawa[4], Kenji Suzuki[1], Hilofumi Yamamoto[5]

1 Biomedical AI Research Unit, Tokyo Institute of Technology, Kanagawa, JP
2 Enspirea LLC, Chicago, USA
3 Laboratory of Regenerative Medicine, Jikei Medical School, Tokyo, JP
4 Laboratory for Future Interdisciplinary Research of Science and Technology, Tokyo Institute of Technology, Kanagawa, JP
5 Laboratory of Linguistics, Tokyo Institute of Technology, Tokyo, JP

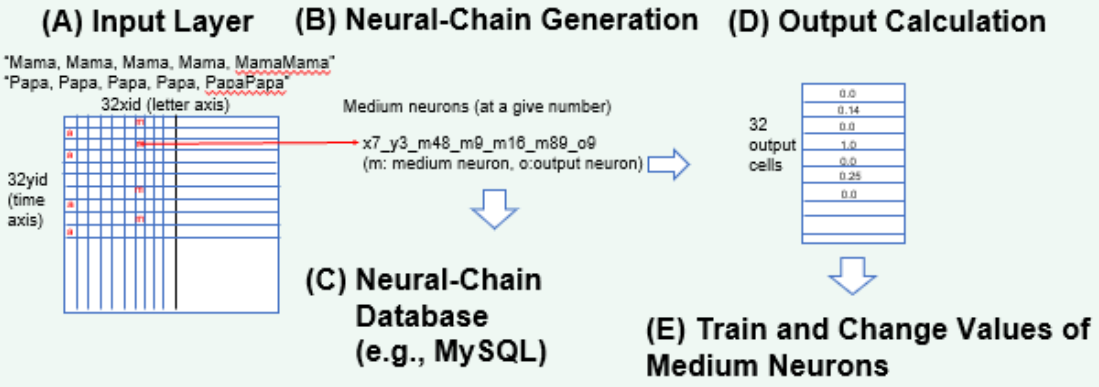
Highlights

- **Input word segment containing 32 characters is assigned to an input field of 32 x 32 cells for each cycle.**
- **A neural chain initiated by the input cell is created using a given umber of medium neurons.**
- **The value of each medium neuron is trained and changed according to usage of the medium neuron.**
- **Medium neuron values are summed to calculate a corresponding output cell (32x) value.**

Summary

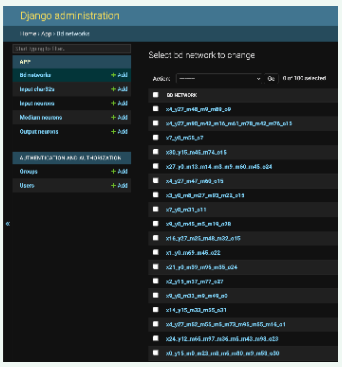
First language acquisition process is of interest, but has not been simulated in a programming-based neural network. In this study, we tried to construct a novel neural-chain- and database-based neural network to simulate a kind of column-based neural architecture in developing infant brain. First, each 32-character-containing input word segment (e.g., “Mama,Mama---”, “Papa,Papa---”) was mapped to an input layer containing 32 x 32 cells. A neural chain initiated by the input cell is created using a given umber of medium neurons (e.g., 7 corresponding to the number of human cortex layers). The value of each medium neuron was trained and modified according to the frequency of use in all the neural chains. Then, an output cell (one of an array of 32 cells) is ended by the neural chain, and the output value was calculated by adding the values of medium neurons in the corresponding chain.

Graphical Abstract

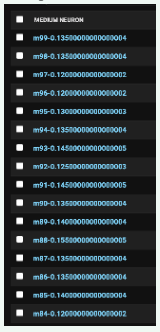


(**A**) Input layer structure. The input layer has two axes (i.e., the letter axis corresponding each single alphabet; the time axis corresponding to the order of the letter in the word segment). (**B**) Generation of neural chains. Each neural chain is initiated by the input cell, chained using a given number of medium neurons. The neural chain is ended by the corresponding output neuron (32x).(**C**) Each neural chain is stored in a neural chain database (MySQL) as well as individual medium neurons and their values are registered and updated. (**D**) The value of the corresponding output neuron is calculated by adding the values of medium neurons in the corresponding neural chain. (E) According to the frequency of use of each medium neuron and (statistical) comparison between the input and the output, the values of medium neurons are trained and modified.

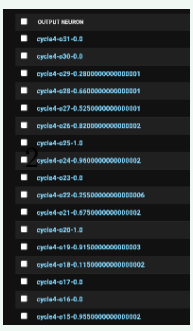
Neural Chains in Database



Medium Neuron Values (Weights)

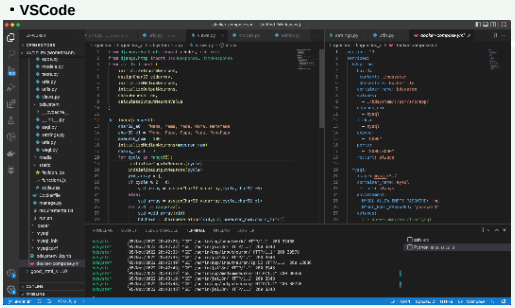


Output Neuron Values (at each cycle)



Programming Environment

- **Python (Django Framework)**
- **Database: MySQL**
- **Docker-compose**
- **VSCode**



Al Thuraiya Al Hashmi
21M58417

Department of Transdisciplinary Science and Engineering, Tokyo Institute of Technology

INVESTIGATING THE CAUSE OF CODE-SWITCHING AND ITS EFFECTS ON LOCAL SOCIETY

1. INTRODUCTION

Code-switching is known as changing languages or dialects throughout a single conversation or a single sentence. This phenomenon often happens with multi-lingual people (Wibowo, Yuniasih, & Nelfianti, 2017)

Why does code-switching occur? How does code-switching affect our society?

2. METHODS

I have conducted a literature review and gathered data through a questionnaire. My sample size for the questionnaire is international bilingual students with a total of 7 students.

3. RESULT

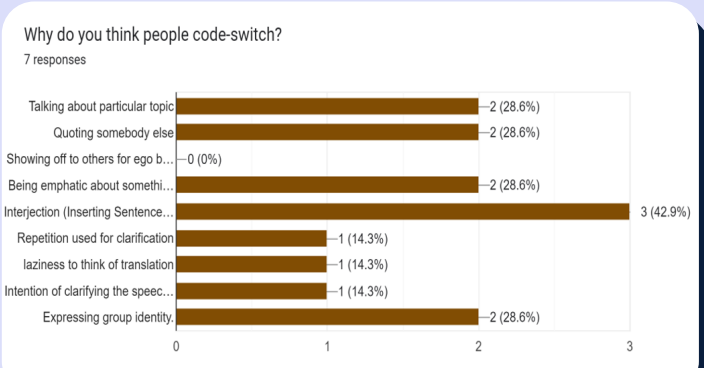


Fig. 1. Reasons for Using Code Switching based on questionnaire participants. Where options were based on (Girsang, 2015) & (Tannen, Hamilton, & Schiffrin, 2015)

- The participant's age was between their 20s-30s. with bilinguals as the majority, 57.1% followed by 28.6% for trilingual and 14.3% for quadrilingual.
- Participants often experienced codeswitching in social events, TV shows, and Social media.
- While 57.1% of participants code-switch, almost all participants are willing to ask for meaning if a foreign language is used.
- 42.9%of participants think that people code-switch for Interjection, while 28.6% agreed that Talking about a particular topic, Quoting somebody else, Being emphatic about something, and Expressing group identity are why people code-switch.

4. DISCUSSION

- Historical origin:

Hybrid Identities (colonization): Multicultural and Interethnic Identities cased by colonization made colonized society learn a new language and use code-switching to prevent the colonizers from understanding them (Tannen, Hamilton, & Schiffrin, 2015) & (HALL & NILEP, 2015).

- Modern day:

Sode-switching is associated with social events where those who network are recommended to learn English as a common language of international business. Furthermore, globalization has enabled the entertainment industry to attract worldwide sales through TV shows, songs, and advertisements. Moreover, people learn and code-switch after reallocating to another country for work or study (Chloros, 2009).

5. CONCLUSION

- Why does code-switching occur? How does code-switching affect our society?
- People are prone to code-switching in modern society, as a result of globalization. Literature review shows that acquiring knowledge, work, or entertainment is influenced by global culture & trends (Boudreau, McDaniel, Teng, Sprout, & Costa, 2022). While the questionnaire reports that code-switching is easy to encounter through news networks, social media, social events, etc.
- Business use code-switching to attract more customers and increase sales through globalization (Shaqiri, 2013). As a result, more people are influenced to use code-switching while referring to certain situations or seeking new jobs, etc. intertwining code-switching into our daily lives.

6. REFERENCE

- Boudreau, D., McDaniel, M., Teng, S., Sprout, E., & Costa, H. (2022). The Global Network. Retrieved from National Geographic : https://education.nationalgeographic.org/resource/global-network
- Chloros , P. G. (2009). Code-switching. Retrieved from Google books: https://books.google.co.jp/books?hl=en&lr=&id=_5TwiJ_A54C&oi=fnd&pg=PR10&dq=why+people+code-switching&ots=QNKcN9Qh_T&sig=PE0C5U7kE0o8r5zn3MiwNRLeKLw&redir_esc=y#v=onepage&q=why%20people%20code-switching&f=false
- Girsang, M. L. (2015). An Analysis of Code Switching and Code Mixing as Found in Television. Retrieved from Universitas HKBP Nommensen: https://uhn.ac.id/files/akademik_files/1712071009_2015_The%20%20Explora%20%20Journal%20%20Journal%20of%20English%20%20Language%20Teaching%20(ELT)%20and%20Linguistics_3.%20An%20Analysis%20of%20Code%20Switching%20and%20Code%20Mixing%20as%20Found%20in%20Tele
- HALL, K., & NILEP, C. (2015). The Handbook of Discourse Analysis, Code-Switching, identity, and globalization. Retrieved from colorado edu: https://www.colorado.edu/faculty/hall-kira/sites/default/files/attached-files/hall-nilep-2015-code-switching_identity_and_globalization.pdf
- Shaqiri, E. (2013). Globalization: The Tendency of Increasing Sales and. Retrieved from HR MARS: https://hrmars.com/papers_submitted/86/Globalization_The_Tendency_of_Increasing_Sales_and_Production_through_Outsourcing.pdf
- Tannen, D., Hamilton, H. E., & Schiffrin, D. (2015). The Handbook of Discourse Analysis second edition. Retrieved from https://www.sscnet.ucla.edu/anthro/faculty/ochs/articles/Ochs_2015_Discursive_Underpinnings.pdf
- Wibowo, A. I., Yuniasih, I., & Nelfianti, F. (2017, September). ANALYSIS OF TYPES CODE SWITCHING AND CODE MIXING. Retrieved from media neliti: https://media.neliti.com/media/publications/227310-analysis-of-types-code-switching-and-cod-1287515d.pdf

Connotation Detection for Classical Poetic Japanese Vocabulary Using Word Alignment Mismatch

Xudong Chen† Bor Hodošček‡ Hilofumi Yamamoto†

† 東京工業大学 Tokyo Institute of Technology ‡ 大阪大学 OSAKA UNIVERSITY

Linguistics D @ Tokyo Tech, 2022.11.16

Introduction

Objectives
- For a specific word (e.g. 「女郎花」, en. *Patrinia scabiosifolia*),
- find its connotation in each of the classical Japanese poems in which it appears.

Proposal
- Schramm's communication model
- Word alignment algorithms

歌番号 226

なにめてゝ
おれるはかりそ
女郎花
われおちにきと
人にかたるな

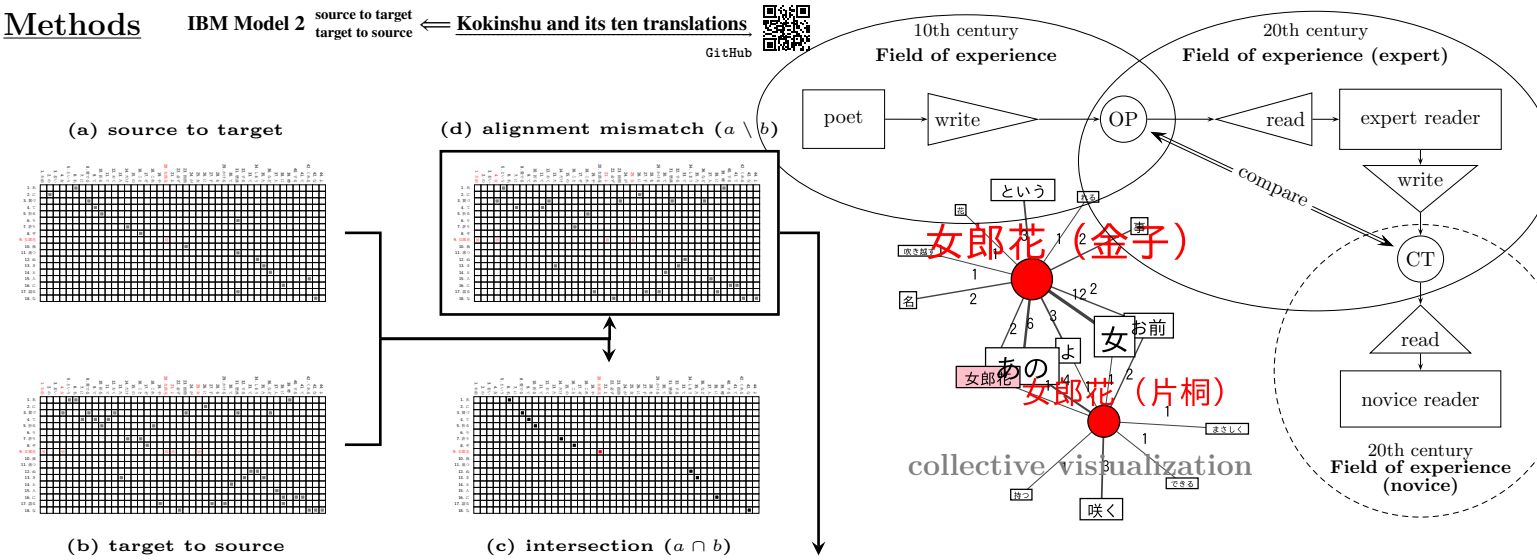
金子訳 (1933)

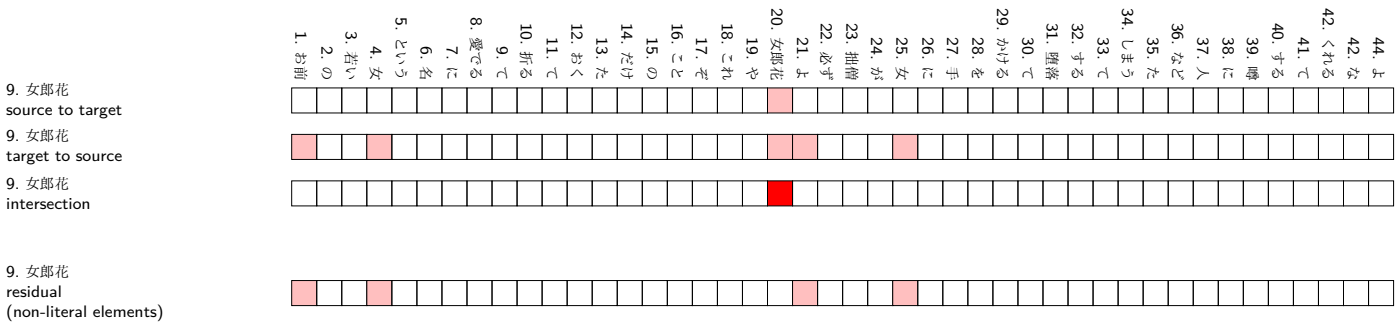
お前の若い女？(young woman)という名に愛でて、
折っておいただけの事ぞ、
これや女郎花よ、
必ず拙僧が女に手を掛けて堕落してしまったなど、
人に噂してくれるなよ。

Methods

IBM Model 2 source to target / target to source ⟸ Kokinshu and its ten translations

GitHub





Extraction process of non-leteral translations for 「女郎花」

Results

Accuracy of IBM Model 2

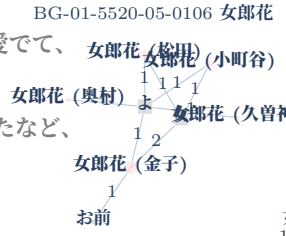
	source → target	target → source	intersection
precision	54.98	11.74	69.77
recall	61.27	52.09	56.12
AER	42.39	84.02	37.24

No.226

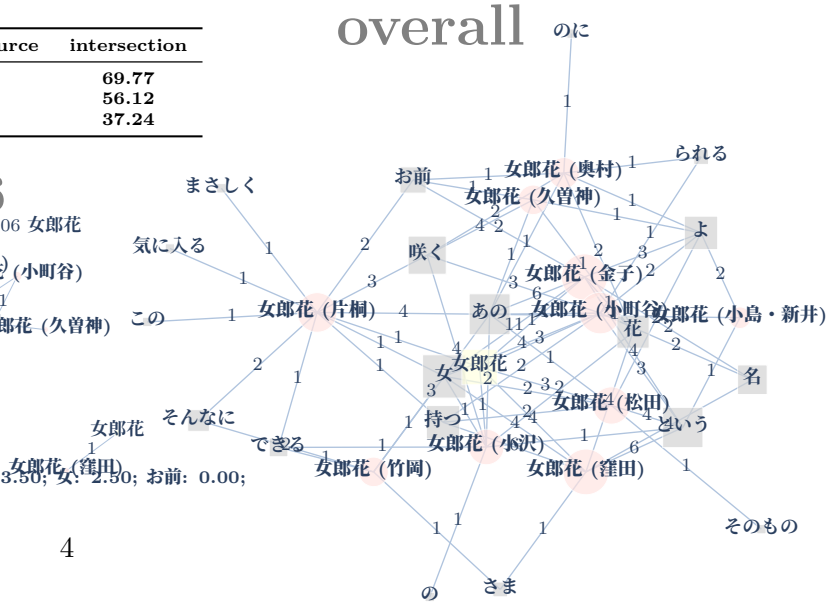
BG-01-5520-05-0106 女郎花

金子訳 (1933)

お前 (you) の若い**女** (woman) という名に愛でて、
折っておいただけの事ぞ、
これや女郎花**よ** (Interjection) 、
必ず拙僧が**女**に手を掛けて堕落してしまったなど、
人に噂してくれるなよ。



top-5 betweenness centrality: **よ: 13.50; 女: 2.50; お前: 0.00;**



5 betweenness centrality: **あの: 71.17; 女: 36.23; 女郎花: 24.45; 咲く: 20.53; よ: 17.10**

Discussion

- Although we were able to extract non-literal elements of each poetic word in each specific poem,
- Many functional elements which did not correspond to connotation were also included.

The relationship between phonemic and ideographic abilities in Chinese characters

YANG XIABIAO
Department of Mathematical and computing science
Tokyo Institute of Technology

1. Introduction

In most Chinese characters, they can be composed of two parts: one is able to indicate its pronunciation(phoneme) and the other one is able to indicate its meaning(ideograph). However, the ability of phoneme and ideograph can differ. We wonder about the relationship between phonemic and ideographic abilities in Chinese characters.

2. Methods

I investigated the phonemes and ideograms of the 50 Chinese characters in the book *Shuowen Jiezi*（说文解字）. Also, I make a list of the 10 components with the highest number of phonemes in *Shuowen Jiezi* and the most commonly used components in modern Chinese to roughly observe the relationship between phonemes and ideograms.

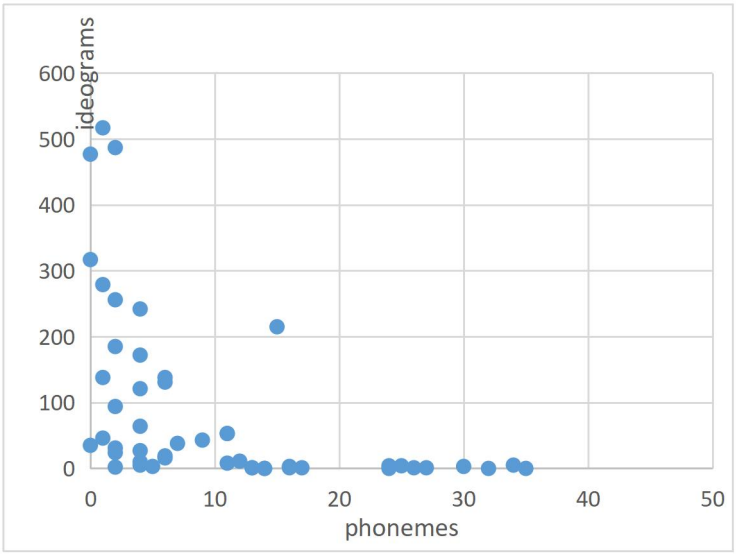
Components	No.ideograms	No.phonemes
水	488	1
艸	476	0
木	460	2
手	284	1
言	282	3
人	259	0
女	251	2
金	210	15
口	201	4
蟲	168	0

Table1: 10 components with the highest number of phonemes

Components	No.phonemes	No.ideograms
今	35	0
各	30	3
且	34	2
分	27	1
占	24	0
干	34	5
合	27	2

Table 2: the most commonly used components in modern Chinese

3. Result



Picture 1: the phonemes and ideograms of the 50 Chinese characters in the book Shuowen Jiezi

We could find from the picture above that if one component be seen as a phonemic part of a character then it rarely represents as a ideographic part.

4. Dicussion

This result is also logically understandable that if the components have both strong phonemic and ideographic abilities, then two different components, AB, can form a script, A can be phonemic and B can be ideographic, or B can be phonemic and A can be ideographic, which can form two different scripts, breaking the constraint relationship between the components of the morpheme.

However, it cannot be ignored that the character “金” seems to be one exception which is shown in the middle of the picture(more exploration next time maybe).

5. Conclusion

The component with a stronger phonemic abilities has a weaker ideographic abilities, and vice versa. There is no component with both stronger phonemic and ideographic abilities.

Reference: 许慎. 说文解字[M]. 艺术中国网, 2012.

Chinese localization tends to keep ‘Role Languages’, while English handles them with more flexibility.

A case study on how localization teams deal with Yakuwarigo

CHEN ZAIFENG

Yakuwarigo (Japanese: 役割語, "role language") is a style of language, often used in works of fiction, that conveys certain traits about its speaker such as age, gender, and class.

When cultural products with Yakuwarigo are introduced to other regions, localization will encounter difficulties. This time I would find some cases from the official localization of *Fate/Grand Order*, a free-to-play Japanese mobile game, to see how the localization teams deal with Yakuwarigo and analyze their effect. The game has lots of characters. Their dialogue uses lots of such things and can be viewed clearly.

- Character 'Benkei'
  - 拙僧が殺めるのだ
  - 由贫僧负责解决。
  - This humble Servant of Buddha will kill you.
- Character 'Nero'
  - 余は楽しい！
  - 余很开心！
  - I am having fun!
- Character 'Gilgamesh'
  - 一人称は『我』と書いて『オレ』と読む。
  - 第一人称写作『我』读作『本王』。
  - （Fan-made version）His first person pronoun writes as Ware (我) and reads as Ore (オレ).
- Character 'Emiya'
  - ん？ たまに一人称が変わるのはなんでかだって？ ……あぁ、そ、そうか、無意識にこぼれてしまうようだな……気を抜くと地が出るというか……『オレ』という呼称は若い頃の癖みたいなものなんだ。
  - 嗯？ 为什么有时候会改变第一人称？ ……这，这样啊。下意识说漏嘴了啊。或者说一放松警惕就暴露本性了么……咱这个称呼是我年轻时候的习惯。
  - You think I sometimes talk funny? Hmm, I never gave it much thought, but I guess I might slip back into my old self every now and then.
- Character O
  - 俺のやり方でいいの？ 誰も幸せにはなれないよ
  - 可以按照我的做法来办吗？没有人能获得幸福哦。
  - 可以按照老子的做法来办吗？没有人能获得幸福哦。
  - (Fan-made version) Are you okay with my methods? Nobody will become happy.

How a Book Influences Human Perception: Case of the Perception on ASD

Wishnu Agung Baroto

Department Social and Human Sciences, Tokyo Institute of Technology

Introduction

- ASD is a neurodevelopmental disorder characterized by social interaction, communication skills, interests, and behavior impairments.
- Many books have been published in Indonesia on ASD treatment to increase public understanding of ASD.
- The question is how books and available information can effectively influence people's perceptions or people understanding of ASD.

Methods

- Text mining on a book titled: *Penanganan dan Pendidikan Autis* (Treatment and Education for Autism) to find what the book is all about based on the frequency of words. Then, utilizing Orange Data Mining to analyze the data and visualize the result.
- Questionnaire on two groups (the reader and non-reader of the book) through Google form to capture perception.
- Analyze the open-ended question in hierarchy clustering using Orange Data Mining.

Findings

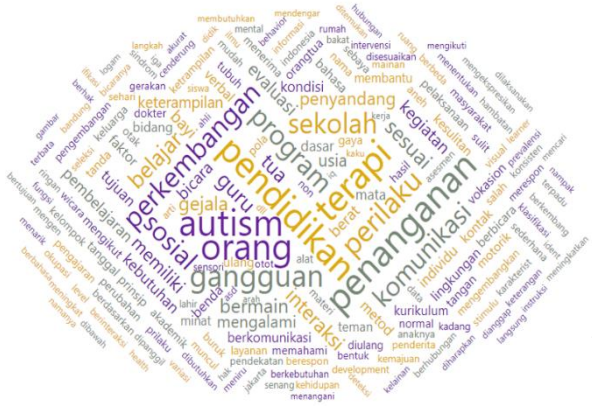
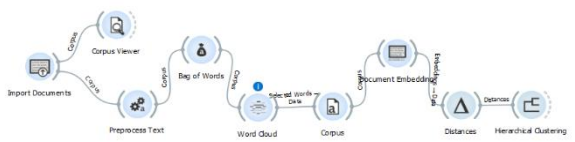


Figure 1: Word Cloud based on Word Frequency of *Penanganan dan Pendidikan Autis* Book



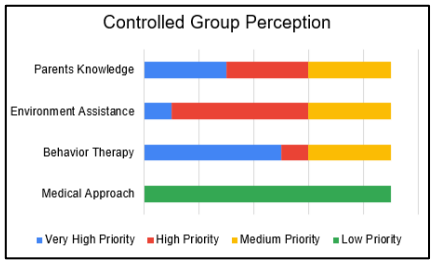


Figure 2: Controlled Group Perception on ASD Treatment

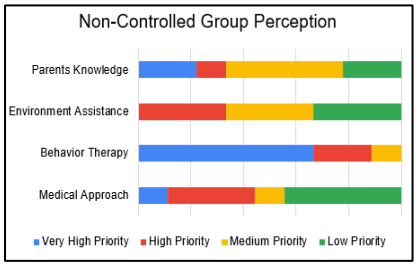


Figure 3: Non-Controlled Group Perception on ASD Treatment

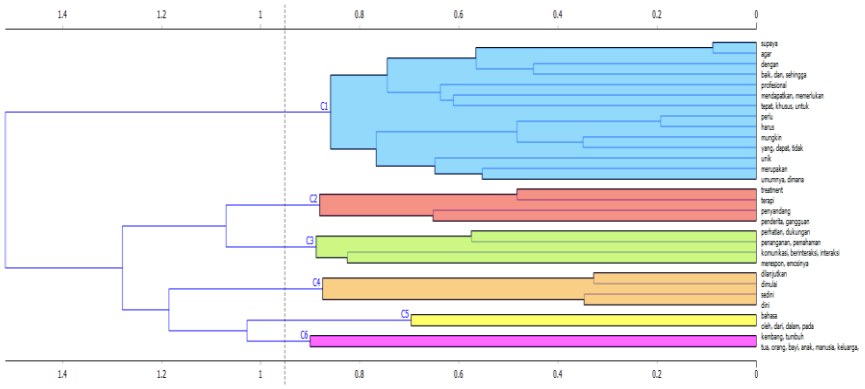


Figure 4: Hierarchy Clustering (Cosine Distance on Orange Data Mining)

Discussion

- The eight most frequent words based on the book are *Pendidikan* (Education), Autism (ASD), *Penanganan* (Treatment), *Terapi* (Therapy), *Orang* (People), *Gangguan* (Disorder), *Perilaku* (Behavior), and *Perkembangan* (Development).
- Based on the word's frequency, the book's content is an education for people about therapy treatment for behavior and development of ASD.
- The controlled group perceives parents' knowledge and therapy as high priority and medicine as classified as low priority, which is like what is implied in the book. However, for the non-controlled group, the results are more diverse.
- Hierarchy clustering analysis on the respondent's answer to "what do you think about ASD" questions can cluster similar meanings or close similarity among words.

Conclusion

- Text mining on a book is beneficial to retrieve emphasized words based on frequency.
- The analyzed book is able to influence people's perception of ASD treatment primarily related to medical treatment.
- Hierarchy clustering on sentences is helpful to group similar

References

- Ousley, O. & Cermak, T. (2013). Autism Spectrum Disorder: Defining Dimensions and Subgroups. Current Development Disorder Reports (2014) 1:20-28.
- Talib, R. et. al. (2016). Text Mining: Techniques, Applications, and Issues. International Journal of Advanced Computer Science and Applications Vol 7 No. 11.

To download the poster

How to pronounce Japanese by using backmasking

Jinfu Zhang / Tokyo Institute of Technology
zhang.j.av@m.titech.ac.jp

Introduction

◆ We were working on how to pronounce Japanese correctly by using backmasking.
◆ By changing the the Roman alphabet to the International Phonetic Alphabet, each phoneme of Japanese can be correctly backmasking.
◆ We focused on some difficult-to-pronounce combinations appearing during playing backmasking and we were looking for solutions.
◆ We paid special attention to the tonal conversion in Japanese, looking for a solution to the combination of high and low tones.

Methods

◆ STEP 1 split sentences into phonemes

Sentence
Syllable
Phoneme
Onset ex.(b, d, g, k)
Rime
Nucleus Ex.(a, i, u, e, o)
Coda Ex.(n)

◆ Phones are the smallest component of a language.
◆ Syllables break when playing backmasking, syllables are not independent.
◆ Phonemes don't break when playing backmasking , phonemes are independent. [1]

Example

ありがとうございます
Syllable
Onset
Onset
a ri ga to u go za i ma su
a r yi g a t o u g o z a yi m a s u

◆ STEP 2 Tones

◆ Tones are very important in Japanese. Same pronunciation with different tones can represent different meanings.
Ex. あめ(high low) 雨 あめ(low high) 飴
◆ **In Japanese, t**he tone of a single character can only be divided into high and low, there is no need to change the tones when backmasking.
◆ Tone is present on Rime, Onsets don't have tone.

The "Three Elements" of Sound

◆**loudness**
◆**pitch(tone)**
◆ **timbre**

High
Low
a r yi g a t o u g o z a yi m a s u
reverse
u s a m yi a z o g u o t a g yi r a

◆ STEP 3 changing the the Roman alphabet to the International Phonetic Alphabet

◆ Because after backmasking there are some syllables that cannot be pronounced in the Japanese syllabary, it is better to change them to International Phonetic Alphabet.

ɯ s a m yi a dz o g ɯ o t a g yi r a

Special case: Occlusive

Occlusive including
resistance ⟶ hold ⟶ release

◆ resistance: The airflow of a consonant is blocked
◆ hold: The airflow is bursting when it is releasing
◆ release: Blowing air[2]

resistance ⟵ hold ⟵ release

After backmasking, this process becomes an inhalation, so no pronunciation is required

◆ So plosives become non-plosives
Ex. と(to) ⟶ ど(do)

ɯ s a m yi a dz o g ɯ o d a g yi r a

plosives become non-plosives

	50音 (清音)	が・ざ・だ・ば・ぱ (濁音)	拗音 (清音)	拗音 (濁音)
母音	あ い う え お			
ローマ字	a i u e o			
IPA	a i ɯ e o			
	か き く け こ	が ぎ ぐ げ ご	きゃ きゅ きょ	ぎゃ ぎゅ ぎょ
ローマ字	ka ki ku ke ko	ga gi gu ge go	kya kyu kyo	gya gyu gyo
IPA	ka ki kɯ ke ko	ga gi gɯ ge go	kja kju kjo	gja gju gjo
IPA		ŋa ŋi ŋɯ ŋe ŋo	← 鼻濁音 →	ŋja ŋju ŋjo
	さ し す せ そ	ざ じ ず ぜ ぞ	しゃ しゅ しょ	じゃ じゅ じょ
ローマ字	sa shi su se so	za ji zu ze zo	sha shu sho	ja ju jo
IPA	sa ʃi sɯ se so	dza dʒi dzɯ dze dzo	ʃa ʃɯ ʃo	dʒa dʒɯ dʒo
	た ち つ て と	だ で ど	ちゃ ちゅ ちょ	
ローマ字	ta chi tsu te to	da de do	cha chu cho	
IPA	ta tʃi tsɯ te to	da de do	tʃa tʃɯ tʃo	
	な に ぬ ね の		にゃ にゅ にょ	
ローマ字	na ni nu ne no		nya nyu nyo	
IPA	na ɲi nɯ ne no		ɲa ɲɯ ɲo	
	は ひ ふ へ ほ	ば び ぶ べ ぼ	ひゃ ひゅ ひょ	びゃ びゅ びょ
ローマ字	ha hi fu he ho	ba bi bu be bo	hya hyu hyo	bya byu byo
IPA	ha ɕi ɸɯ he ho	ba bi bɯ be bo	ça çɯ ço	bja bjɯ bjo
		ぱ ぴ ぷ ぺ ぽ	← 半濁音 →	ぴゃ ぴゅ ぴょ
ローマ字		pa pi pu pe po		pya pyu pyo
IPA		pa pi pɯ pe po		pja pjɯ pjo
	ま み む め も		みゃ みゅ みょ	
ローマ字	ma mi mu me mo		mya myu myo	
IPA	ma mi mɯ me mo		mja mjɯ mjo	
	や ゆ よ			
ローマ字	ya yu yo			
IPA	ja jɯ yo			
	ら り る れ ろ		りゃ りゅ りょ	
ローマ字	ra ri ru re ro		rya ryu ryo	
IPA	ra ri rɯ re ro		r j a rjɯ r j o	
	わ を ん			
ローマ字	wa o n			
IPA	wa o N			

Prepared by Townsman on Feb. 25, 2010

References
[1]:https://zhuanlan.zhihu.com/p/115109163
[2]:http://phonetic-blog.blogspot.com/2009/11/unreleased.html

Conclusion

◆ By learning interesting backmasking, we have a better understanding of Japanese pronunciation structure and methods.
◆ **This method works for all languages, but for languages with more tones, such as Chinese, the backmasking process will become more complicated.**
◆ This method is also suitable for singing, you need to add a process of reversal of reverse the tone.

Comparing Information density of Japanese and English using "The Tale of Genji"

16 November 2022

1 Introduction

As introduced in lesson 3 of this course, Zipf's law can be applied to languages at a high accuracy. Predictions have been made based on this observation, such as the inverse relationship between frequency of a word and its contained information. For example, common words such as "the", "of" do not contain much information, meanwhile less common words like "linguistics" contain much more information.

In this conference poster, a theory of information density of each word in English and Japanese will be developed based on Zipf's law. Then these two densities will be compared together using a common piece of information, which is Chapter 1 of "The Tale of Genji"[1][2].

2 Methods

Form the observation based on Zipf's law above, it can be theoreticized that the amount of information contained in each word and in a set of N words, are respectively:

$$I = \frac{C}{f} \Rightarrow I_{\text{total}} = C \sum_{i=1}^{N} \frac{1}{f_i}$$

With occurrence frequency f taken from [3], a table of information contained in each word for English and Japanese can be constructed as a function of the constants C_E for English and C_J for Japanese.

To calculate C_E and C_J, the chapter 1 of "The Tale of Genji" [1][2] is used, assuming the same amount of information for both English and Japanese versions of the chapter, which are both defined as 1.

$$I_{\text{Ch.1, The Tale of Genji}} = C_E \sum_{i=1}^{N_E} \frac{1}{f_i^E} = C_J \sum_{i=1}^{N_J} \frac{1}{f_i^J} = 1$$

3 Result

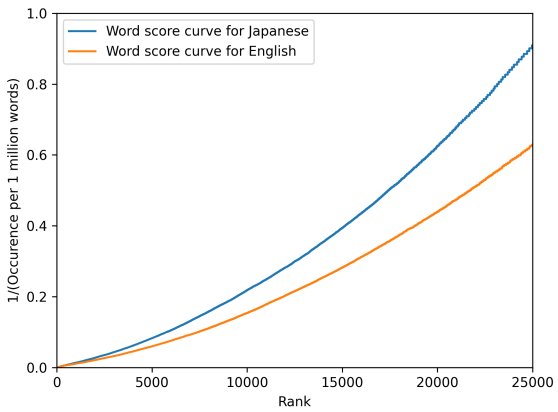


Figure 1: $\frac{1}{f}$ vs. Rank of words

Figure 1 shows the "word score" $\frac{1}{f}$ for Japanese is higher than in English for the first 25000 common words.

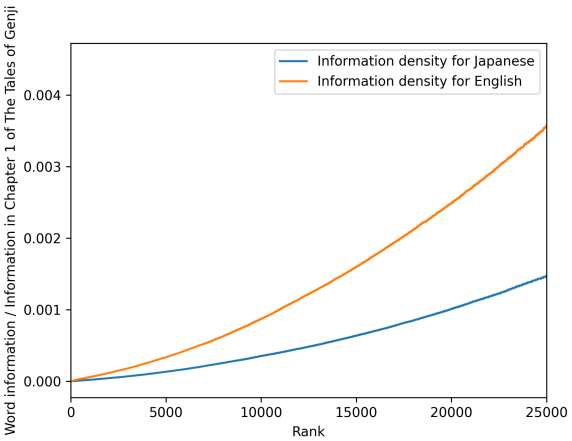


Figure 2: Information score

After adjusted using "The Tale of Genji", Figure 2 shows the information density of English is higher than Japanese.

4 Discussion

It could be misleading to just use the inverse of occurrence frequency of each word to be its contained information. If it were true, then the information in each word of Japanese would be higher than English for the most common words. However, our study found out that the reverse was true: English is higher than Japanese for information density.

5 Conclution

By using "The Tale of Genji" as a middleground for comparison, from the graphs, we can see that for the first 25000 common words, English contain more information than Japanese.

References

[1] M. Shikibu. (). The project gutenberg ebook of the tale of genji, [Online]. Available: `https://www.gutenberg.org/cache/epub/66057/pg66057.txt`.

[2] ——, (). Genji monogatari modernized, [Online]. Available: `http://jti.lib.virginia.edu/japanese/genji/modern.html`.

[3] U. o. L. Centre for Translation Studies. (). Use of corpora in translation studies, [Online]. Available: `http://corpus.leeds.ac.uk/list.html`.

Word Embeddings

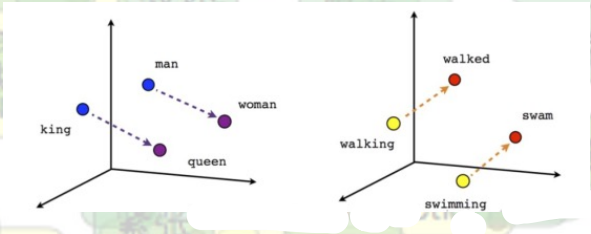
Nie Han 20M18544
Electrical and Electronic department, Tokyo Institute of Technology

1 . What is word embedding

How to describe word meaning to computer?

Code as vectors!

Make computer Understand human language!



Use linguistics to categorize similar topics!

Mathematical Expression

Word similarity of adjustable Nth Dimension

X: A matrix of all word vectors in the corpus
X_{i*}: word vector for the ith word in the dictionary
$M_n(X) = (XX^T)^n$: n-order similarity measure formula
$X^T X = Q\Lambda Q^T$: SVD (singular value decomposition)
$W = Q\Lambda^{\alpha}$: Linear Transformation Matrix
*The main purpose is to **adjust alpha** to get different similarity measures, and then you can display word similarity of different dimensions*

Task Solving

Semantic analogy: *solving problems like what is the word that is similar to France in the same sense as Berlin is similar to Germany?*
Syntactic analogy: *solving problems like what is the word that is similar to small in the same sense as biggest is similar to big?*

2 . Methods

- **Bag-of-words: one-hot, tf-idf, textrank, etc.**
- **Topic models: LSA (SVD), pLSA, LDA**
- **Fixed representations based on word vectors: word2vec, fastText, glove**
- **Dynamic representation based on word vectors: ELMO, GPT, bert**

3 . Examples

Case I: Capture word meanings by similarity[1]

Linguistic Perspective: word similarity

1. **Genuine similarity & Relatedness:** The former is also called functional similarity or just similarity such as *car* and *automobile* is true similarity; while *car* and *road* are associated, also known as associative similarity, topic similarity or domain similarity.
2. **Semantics & Syntax Dimensions:** like *sing* and *chant* are semantically similar, while *sing* and *singing* are syntactically similar.

4 . Morphology [2]

Morphology sensitive embedding : It treats the word *as a sum of character n-grams* representation.

"where" = ⟨wh, whe, her, ere, re⟩

Effective for morphologically rich languages like German, French, Spanish, Russian, Czech(Cs)

Attract the inflectional morphology & Repel the derivational morphology

Inflectional ↘
Derivational ↗

English	German	Italian
(discuss, discussed) (laugh, laughing)	(schottisch, schottischem) (damalige, damaligen)	(golfo, golfi) (minato, minata)
(dressed, undressed)	(stabil, unstabil)	(abitata, inabitato)

5 . Conclusions

Morphology makes Word Embedding more Effective

[1] Artetxe, Mikel, et al. "Uncovering divergent linguistic information in word embeddings with lessons for intrinsic and extrinsic evaluation." *arXiv preprint arXiv:1809.02094* (2018).
[2] Tanay Gahlot, . "Moving beyond the distributional model for word representation."

Difference between languages about perception toward Metaphor

Kaoru Yamamoto
Engineering Design Course, Tokyo Institute of Technology

1． Introduction

How does people deal with metaphor? Metaphor is the stretch of meaning. It have a strong cultural component, and metaphorical use of language is language creativity as its highest.
So the question is does the metaphorical impression differs if we translate in different languages which contains different culture?

2． Methods

Analysis of previous study. Research of Kumakura (熊倉2007), that compared difference between the perception toward metaphorical expression in Japanese and in English.
The experiment method is by showing Japanese and English metaphorical expression(37 expressions) to each language speaker(each consist of 4) and ask weather each expression is metaphor or not. The evaluation was done by questionnaire. (Fig1)

3． Result

Although both in Japanese and in English the perception toward metaphor were quite in common, there were some difference between Japanese and English speaker. (Table 1)

Q. Do you think these sentences are Metaphorical expression?

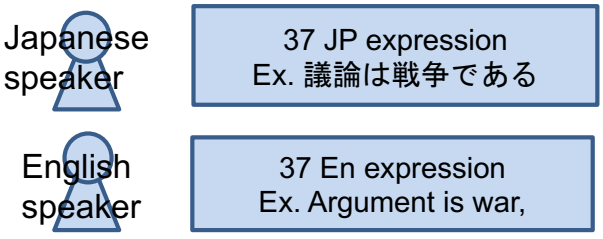


Fig. 1 ： Experiment

4． Discussion

From the result in test, we could say that if there are difference toward metaphor in different language, the verbal communication gives people different impression. For example, translation text. Even if author in certain country intend to use metaphorical expression, when the text is translated to other language, it may not be metaphorical in translated language reader.
However we should consider that the sample size was small in the experiment so we cannot define that the result is significant.

5． Conclusion

From the result, many Japanese expression gives same metaphorical impression in English when it is translated.
Further, Japanese have quite common perception toward metaphor by speakers compared to English.
However, we could not say with possibility.

Reference:
1.熊倉(2007)メタファーの日米比較：日常化したメタファーの考察,岩手大学英語教育論 9 11-32.
2.Fromkin et al(2003) : An Introduction to Language Seventh Edition, Wadsworth
3.吉岡(2021)：フィールド言語学者、凄もる, 創元社

Table 1: Result

Condition	Result number	Supplementary
Expressions graded the same in Japanese and English	22	
Expressions graded differently in Japanese and English	8	Number of Expressions that results were divided among the raters JP：2 EN：6

Analyzing ancient language or alien language with attention maps

Gao Tianlang/Tokyo Institute of Technology

Introduction

- In natural language processing field, deep learning models with the idea of Transformer is widely used. Transformers use the idea of "attention" train models and can show the weight of attention of each word in a sentence.
- For unknown ancient language or alien languages, the most difficult problem is that we have no ideas how the words are linked. Thus, the weight of the transformer could show whether two particular words are connected of not.
- Model: Famous NLP model "BERT" with dataset: WikiText-2

Result

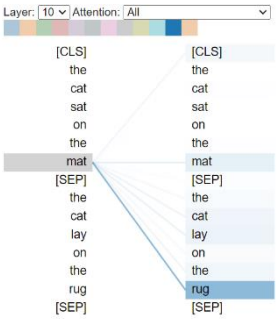
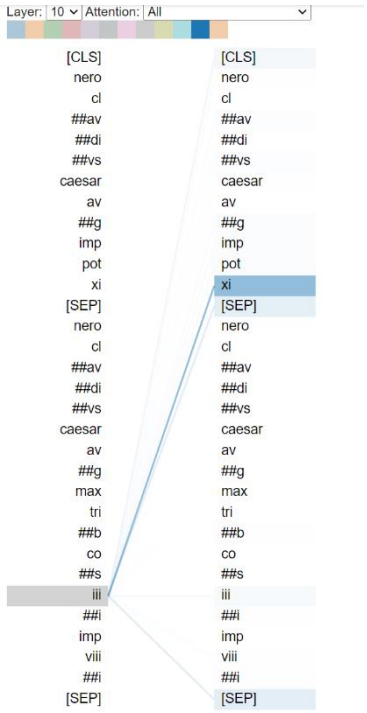


Fig.1 The attention map of two sentences. (The cat sat on the mat) and (The cat lay on the rug).



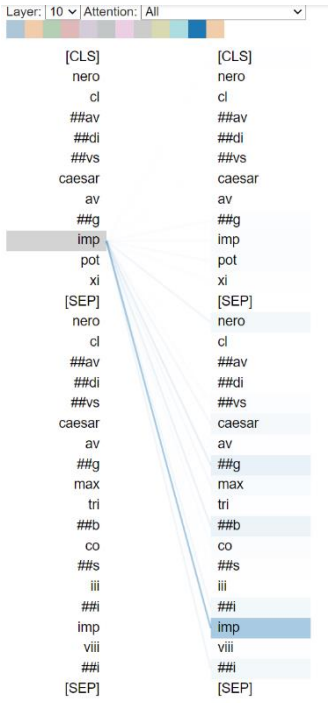
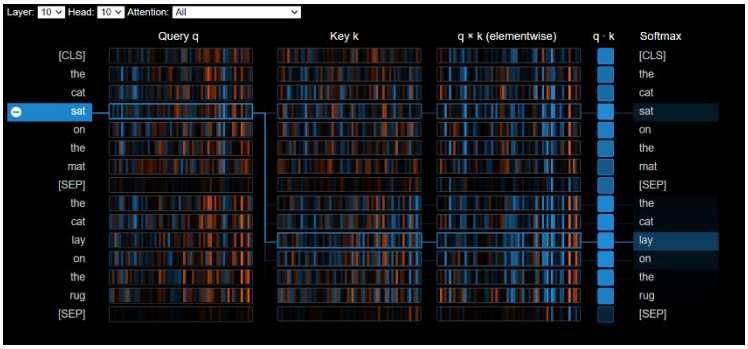


Fig.3 Ancient Latin, (NERO CLAVDIVS CAESAR AVG GERMANICVS IMP PONT MAX TRIB POT XI COS IIII IMP VIIII) whose meaning is (Nero Claudius Caesar Augustus, Imperator, Pontifex Maximus, in his 11th year of Tribunician Power, Consul four times, Imperator nine times). I separate it into two sentences, try to look up the attention map.



sentence_c = "Nero clavdis caesar avg imp pot xi"

sentence_d = "Nero clavdis caesar avg max trib cos iiii imp viiii"

Fig.2 Details of the BERT attention map

Future work

- BERT performs better when pretrained models are better, so we can use massive ancient languages as dataset to train the pretrained model. Then the attention map would have better result.
- The attention map has many other patterns to explore, for the pattern I showed, it is "the most related word in the other sentence".

The meaning expression of the emoji (Smiling Face with Open Mouth and Cold Sweat) in different cultures.

Introduction

I have collected some application scenarios of this emoji on twitter and selected four classical application scenarios.

I analyzed the meaning expressed by this emoji in each paragraph.

Last I compared these meaning to figure out weather these meaning are same.

figure1 Smiling Face with Open Mouth and Cold

Application scenarios and analysis

Tweets 1

In tweets 1, the twitter said the horse shoudn't take sand bath and add the emoji. But the context expression without impatient,so the emoji express the twitter feeling ovely but helpless about this behavior.

Emotional Tendency: Positive

Tweets 2

In tweets 2, the twitter said the leaflet ware printed a lot in school. Follow the sentence, the emoji was added.Expressed the dissatisfaction with this behavior.

Emotional Tendency: Negative

Tweets 3

In tweets 3, the twitter said the Simons sent Shark a Mad DM.Then the emoji is added. The emoji expressed the shame on this behavior.

Emotional Tendency: Negative

Tweets 4

In tweets 4, the twitter said the Curry's achievements hit a lot of anti-fans in the face. Expressed the shame on anti-fan's comments.

Emotional Tendency: Negative

Conclusion

Japanese culture use the emoji to express both positive and negative emotion.

American culture use the emoji to express only negative emotion.