# A study on individual differences in contemporary Japanese colloquial expressions

Griffen Schwiesow[1,a)] Hilofumi Yamamoto[2,b)]

**Abstract:** This paper addresses the issue of individual variety in colloquial speech of contemporary Japanese. After conducting short interviews with native Japanese speakers across a variety of controlled groups (such as age and gender), the expressions and speech styles used by each individual were analyzed and the variances were compared to determine their significance. These findings contribute to automatic Japanese sentence recognition systems.

## 1. Introduction

This paper addresses the differences in contemporary Japanese expressions between colloquial and formal speech styles. Japanese expressions may differ across the language medium, such as speech or writing, as well as the impact demographics make on these differences. To clarify this, we will examine if the word order, type of tokens, and use of interjections are consistent across spoken and written Japanese.

## 2. Methods

We will attempt to extract colloquial patterns of Japanese phrases. There will be three major steps involved in this research: data collection, data parsing, and data analysis.

We will collect three types of data. Every time we collect a data sample, we will ask the participants to submit their answers to the same three questions to an online form (**Table 1**). However, the way in which we will request the participants to answer will vary: 1) type answers as a written text; 2) record answers in formal style, and then transcribe the recording; 3) speak answers in casual style into voice-recognition software. We asked these three specific questions to address the cognitive and affective aspects of language [3].

**Table 1** Questions asked to research participants

| No. | Type | Question sentence |
|---|---|---|
| 1 | Fact | What did you do in today's class?<br>*Kyō no jugyō de nani wo shimashitaka* |
| 2 | Cognition | What did you understand in today's class?<br>*Kyō no jugyō de nani ga wakarimashitaka* |
| 3 | Affection | What did you think was interesting in today's class?<br>*Kyō no jugyō de nani ga omoshirokatta desuka* |

We will examine the data in terms of the language medium (written or spoken), formality, and the gender of the speaker. After collecting the data, we will use MeCab, an open-source text segmentation library, to tokenize the samples and parse the responses into tokens. We will eliminate nouns and verbs in order to focus on the colloquial utterances produced in Japanese speech and then classify the tokens according to their appearance in written and spoken data samples [4]. From this, we will gather multiple sets of tokens, such as "written- or spoken-only" tokens which will allow us to analyze colloquial speech patterns.

## 3. Results

The steps outlined above allowed us to look at various aspects of the language and we were able to draw three main results from the analysis. We found that each language medium had its own unique tokens, but that the co-ocurrence patterns of tokens in each language medium revealed more about the language than the unique tokens alone. Further, we looked at how disparate the tokens were for each gender by finding the tokens unique to each gender.

### 3.1 Unique tokens

After classifying tokens and excluding nouns and verbs, we found the most used tokens for each category (**Table 2**).

**Table 2** Top 10 most common tokens unique to each language medium

| Rank | Spoken | Written |
|---|---|---|
| 1 | えー | あり |
| 2 | とか | を通して |
| 3 | って | に関する |
| 4 | っていう | なけれ |
| 5 | けど | のみ |
| 6 | だけ | すみません |
| 7 | えーと | 面白 |
| 8 | たい | そこで |
| 9 | … | ものの |
| 10 | まず | 興味深かっ |

The most common tokens found in spoken Japanese were fillers and casual variants of grammatical constructs (for example, *toka* appears in spoken Japanese for which the formal equiv-

[1] University of Washington
4014 University Way NE, Seattle, WA 98105-6203
[2] Tokyo Institute of Technology
W1-9 2-12-1 Ōokayama, Meguro, Tokyo 152–8550, Japan
[a)] griffens@uw.edu
[b)] yamagen@ila.titech.ac.jp

alent is *nado*; meaning “etcetera”). However, in written and formal Japanese, fillers and other grammatical constructs were less common. Instead, content words, such as verbs and adjectives, appeared as the most common tokens.

### 3.2 Co-occurrences

We examined the most common co-occurrences of tokens in Japanese phrases, which are described in **Table 3** below.

**Table 3** Top 10 cooccurrences unique to each language type

| | Spoken | | Written | |
|---|---|---|---|---|
| Rank | Frequency | Pattern | Frequency | Pattern |
| 1 | 15 | とか–し | 2 | あり–ものの |
| 2 | 14 | ん–し | 2 | のみ–あり |
| 3 | 14 | えー–し | 2 | に関する–あり |
| 4 | 13 | まず–し | 2 | なかなか–あり |
| 5 | 13 | けど–し | 1 | そこで–面白 |
| 6 | 13 | って–し | 1 | ものの–そこで |
| 7 | 13 | だけ–し | 1 | 興味深かっ–面白 |
| 8 | 12 | とか–ん | 1 | あり–そこで |
| 9 | 12 | たい–し | 1 | に関する–そこで |
| 10 | 11 | とか–けど | 1 | に関する–ものの |

In the spoken Japanese samples, the token *shi* occurred in eight of the top 10 most frequent coocurrences. *shi* is used to indicate an enumeration or a list and can be used to end a sentence similarly to *toka*, meaning “etcetera.” However, in the written samples, *ari* and *sokode* were the most commonly seen tokens in co-ocurrences.

### 3.3 Gender differences

Finally, in **Table 4**, we can see the tokens unique to gender. Since we already know that various cues, such as vowel length and voice tone, give indications to a speaker’s age [2], we wanted to explore if there was lexical difference between male and female speakers.

**Table 4** Top 10 most common tokens unique to each gender

| Rank | Male | Female |
|---|---|---|
| 1 | だっ | すごく |
| 2 | ··· | あ |
| 3 | より | なあ |
| 4 | ず | 結構 |
| 5 | 高い | を通して |
| 6 | どういう | |
| 7 | 例えば | |
| 8 | とても | |
| 9 | あり | |
| 10 | えーっと | |

These findings show that there is little lexical difference between men and women.

### 3.4 Reflecting on the experience

As part of our research, we had the participants reflect on the difference in their writing and speech. A common theme found in their reflections was that when they speak, interjections and other filler words without a clear meaning are used. Further, the word order of spoken Japanese appeared to fluctuate while the written phrases followed the clearly defined grammatical rules. However, the fluctuation of word order and insertion of filler words made speech sound and flow more naturally. One participant commented, “... words that are not directly related to the content, such as ‘*ē*’ or ‘*toka*’, are often inserted in the spoken language, but I felt that they make the dialogue as communication natural.”

## 4. Discussion

Our findings suggest that there are clear differences between the written and spoken lexicon but the differences in lexicon across gender is less pronounced. According to the participant reflections, we observed that people are not often acutely aware of how their language differs when they change language mediums; even when people are speaking in a more polite way, they still ended up using expressions that are not normally considered as polite language.

### 4.1 Fillers are used in spoken word

While the token *shi* was not one of the most common tokens on its own, it appeared in eight of the top 10 co-ocurrence patterns. Many participants commented that words such as *shi* and *toka* were used without a solid meaning but to make time for thinking while speaking. These filler words also show how dynamic the word order of Japanese sentences can be when spoken. By inserting fillers such as *shi* in a sentence, the Japanese sentence can end, leaving the listener to understand contextually what the speaker is saying.

### 4.2 Word order is not strict in spoken word

While Japanese is generally considered an SOV (subject-object-verb) language, the use of fillers helped speakers adjust this order to create sentences by saying what came to their head first. A participant noted in their reflection that, “... in spoken language, even predicates, which are usually at the end of sentences, are disorganized.”

### 4.3 Suspended verbs are used in written text

In written Japanese, especially concerning formal works, the commonly used *te-form*, a verb conjugation used to connect sentences, is replaced with the suspended form, *ren’yōchūshihō* [1]. This is exemplified by the prevalence of *ari* in the written samples, which is the suspended form of the verb *aru* (meaning “to be”).

## 5. Conclusion

After gathering samples of spoken and written Japanese from native speakers, the data uncovered clear lexical differences between spoken and written Japanese, as well as casual and formal styles. Further, the word order of Japanese is less strict when spoken.

### References

[1] Kitajo, M.: Gendai nihongo dōshi no renyō-kei to te-kei no danwa kinou no chigai, *Nihon gengo gakkai dai 114-kai daikai* (1997).

[2] Ptacek, P. H. and Sander, E. K.: Age Recognition from Voice, *Journal of Speech and Hearing Research*, Vol. 9, No. 2, pp. 273–277 (online), DOI: 10.1044/jshr.0902.273 (1966).

[3] Schwarz-Friesel, M.: *Language and emotion*, pp. 157–174 (online), DOI: 10.1075/ceb.10.08sch, John Benjamins Publishing Company (2015).

[4] Yamamoto, H.: Lexical Modeling of Yamabuki (Japanese Kerria) in Classical Japanese Poetry, *JADH2013 DH-JAC2013 Conference Abstracts* (2013).