How Differently Do We Talk? A Study of Sentence Patterns in Groups of Different Age, Gender and Social Status

Michal Ptaszynski † Dai Hasegawa ‡ Fumito Masui †
Hiroshi Sakuta ‡ Eijiro Adachi §

† Department of Computer Science, Kitami Institute of Technology
{ptaszynski,f-masui}@cs.kitami-it.ac.jp
‡ Department of Integrated Information Technology, College of Science and Engineering,
Aoyama Gakuin University {hasegawa,sakuta}@it.aoyama.ac.jp
§ Department of Molecular Morphology, Graduate School of Medical Sciences,
Kitasato University jsmbm@kitasato-u.ac.jp

Abstract

In this paper we present our study in differences of how people talk. We study the differences by comparing frequent sentence patterns appearing in conversations between people of different age, sex, and social status. In the comparison we used a conversation corpus containing nearly hundred different conversations. Firstly, we defined sentence patterns as ordered combinations of sentence elements. Next, we automatically extracted lists of such patterns from conversations and used them in a text classification task. The overall results are compared in terms of Precision, Recall, and F-score.

1 Introduction

Comparative studies of differences in communication strategies within conversations have been researched from both qualitative and quantitative perspectives. The qualitative research can be most often found in linguistics and comparative linguistics, and focuses on thoroughly discussing a small number of the most evident differences in vocabulary, or sentence structure depending on the interlocutors, the situation, etc. A disadvantage of this kind of studies is that they rarely present any quantitative results. On the other hand the research presenting quantitative results can be found in corpus and computational linguistics. Corpus linguistic studies allow providing accurate quantitative results showing which words appear more often in which corpora. Unfortunately such studies are usually based on comparing either words or short n-grams (bigrams, trigrams). Actual patterns in language are usually more sophisticated than n-grams. For example, a pattern more sophisticated than n-gram, can be found in the following sentence in Japanese *Kyō wa nante kimochi ii hi nanda !* (What a pleasant day it is today!). The sentence contains a pattern *nante * nanda !*[1]. Therefore it would be desired to include such patterns with disjointed elements in the analysis as well. Finally, computational linguistics research often focuses on providing numerical values representing the performance of a machine learning classifier with the corpora used as training and test material. Although such results could be useful in linguistic studies, training of the classifier requires carefully selected training samples, which means one already has to know the patterns on which the classifier could be trained. Although this approach provides accurate numbers interpretable for the need of corpora comparison, the machine learning approach by the definition does not allow any new findings and unexpected discoveries within a corpus

In our studies we aimed at achieving all of the above mentioned kinds of results. We needed quantitative results in the form of accurate and consistent numbers interpretable as a rate of difference between corpora. We also wanted to know which patterns are used more frequently in which corpus, and we did not want to limit the research to single words or n-grams. Moreover, we wanted to be able to perform a qualitative analysis of the behavior of patterns across corpora. To achieve our goal we used a method known as Language Combinatorics, proposed by Ptaszynski et al. [2]. The method allows extraction of sophisticated patterns from sentences using words to create n-long combinations of sentence elements. It also allows classification of sentences with the use of the extracted patterns. We applied this method in the task of comparing corpora of conversations between people of different age, sex and social status.

Outline of this paper is as follows. In section 2 we present the methodology employed in our research. We describe the system of Ptaszynski et al. [2] and explain how we apply it to corpus comparison. Section 3 presents the corpus used in the research in general and the specific samples used in experiments in particular. Section 4 shows the results of experiments and discussion. Finally the paper is concluded in section 5.

2 Methodology

2.1 SPEC

Sentence **P**attern **E**xtraction ar**C**hitecturte is a system created by Ptaszynski et al. [2] on the assumptions of the Language Combinatorics approach. The system automatically extracts frequent sentence patterns distinguishable for a corpus (a collection of sentences). Firstly, the system generates ordered non-repeated combinations from the elements of a sentence. In every n-element sentence there is k-number of combination groups, such as that $1 \geq k \geq n$, where k represents

[1] equivalent of *wh*-exclamatives in English [1]; asterisk "*" used as a marker of disjoint elements

all k-element combinations being a subset of n. The number of combinations generated for one k-element group of combinations is equal to binomial coefficient, represented in equation 1. In this procedure the system creates all combinations for all values of k from the range of $\{1, ..., n\}$. Therefore the number of all combinations is equal to the sum of all combinations from all k-element groups of combinations, like in the equation 2.

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \quad (1)$$

$$\sum_{k=1}^{n} \binom{n}{k} = \frac{n!}{1!(n-1)!} + \frac{n!}{2!(n-2)!} + ... + \frac{n!}{n!(n-n)!} = 2^n - 1 \quad (2)$$

Next, the system specifies whether the elements appear next to each other or are separated by a distance by placing an asterisk ("*") between all non-subsequent elements. SPEC uses all original patterns generated in this procedure to extract frequent patterns appearing in a given corpus and calculate their weight.

If the initial collection of sentences was biased toward one of the sides (e.g., more positive sentences, or the sentences were longer, etc.), there will be more patterns of a certain sort. Thus agreeing to a rule of thumb in classification (fixed threshold above which a new sentence is classified as either positive or negative) might be harmful for one of the sides. Therefore assessing the threshold is required for optimizing the classifier. All of the above mentioned modifications are automatically verified in the process of evaluation to choose the best model. The metrics used in evaluation are standard Precision (P), Recall (R) and balanced F-score (F).

2.2 Corpora Comparison with SPEC

We propose an application of SPEC in comparison of corpora. SPEC provides enough information to perform both qualitative and quantitative comparison. Below we describe in detail which information is used for what purpose.

One of the information provided by SPEC is the result of an automatic classification. SPEC performs the classification by using two provided corpora, presumably of opposite characteristics (e.g., "positive reviews" and "negative reviews"). In the classification process the two corpora are first divided into n parts applied in an n-fold cross validation test (10-fold by default). The results averaged from all tests represent the overall score of the classifier.

When the two compared corpora are exactly the same for all performed tests, the results of the classifier for Precision will be equal to zero for threshold t higher than zero and 0.5 for threshold t equal or lower than zero. Similarly the Recall will be zero for $t > 0$ and 1 for $t \leq 0$. Finally, balanced F-score will have the values 0 and 0.67 for the same threshold range. Any result different to the above will mean that the two corpora are not the same. Moreover, when the two corpora are exactly different, meaning none of the patterns extracted from one corpus appears in the other, the results for all metrics would be equal 1. Thus we can consider the result of the classification as a rate of similarity between the two compared corpora.

In the process of pattern generation SPEC generates a list of patterns which contains patterns appearing uniquely for one of the sides (e.g. patterns unique for positive sentences) or in both (ambiguous patterns). Moreover, an ambiguous pattern could appear more frequently in one of the corpus and thus its weight will be biased toward 1 or -1 (but not reaching it). As a special case, an ambiguous pattern can appear at exactly the same rate in both corpora. Such a pattern will necessarily have a weight equal to 0 (late called zero-pattern). The weights of patterns can be interpreted as a probability rate of how often a certain pattern appears in either of the corpora. Therefore analysis of the patterns and their weights can contribute to the corpus linguistic studies.

SPEC extracts all patterns automatically. Within the extracted patterns those having the weight 1 or -1 appeared only in one of the two compared corpora. Therefore analyzing those specific patterns and the sentences in which they appeared could provide interesting linguistic discoveries. Since the patterns extracted automatically represent all probable frequent patterns hidden in the two compared corpora, we can assume that if the corpora cover a representative sample of the compared feature, the patterns already known to linguists should also be included in the weighted pattern list. Moreover, we can expect new patterns unknown before the comparison was made. Some of them will surely be data-dependent. However, we can assume that a filtering in the form of n-fold cross validation will retain only those patterns which were useful across all tests.

3 Datasets for Experiment

The methodology described in the section 2 was employed to compare corpora of conversations between people of different characteristics. We used conversations included in the BTSJ corpus created by Usami [3].

3.1 BTSJ Corpus

The BTS (Basic Transcription System) corpus [3] applied in this research is a corpus containing conversations between people of different age, sex and social status. It contains 99 conversation transcripts covering 1,604 minutes of talking. The records contain conversations between two Japanese native speakers, or between a Japanese native speaker and a Japanese language learner. Since we wanted to avoid any potential language mistakes, in our study we used only the former part. The conversations are performed either between friends or people who first met. Some conversations represent small talk, while others are on a specific topic. There are conversations between men only, women only, or mixed. Most of the conversations has been performed between students. Each of the above features of conversations can be considered as opposite features. Therefore we can extract conversation subsets for which only one feature would differ. By comparing such subsets with the use of SPEC we could extract those sentence patterns frequent and characteristic only for the one differing feature.

3.2 Datasets from BTSJ Corpus

From the BTSJ Corpus we extracted several conversation sets. Unfortunately, the corpus, although containing numerous conversations, does not contain enough conversation variations to compare all features separately. Therefore we extracted only those for which the experiment could be performed. We focused only on small talks. In

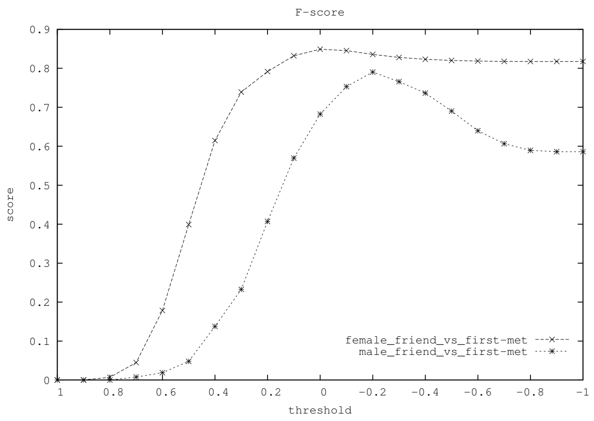

Figure 1: F-scores for different datasets across the whole threshold span.

particular we extracted 24 sets for conversations between female students. Half of those conversations was performed by friends and half by two interlocutors unknown to each other before the conversation. Next we extracted 12 conversation sets for similar conditions, but for male students. Having these conversation sets we were able to perform an experiment to compare how the way of talking differs for female and male students when they talk to their friends or to unrelated peers. The summary of all conversation sets used in the experiment is represented in Table 1.

4 Experiment and Discussion

4.1 General Observations

The general overview of the samples provides the following discoveries. On first met male interlocutors exchange more information than in friend's conversations. For female interlocutors on the contrary, friend's small talk is on average about twice as long as on first met, when the average number of sentences in conversations is compared. Males use longer sentences in general and exchange turns less often than females. Females use backchannel more often, which might suggest that for males it is more important to convey specific information rather than keep the conversation going as it is with females. Although these findings have to be interpreted within this closed data, they support other findings [4, 5].

Table 1: Summary of the conversation sets.

Small talk conversations		No. of samples	Avg. sent. length	Avg. sentences per conversations
Female-student	first met	12	12.7	288.9
	friends	12	9.3	550.0
Male-student	first met	6	12.4	326.5
	friends	6	14.5	245.3

4.2 Feature Differences

We compared those corpora which differed in one feature, namely, whether the conversations were performed by friends or by strangers. Comparison of the results achieved by the classifier shows that higher F-scores were achieved for female rather than male interlocutors. Higher F-score means that the compared conversation sets were easier to distinguish, or, that for females there were larger differences between conversations with friends and strangers. In general terms this means that comparing to men, women talk more differently to a person they just met than to friends.

In particular, the highest F-score achieved by the classifier for male conversations was F = 0.79 with Precision = 0.74 and Recall = 0.85, while for women the highest F-score reached 0.85 with P = 0.79 and R = 0.96. More detailed comparison of Precision and Recall rates for male and female conversation sets are represented in Figure 2 and 3.

This difference is similar to the differences in numbers of sentences appearing in each set. There were much larger differences for female interlocutors (about two times more sentences exchanged with friends comparing to strangers), while for males the differences were not that large (only one third with more sentences exchanged with strangers rather than with friends).

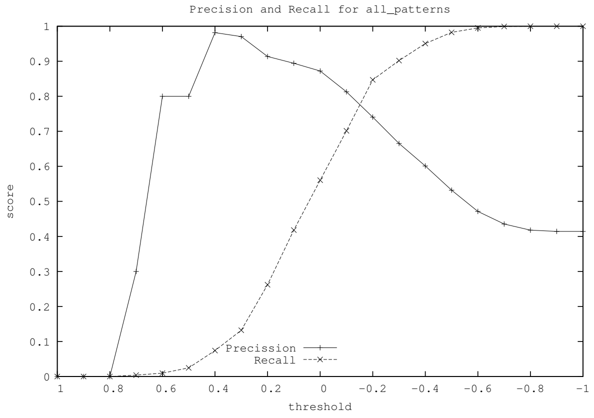

Figure 2: Precision and Recall with Break-Even Point for male students conversation dataset.

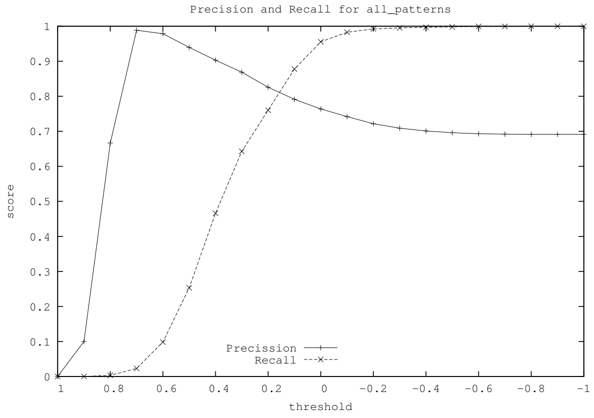

Figure 3: Precision and Recall with Break-Even Point for female students conversation dataset.

4.3 Detailed Analysis

Next we analyzed specific patterns characteristic to each of the compared side of the corpora. Below we discuss some of those patterns with example sentences. As the first interesting finding, we noticed that there are the same patterns for both male and female students in similar situations. For example, the pattern *nanka * na* appears in friend conversations for both sexes. Example sentences containing this pattern are given below The first two examples are for female students. The latter two are for male students.

Example 1. *Nanka... bannō nabe mitai na yatsu.* (Something like a... universal cooking pot.)

Example 2. *Nanka sugoi kōseinō na sukyanā da to–* (Oh its like an amazingly high-performance scanner!)

Example 3. *Nanka gakugaku, mitai na.* (Something, like a sound of knocking.)

Example 4. *Intānetto to shite wa, nanka kekkō, fusoku na toko mo aru.* (So when it comes to the Internet, it has like pretty a lot of deficiencies.)

There are also similar patterns for both sexes appearing in conversations under the "first met" condition, such as the pattern *so * desu*, often taking the form *so nan desu* in the example sentence below. This sentence appears in both female and male student conversations.

Example 5. *Aaa, sō nan desu ka* (Oh, so that is the case [I understand now])

The presence of the same patterns suggests such patterns are characteristic for social distance rather than for the sex of interlocutors.

Except patterns which appear for both sexes there are also patterns specific for a particular sex. For example self referential expressions like *ore* for boys and *atashi* for girls (both meaning "I/me"). See the below two sentence examples.

Example 6. *Ore 1-kai mo nai kara ne.* (I[masculine] haven't [done it] even once, you know.)

Example 7. *Nanka atashi, tento tte sugoi suki.* (Oh, I[feminine] just love tents so much.)

Moreover, there are patterns characteristic for one sex which are not by the definition specific to the sex. For example, a pattern *sō sō sō!* (affirmative interjectional expression meaning "yes, yes that's right!") does not contain any gender-specific vocabulary (like in the case of *ore* vs. *atashi*). However, the actual language use shows that although the pattern is often used by female interlocutors, it does not appear in male conversations. On the other hand a pattern similar in meaning *hai hai hai* ("yes, yes, yes") is used by males, but does not appear in female conversations.

The Language Combinatorics approach allows extracting frequent expressions, phrases and words people tend to use in conversations. Most of corpus linguistic research so far [4, 5] has focused mainly on topics of conversation. Therefore our method can greatly contribute to linguistics by providing analysis of ways of talking not researched thoroughly before. It has to be added that the method is also capable of extracting conversation topics. For example, for friend-students of both sexes, the topic of "an exam" was equally frequent. However, a topic of "a marriage" appeared only in female student conversations. Also "food" and, surprisingly, "alcohol" as well appeared only for girl-students. On the other hand "newspapers" were the boys-specific topic.

5 Conclusions and Future Work

We presented our study in differences of how people talk. We study those differences by comparing frequent sentence patterns appearing in conversations. We employed the Language Combinatorics approach and defined sentence patterns as ordered combinations of sentence tokens. Next, we automatically extracted lists of frequent sentence patterns from the conversations and performed a text classification experiment using those patterns. The overall results of the classifier interpreted to explain the differences between the conversation sub-corpora.

We found out that male interlocutors use longer sentences and exchange turns less often than females. Moreover, when it comes to differences of talking to friends and newly met people, for females the differences are much grater than for males. Investigation of patterns specific for each kind of conversation shows that some patterns appear for both male and female interlocutors in similar situations, which suggests these patterns could be typical for linguistically expressed social distance. There were also patterns specific for a particular sex. These included words typically assigned to only by one sex (like *ore* for man and *atashi* for women), but also patterns which although could be used by anyone, in practice were used only by one side.

In the near future we plan to perform further analysis of other conversations as well. We also plan to perform corpora comparison this way on different kinds of corpora, not limited to conversations.

Acknowledgments

This work was supported by Strategic Information and Communications R&D Promotion Programme (SCOPE), the Ministry of Internal Affairs and Communications, Japan.

References

[1] Kaori Sasai. 2006. The Structure of Modern Japanese Exclamatory Sentences: On the Structure of the *Nanto*-Type Sentence. *Studies in the Japanese Language*, Vol, 2, No. 1, pp. 16-31.

[2] Michal Ptaszynski, Rafal Rzepka, Kenji Araki and Yoshio Momouchi. 2011. Language combinatorics: A sentence pattern extraction architecture based on combinatorial explosion. *International Journal of Computational Linguistics (IJCL)*, Vol. 2, Issue 1, pp. 24-36.

[3] Mayumi Usami (Ed.). 2007. *BTS ni yoru nihongo hanashikotoba kōpasu 1 (hatsutaimen, yūjin; zatsudan, tōron, sasoi)* [Conversation corpus of spoken Japanese using the Basic Transcription System (first meeting, friend's conversation, small talk, discussion, invitation)] (In Japanese), Tokyo University of Foreign Studies, Tokyo, Japan.

[4] Adelaide Haas. 1979. Male and female spoken language differences: Stereotypes and evidence. *Psychological Bulletin*, Vol. 86, No. 3, pp. 616-626.

[5] Lynette Hirschman. 1994. Female-male differences in conversational interaction. *Language in Society*, 23, pp 427-442.