Linguistics E conference

2024.01.31

January 30, 2024

Viewing Puns from Linguistic Perspective

Garry Pranata Kusuma
Electrical and Electronic Engineering, Tokyo Institute of Technology

1. Introduction

Language is a complex and dynamic system that often engages in playful and creative expressions. Among them, puns are a unique form of wordplay that hinges on the exploitation of multiple meanings, homophony, and clever manipulation of language. This study seeks to classify puns based on how they manipulate language into a medium of jokes.

2. Methods

A set of English puns is gathered from various sites (e.g., Reddit, news websites) to ensure diversity in the dataset. The selected puns are dissected to identify the structural elements. Criteria for classification were established based on common linguistic features found in puns. Categories were defined according to the nature of wordplay.

3. Result

Based on the 50 English puns, there are 5 categories. They are summarized in Table 1.

Table 1. Categories of Puns

Category	Example
Homonymy	Why do Buddhist monks avoid sending word documents? They're supposed to avoid attachments.
Homophony	Want to hear a joke about paper? Nevermind it's tearable.
Paronymy	What do you call a fake noodle? An Impasta.
Context	Why should you stay away from artists? They're sketchy.
Phrasal Ambiguity	I'm afraid of speed bumps, but I'm slowly getting over it.

4. Discussion

The 5 categories presented are the result of the author's observation. To elaborate some, in homophony-based puns, the word which becomes the object is replaced with a word with identical pronunciation. On the other hand, paronymy-based puns use a word (or a combination of words) to form a new word with similar pronunciation. Context-based puns use words which only seem funny given the context of the sentence.

Some categories of puns are easier to understand when written. Homophony-based puns are hard to understand as oral jokes since it is difficult to tell where the joke is.

In her paper, Giorgadze [1] came up with 3 categories:
1. Lexical-Semantic Pun
Lexical-Semantic Pun results from similar words whether in pronunciation or spelling. Homonymy, homophony, paronymy fall into this category.
2. Structucal-Semantic Pun
It arises a phrase which can be parsed in more than a way. In this study, puns in this category are put under homonymy category.
3. Structural-Syntatic Pun
This comes from ambiguity of words or phrases.

5. Conclusion

Based on the gathered dataset, puns can be categorized into homonymy-based, homophony-based, paronymy-based, context-based, phrasal ambiguity-based. Some categories might be more popular as written puns since some are more difficult to understand orally.

Sources: https://justpaste.it/ego3o

[1] Giorgadze, Meri (2014) *Linguistic Features of Pun, Its Typology and Classification*, European Scientific Journal, 2, 273-275.

A Study on Chinses Grammar Changes By Comparasion of the Distribution of Quantifier between Classical Chinese and Mordern Chinese Novel

CHOEI
EE, School of Engineering, Tokyo Insttute of Technology

I. Introduction

Benford's law, the law of anomalous numbers, or the first-digit law, is an observation that in many real-life sets of numerical data, the leading digit is likely to be small. On the other hand, quantifiers are a kind of word to express number, in Chinese language, the quantifiers are not only used to log numbers, but also play other roles. This study will analyse how classic chinese and mordern chinese satisfy the Benford's law, and explain the exceptions occurs.

II. Methods

One hand, We collect the Chinese character of 1~9（一,二,三,四,五,六,七,八,九）from the dataset consists of 全唐詩, 史記, 漢書, 舊唐書 as the classical chinese materials, and *The Three-Body Problem* by Liu Cixin as the mordern chinese material. Materials written at multiple eras and types will be used to obtain a macroscopic result. On the other hand, we calculate the Benford's law in 10 digits as follows, for $d \in [1, 9]$

$$P(d) = \log_{10}\left(1 + \frac{1}{d}\right) \tag{1}$$

then we draw the error of our data from the Benford's law as follows

$$E(d) = \|F(d) - P(d)\|^2 \tag{2}$$

III. Result

Collecting the data and drawing (2) for two datasets as Figure 1, which shows that the data of classical chinese basicly satisfies the Benford's law well, but for the case of mordern novel, the frequency that 1 occurs doesn't satisfy that properly.

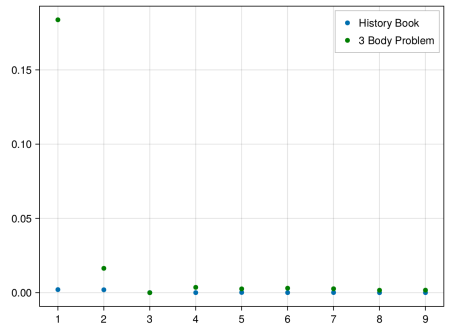


Figure 1: Errors from Benford's law

IV. Discussion

For the case of number 1 in *the three-body problem*, it shows a abnormally distincts from the Benford's law. this is because the phenomenon called "歐化" which is taking place in Mordern Chinese. that means mordern chinese grammar is affected by the Grammar of European Languages. In Chinese there is no conception of article.Affecting by Word-by-word translation that does not conform to traditional Chinese habits, people started to use "一個"(literally means "one") as the indefinite artical.

V. Conclusion

By statistical analysis of this view, we learned that not only the vocabulary but also the grammar could deeply affected by other language. And the similar phenomenon can be observed in Japanese as well: $1 + 1$ is expressed as "一 足す 一" which doesn't match the regular Japanese grammar.

References

Quantifying the Border of "Language Death"

Nathan Van Alstine | vanalstn@elsi.jp
M2 Student - ELSI, Department of Earth and Planetary Sciences, Tokyo Institute of Technology

Introduction

The death of a language is generally defined as by when a language no longer has any native speakers. While this may represent it self as a clear definition, it remains unclear at what stage a language is considered different enough from its historical counterparts to be unrecognizable to modern speakers. That is - at what stage is a language deliniated enough to consider it a "different" language. This presents itself with English in its previous stage of **Old English**, otherwise known as Anglo-Saxon (5th centure) which is overall undecipherable by modern English speakers except for some characters and words. This similarly presents itself with Old Japanese, or Jodai Nihongo (8th century), which is one of the oldest forms of the language and similarly nearly audibly unrecognizable to modern Japanese. The Global Language Monitor (GLM), and Google Corpus similarly estimate a little over 1,000,000 words in the English language, and that anywhere from 1,000-4,000 are added each year. Japanese is counted less at around 500,000 words with around 3,500 new words and 1,000 words erased per year (word erasure, limited database , or counting differences may make up this large difference). With this rate of change in mind, where does a quantifiable deliniation occur that a language becomes a "version" that is no longer understandable to modern native speakers?

Methods

-Different datasets were examined from multiple language databases
-English data pulled from:
  -Global Language Monitor (GLM)
  -Google Corpus/Google Books corpora (& Harvard Analysis)
-Japanese data pulled from:
  -日本国語大辞典 (Japan National Dictionary)
  -国立国語研究所 (NINJAL)
-Average word addition is taken from Google Corpus (EN) and Kojien changes per edition (JP)

Results

-Word counts are shown with lower to higher estimates for both languages
-Eras are plotted and were understood to be matched with different quantities of word change
-New Eras in English occur every 450-500 years, or every 1,187,500 (average) new words, and 325-375 years or every 787,500 (average) new words for Japanese respectively.

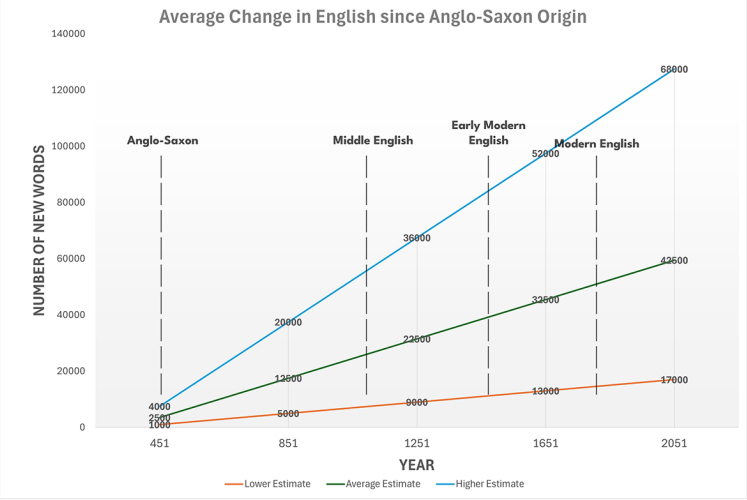


Figure 1. Lower, Average, and Higher estimates of new words added per year for English since its origins as Anglo-Saxon in the year 451. Plotted together with eras of the English language. Scale is 100x.

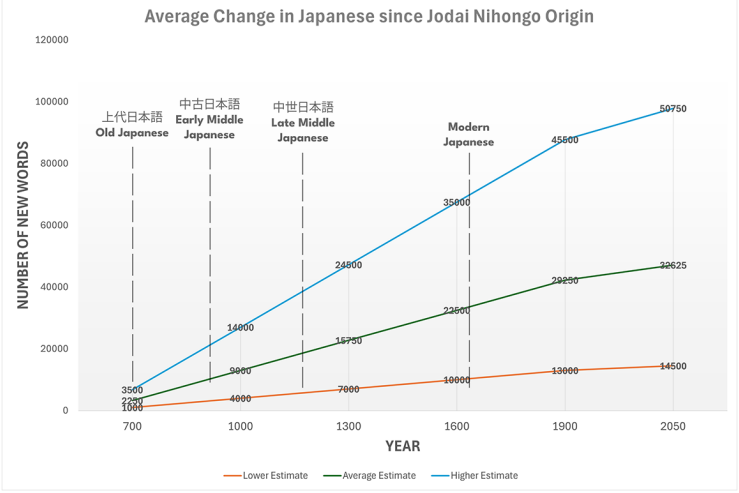


Figure 2. Lower, Average, and Higher estimates of new words added per year for Japanese since its origins as Jodai Nihongo in the 8th century. Plotted together with different eras of the Japanese Language. Scale is 100x

Discussion

It is difficult to qualify where the modern version of our languages are 'different enough' for an older version to be dead. According to Figure 1 and 2, new eras occur roughly when new word count has slightly surpassed the previous count - 1 million~ for English and 500,000~ for Japanese. Dynamic evolution of words in a language generally stems from previously existing words such as threw derivation, compounding, repurposing, eponyms, and more. This genetic relation may explain why even though a few hundred years may make the difference in over 1,000,000 words, English and Japanese of 300-400 years ago is still easily understandable to native speakers. Furthermore, as word change does not directly ascribe grammatical change, which is much slower, this may further explain familiarty.

Conclusion

Eras of a language have different lengths, likely due to cultural shifts or lack thereof occuring at the time. Anglo-Saxon to the middle ages is large as this was the "Dark Ages" before the Rennaisance. Old Japanese to Modern Japanese is rather quick possible due to warring periods, following the relatively more stable Edo Era. Modern language change now faces a new variable, which is the advent of the internet, allowing for more rapid introduction of cultural exchange, words, and "internet language/meme speak". Generational gaps are immediately observed presently with the introduction of new words in English such as "skibidi" or "rizz" which is generally used within Gen Z and Alpha in America, or アレ(阪神) or ナートゥ in Japanese which may not be understood by older generations that are not so commonly on the internet. Furthermore, Japanification of English, or Englification of Japanese is a new phenomenon as well as cultures intermerge - INTERNET YAMERO is a very popular case of this as one of the most popular music videos of Japan in 2023 and representitive of a youth culture shift in Japan. With these cultural shifts and word quantities exemplified, a "new era" of both languages should be expected around the early 2100s.

References

Data Provided by:
https://www.english-corpora.org/googlebooks/#
https://repository.ninjal.ac.jp/
https://kotobank.jp/word/%E6%97%A5%E6%9C%AC%E5%9B%BD%E8%AA%9E%E5%A4%A7%E8%BE%9E%E5%85%B8-679303
https://languagemonitor.com/

Understanding れる・られる from the middle voice

Akazawa Shunpei
Physics, Tokyo Institute of Technology

１． Introduction

Why we use the same れる・られる as “passive” and “spontaneous” in Japanese?

２． Methods

Focusing on the middle voice, and Japanese ancient grammar, and extracting the essence common to these two meanings.

３． Result

Both passive and spontaneous mean “some action materializes within the agent”, which the exact meaning of the middle voice.

These two can be understood in the same framework.

４． Discussion

れる・られる can also mean “potential/capability (できる)” and “honorific (尊敬)”. We could integrate them into the frame of the middle voice.

５． Conclusion

Why れる・られる are used as “passive” and “spontaneous” can be understood in a unified way by considering the middle voice.

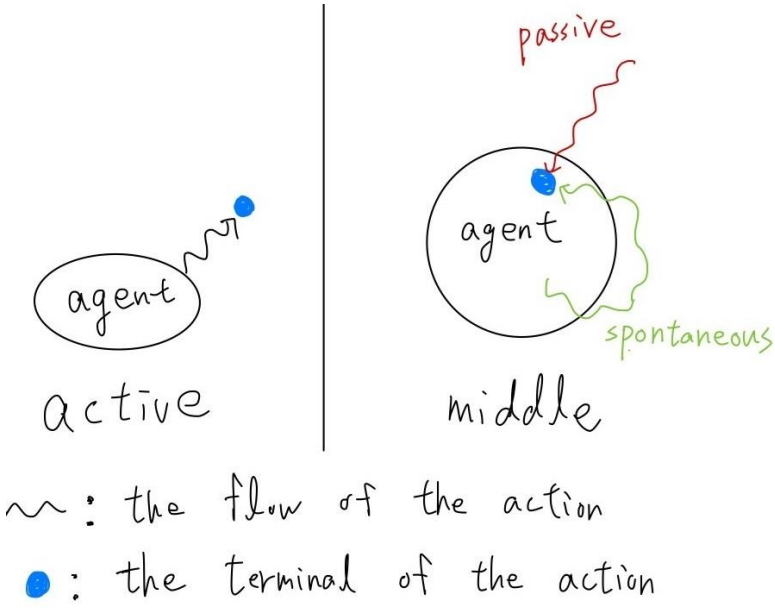


Fig: the active voice and the middle voice

Reference:
[1]國分功一郎, 中動態の世界 意志と責任の考古学, 医学書院, 2017 p177-p191
[2]金谷 武洋,日本語文法の謎を解く—「ある」日本語と「する」英語,ちくま新書

The important component in funny answers for Japanese Ohgiri

Kanamori Yuki
Civil Engineering, Tokyo Institute of Technology

1. Introduction

What component in Ohgiri answers makes them funny ?
Ex Q) Give me an example sentence with “be about to”.
A. My dad is about to use the front teeth.

2. Methods

- Classification by XGBoost model[1)]

--Questionnaire--

- 20 Ohgiri answers
- Rate the followings on a scale of 1 to 5
- Add one feature

<Output>
- Is the answer funny?

<Input>
- Suitability (F1)
- Ease of imagination (F2)
- Novelty (F3)
- Emotional (F4)
- Number of letters (F5)

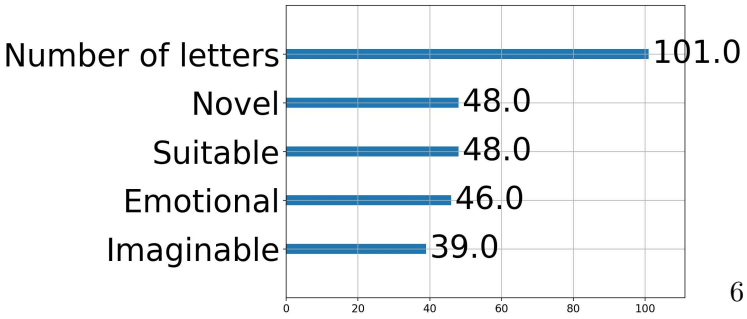


Fig.1: Feature importance of components

3. Result and Discussion

- Number of letters most connects to the amusingness (for various combinations of Train & validation data)

-After 10-fold CV-

- Models with only the number of letters have the worst prediction accuracy.

4. Conclusion

For funny answers:
1. Consider several components
2. Think about the number of letters as the 1st priority

Reference:
1） Yuki Nakagawa et. al. (2019), Crowdsourcing analysis of the components of amusingness in Japanese Oogiri, Papers from annual conference of the association for Natural Language Processing, 25th, pp.233-236.

Table 1: log loss mean for test data of models

Input	log loss mean
F1	1.388
F2	1.351
F3	1.422
F4	1.503
F5	1.533
F1 F2 F3 F4 F5	1.339

Quantitative Research on the Predictability of Hanzi Reading

Masaki Nagata
Dept. of Mathematical and Computing Science,
Tokyo Institute of Technology

1. Introduction: Background

- It is widely known that the languages used in countries within the Sinosphere, a.k.a. the Chinese cultural sphere (漢字文化圏), share a substantial amount of vocabulary derived from Chinese
- Each language has corresponding readings for each Chinese character (漢字, hanzi)
- Very often, you can guess how to read a hanzi in the language from knowledge of other languages

2. Introduction: Question

How accurate can we guess the reading of given hanzi in specific language, given readings in other languages?
More specifically: given 3-tuple of reading of hanzi with one of them unknown, can we fill the blank accurately?

金 : (kin, jīn, geum)

禁 : (kin, ?, geum)

→ How does 禁 read in Chinese?

Fig. 1. Example: 金 and 禁
The 3-tuple in the figure represents the readings of the hanzi in each three languages: Japanese, Chinese, and Korean

3. Method

Use the public database and library of hanzi reading data to predict the reading of hanzi and calculate the accuracy of prediction.

- Database: DB_elm (1,000 Kanjis learned in Japanese elementary school), DB_adv (1,031 Kanjis commonly used in Japan with DB_elm excluded), DB_all (DB_elm + DB_adv, 2,031 Kanjis in total)

4. Result

Table. 1. Prediction accuracy (Trained with DB_elm)
Row: The language to predict the reading of
Column: The database used for testing

DB_elm	DB_elm	DB_adv	DB_all
Japanese	85.9%	36.5%	60.8%
Korean	90.7%	36.8%	63.3%
Chinese	82.6%	34.1%	58.0%

Table. 2. Prediction accuracy (Predicts Japanese)
Row: The database used for training
Column: The database used for testing

Japanese	DB_elm	DB_adv	DB_all
DB_elm	85.9%	36.5%	60.8%
DB_adv	39.5%	89.0%	64.6%
DB_all	82.1%	84.6%	83.4%

5. Discussion & Conclusion

- The accuracy of extrapolation (predicting hanzi not in training data) is around 40%
- Accuracy can be improved with more uniformly sampled database
- Further research can be done by comparing the performance between the prediction using two languages (this research) and that using only one

Reference

[1] Wikipedia (2022) 常用漢字一覧, Webpage link.
[2] S. Byeon (2021) hanja: 한자-한글 변환 라이브러리, GitHub link.
[3] T. Roten (2023) Dragon Mapper, GitHub link.

Template of Conference Poster

Zhang Xiulin 張修麟
Department of Social and Human Sciences, Tokyo Institute of Technology

1． Introduction

The principle of least effort argued by Zipf has become one of the most famous linguistic rule, and polysemy had been conceived as one of its consequence(Cancho et al., 2003). In this study, I tried to prove how could the principle of least effort lead to polysemy, by game-theoretic simulations.

2． Methods

Philosophers and linguists has developed signaling game as a powerful way of modelling the emergence and evolution of language. Within a signaling game with reinforcement learning, the senders and receivers of signals are inclined to reach a signaling system where perfect information is finally communicated, which is motivated by the payoff of successful communication. However, sometimes the learners fall into partial pooling equilibriums where polysemy of certain signals exist. My simulation will show that the weight of Roth–Erev reinforcement learning is responsible for these situations

3 . Result

The result is shown on the right graph: the bigger the initial weight is, the more likely the learning processes will fall into situations of polysemy.
The result is first found by Skyrms(2010) but he didn't realize its significance to do with polysemy and the principle of least effort. I made 1000 trials by 10000 iterations.

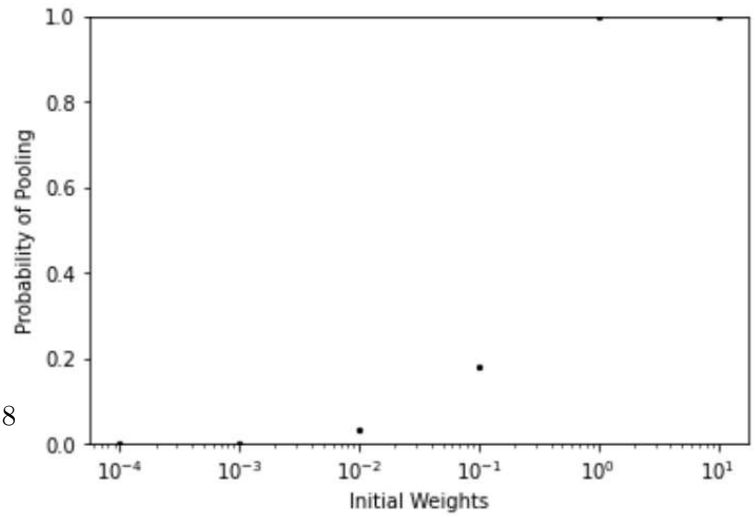
4 . Discussion

The weight of Roth–Erev reinforcement learning represents how difficult it is to learn something new. When weight is large, the learner is then overwhelmed by the reinforcement mechanism and is reluctant to adjust his pattern of signal use. According to the principle of least effort, polysemy can reduce the increasing of new words to some extent, so that the learner gets rid of the burden to learn more.

5． Conclusion

The study have shown how the principle of least effort could result in polysemy, by the simulation results.

Reference:
Cancho R F I, Solé R V. Least effort and the origins of scaling in human language[J]. Proceedings of the National Academy of Sciences, 2003, 100(3): 788-791.
Skyrms, Brian. Siganls, 2010, 97.



The Impact of Language Skills on Friendship Building during the First Meeting

Yuan Yang
Global Engineering for Development, Environment and Society
Tokyo Institute of Technology

1． Introduction

How language skills influence friendship formation in initial meetings. To explore preferences regarding communication styles, humor, personal experiences, and the significance of clarity and kindness in conversations.

2． Methods

Participant Demographics, Scenario Setting, Preference Assessment, Friendship Factors, Depth of Conversation, Articulation and Friendship, Personal Sharing Impact, Interest and Connection, Humor's Role, Talking vs. Listening, Boasting Impact, Patience and Listening

3． Result

Gender
Among men, 20 chose scenario A and 9 chose scenario B.
Among women, 14 chose scenario A and 4 chose scenario B.
Nationality
Among the Chinese, 20 people chose Scenario A and 11 chose Scenario B.
Among Vietnamese, 4 people chose scenario A and 2 people chose scenario B.
Among Japanese, 4 people chose scenario A-
Among British people, 2 people chose scenario A
Among Canadians, 2 chose scenario A
Among Americans, 2 people chose scenario A

4． Discussion

In the gender comparison, men are more likely to choose Scenario A, that is, they like to interact with people who are direct, confident, and likely to brag, while women are more likely to choose Scenario B, that is, they are more willing to interact with people who are patient, listening, and speak kindly. people build friendships.
Among the Chinese and Vietnamese samples, more people chose scenario A, that is, they preferred straightforward and confident communication. Among the samples from Japan, the United Kingdom, Canada and the United States, more people chose Scenario A, but the samples are small, and the results still need to be interpreted with caution.

5． Conclusion

Cultural diversity shapes communication preferences in friendship. Gender and age influence priorities, with young men valuing directness and humor, and young women emphasizing kindness and patience. Universal preferences include respect, kindness, patience, and genuine interest in personal experiences, reflecting a shared human inclination towards meaningful communication.

Information Transmission Efficiency during Dictation, 3-Language Survey

Jacob Lindahl
Mathematics & Computing Science, Tokyo Institute of Technology

1. Introduction

We continue the research from the previous quarter. Instead of surveying many languages across a few texts, we survey a few languages (English, Russian, Japanese) across many texts.

2. Method

The primary data sources for this survey were well-known, classic, written works in various languages, with target language translations having audio versions available online.

3. Results

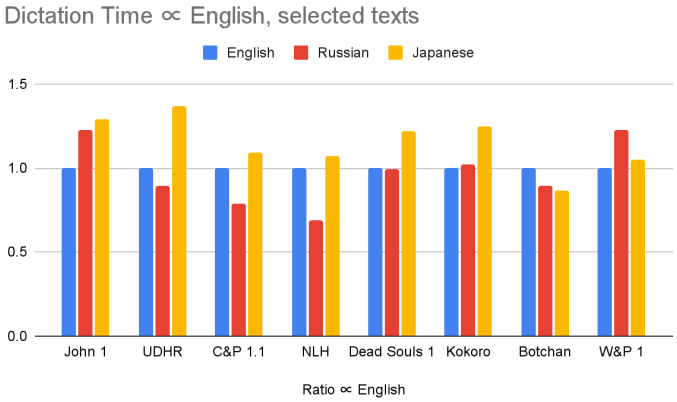
We discover a small pattern across all sampled texts: Russian takes about 97% of the time of English, and Japanese 115% the time of English, to dictate the corresponding translated text. The median proportion follows $-\ 0.011|P|\ +\ 1.37$, negatively correlated with the cardinality of the phonemic inventory.

4. Discussion

Its prolific authors have established Russian as a legendary progenitor of classic literature, often credited for being precise, colorful, and descriptive. The stereotype of information-dense Russian vocabulary is re-buttressed by the results of this survey. Japanese requires more time than the other members of this survey. Although other sources credit Japanese as one of the fastest-spoken languages, the dictation cadence observed does not bear this out.

5. Conclusion

Russian is a slightly more efficient spoken language than English, and significantly more so than Japanese, when dictating text.



References.

Coupé, Christophe, Yoon Mi Oh, Dan Dediu, and François Pellegrino. "Different Languages, Similar Encoding Efficiency: Comparable Information Rates across the Human Communicative Niche." *Science Advances* 5, no. 9 (September 4, 2019): eaaw2594. https://doi.org/10.1126/sciadv.aaw2594.

Horn, Heather. "Russia Has Great Literature--Here's Why." *The Atlantic* (blog), December 31, 2010. https://www.theatlantic.com/culture/archive/2010/12/russia-has-great-literature-here-s-why/339282/.

No Longer Human - Osamu Dazai, 2022. https://www.youtube.com/watch?v=zo51SMB54Vk.

"Universal Human Rights Initiative - YouTube." Accessed January 29, 2024. https://www.youtube.com/@UniversalHumanRightsInitiative.

Van Kuyk, Steven, W. Bastiaan Kleijn, and Richard C. Hendriks. "On the Information Rate of Speech Communication." In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5625–29. New Orleans, LA: IEEE, 2017. https://doi.org/10.1109/ICASSP.2017.7953233.

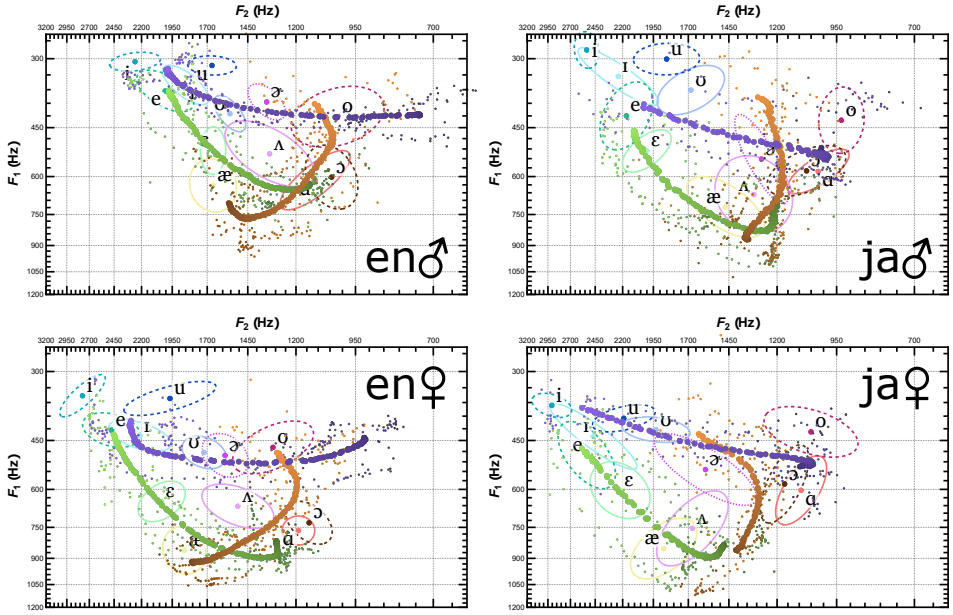
Аудиокниги Клуб - Слушаем Онлайн! "Дадзай Осаму - Исповедь Неполноценного Человека," May 2, 2022. https://akniga.org/dadzay-osamu-ispoved-nepolnocennogo-cheloveka.

【朗読】太宰治『人間失格』語り：西村俊彦, 2019. https://www.youtube.com/watch?v=Kl3EZTP8FEQ.

Locally Looking at Diphthongs

A Comparative Study of English L1 and Japanese L1 Speakers

Department of Earth and Planetary Sciences, Tokyo Institute of Technology
Austin Taylor



Plots in formant space of the standard English diphthongs
aɪ, aʊ, ɔɪ

The vowel progresses from the dark to saturated, normalized for total length

Small points are the individual formant data captured from recording

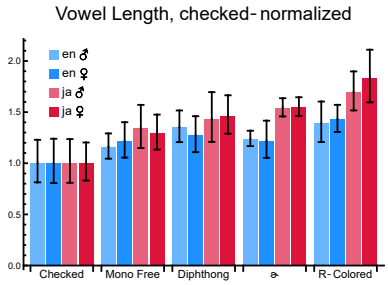
Large points are a moving average, with a 25% window length

Background

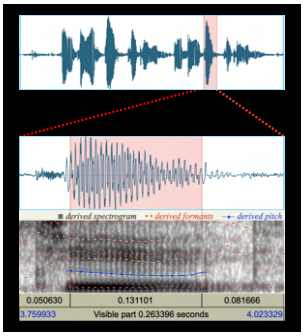
- Previously, I compared both vowel lengths and formants of English's canonical monophthongs
- We found that Japanese L1 speakers had overlapping formants, and varied length more
- **This time, I'd like to see if there's any major difference between the major diphthongs between Ja and En natives**

Vowel Lengths

Both groups tend to show diphthongs around 1.5 times the length of the shortest vowels



Methodology



4 groups of 3 speakers

19 English vowels recorded

Vowels are analyzed with Praat, capturing length and taking formant data across the length

Data is normalized for speaker's lengths and statistics are taken from the logs of these values

Discussion

There are some differences in the path that the diphthongs take between the language groups. However, all 3 are clearly defined and unlikely to be confused for one another.

The data here is highly qualitative. It would be nice to have some sort of quantitative model that could be more rigorously tested.

It would be nice to look also at r-colored vowels, which are diphthongal and have F3 variance. e and o would also make for good candidates.

Acknowledgements

Yazawa et al., *Spectral and temporal implementation of Japanese speakers' English vowel categories: A corpus-based study*, 2023
Kato et al., *Analysis of L2 English Vowel Production by Native Japanese Children in Domestic Elementary School*, 2019.
Oh et al., *A one-year longitudinal study of English and Japanese vowel production by Japanese adults and children in an English-speaking setting*, 2011
Paul Boersma and David Weenik, Universiteit van Amsterdam for producing and offereng Praat on GPL version 3.

東京工業大学 地球生命研究所