Differences in Speaker Individualising Information between Case Particles and Fillers in Spoken Japanese

Shunichi Ishihara
Department of Linguistics
Australian National University

shunichi.ishihara@anu.edu.au

Abstract

This study investigates idiosyncrasy manifested in language use in spoken Japanese. For this purpose, we use speaker classification techniques as analytical tools. More precisely, focusing on Japanese case particles and fillers, of which the linguistic functions differ significantly, we aim to investigate 1) the extent of speaker idiosyncrasy in the selection of certain case particles/fillers over others in Japanese monologues, and 2) the differences, if any, between case particles and fillers in the degree of speaker-individualising information. We discuss what contributes to the identified differences between case particles and fillers. This study will contribute to the further development of automatic speaker recognition systems and authorship analysis studies.

1 Introduction

We intuitively know that different people speak/write differently, even when they try to convey the same message. We also know that people tend to use their individually-selected, preferred words despite the fact that, in principle, they can use any word at any time from the vocabulary built up over the course of their lives. Every speaker of a given language has their own distinctive and individual version of the language – which is often referred to as their idiolect (Halliday et al. 1964).

Linguistic idiosyncrasy has been studied in both spoken and written languages (yet, more extensively on written languages) (Baayen et al. 1996, Burrows 1987, Doddington 2001, Ishihara and Kinoshita 2010, Weber et al. 2002). Many of these studies were based on the unique lexical usage of authors (Holmes et al. 2001, Juola and Baayen 2005), assuming that word selection is unique to the individual author, and that their preferred selection is consistent over time (Holmes 1992). In particular, function words are often used as a feature to quantify the unique lexical usage of individual authors, and it has been attested that function words carry author-individualising information (Binongo 2003, Holmes, et al. 2001). In addition to function words, fillers (such as English "um", "you know", and "like"), which are unique to spoken languages, have also been reported to carry speaker idiosyncratic information (Ishihara and Kinoshita 2010, Weber, et al. 2002).

Although the above studies demonstrated that function words and fillers carry speaker/writer idiosyncratic information, the degree/characteristics of the individualising information that they carry may be different as the type of linguistic information they provide is significantly different. We will investigate this in this study. For that purpose, we use case-particles and fillers appearing in Japanese monologues. Case particles are representative function words in Japanese. We use Japanese monologues because many of the previous studies used English as the target language, and research on idiosyncrasy in spoken languages are relatively fewer than those on written languages.

That being said, the current study will investigate 1) the extent of speaker idiosyncrasy in the selection of certain case particles/fillers over others in Japanese monologues, and 2) the differences, if any, between case particles and fillers in the degree of idiosyncrasy.

In order to answer the above questions, we will conduct a series of speaker classification tests. The hypothesis is that the more consistent the individual speaker's selection of words (e.g. particles) is, and the more significantly words selected by one speaker differ from those selected by another, then the more accurate the speaker classification results will be.

1.1 Case Particles and Fillers

Case particles (kaku-joshi), which are function words, provide the grammatical relationship be-

Shunichi Ishihara. 2013. Differences in Speaker Individualising Information between Case Particles and Fillers in Spoken Japanese. In *Proceedings of Australasian Language Technology Association Workshop*, pages 118−122.

tween the predicate of a sentence and the noun phrases appearing in the sentence. In Example 1), the case particles, **-ga**, **-de** and **–o**, are the subjective (SUBJ), instrumental (INS), and accusative (ACC) markers, respectively.

ani-**ga** boo-**de** watashi-**o** tataita Ex 1)
elder.brother-**SUBJ** stick-**INS** I-**ACC** hit.past
My elder brother hit me with a stick.

Fillers function as placeholders when fluency fails and one is searching for a desired expression (Martin 2004:1041). In the database we used for this study, a filler tag is assigned to the preselected words which have the function of 'filling up gaps in utterances'. Fillers are unique to spoken languages.

2 Methodology

Two kinds of comparisons are involved in speaker classification tests. The first is *Same Speaker Comparison* (SS comparison) in which two speech samples produced by the same speaker need to be correctly identified as being from the same speaker. The second is, *mutatis mutandis*, *Different Speaker Comparison* (DS comparison). These comparisons were conducted separately for case particles and fillers. Since the comparisons are yes-no basis, the baseline for these comparisons is 50%.

2.1 Database and Speakers

For this study, we used the monologues from the Corpus of Spontaneous Japanese (CSJ) (Maekawa et al. 2000), which are categorised as *Academic Presentation Speech* (APS) or *Simulated Public Speech* (SPS). APS was mainly recorded live at academic presentations, most of which were 12-25 minutes long. For SPS, 10-12 minute mock speeches on everyday topics were recorded. We selected our speakers from this corpus based on three criteria: availability of multiple and non-contemporaneous recordings, spontaneity (e.g. not reading) of the speech, and speaking in standard modern Japanese. Spontaneity and standardness of the language were assessed on the basis of the rating the CSJ provides. Thus, only those speech samples which are high in spontaneity and uttered entirely in Standard Japanese were selected for this study. This resulted in 416 speech samples (208 speakers: 132 male and 76 female speakers x 2 sessions). From the 416 speech samples, 208 SS and 86112 DS comparisons are possible. From these selected speakers, 64 case particles and 49 fillers were identified.

2.2 Vector Space Model

First of all, the identified words were sorted by their occurrences in descending order. Then, using the sorted order and the occurrences of the identified words, each speech was modelled as a real-valued vector in this study. If n different words are used to represent a given speech S, the dimensionality of the vector is n. That is, S is represented as a vector of n dimensions ($S = (F_1, F_2 \ldots F_n)$, where Fn represents the nth component of S and Fn is the frequency of the nth word). For example, if 5 words (e.g. *ah*, *like*, *OK*, *yes*, *all right*) are used to represent a speech sample (x), and the frequency counts of these words in the speech sample are 3, 10, 4, 18 and 1, respectively, the speech sample x is represented as given in 1).

$$\vec{x} = (3,10,4,18,1) \quad 1)$$

In this study, the speech samples are modelled using different vector dimensions. This is to see how the performance of the speaker classification system is influenced by the number of dimensions.

2.3 Term Frequency Inverse Document Frequency Weighting

The usefulness of particular words is determined by their uniqueness as well as by how frequently they occur. The *tf·idf* (term frequency inverse document frequency) weight, of which formula is given in 2), was used to evaluate how unique a given word is in the population, and weight was given to that word to reflect its importance to speaker classification (Manning and Schütze 2000)

$$w_{i,j} = tf_{i,j} * log(\frac{N}{df_i}) \quad 2)$$

In 2), term frequency ($tf_{i,j}$) is the number of occurrences of word i (w_i) in the document (or speech sample) j (d_j). Document frequency (df_i) is the number of documents (or speech samples) in the collection in which that word i (w_i) occurs. N is the total number of documents (or speech samples).

2.4 Cosine Similarity Measure

The similarity (=difference) between two speech samples, which are represented as vectors ($\vec{x}, \vec{y}$), was calculated based on the cosine similarity

measure (Manning and Schütze 2000). This particular method (e.g. instead of measuring the distance between two vectors) was selected because the durations of the speech samples are all different. Note that for the experiments of this study, the length of the vectors were standardised by only considering the *X* most frequent case particles and fillers across the speakers.

$$cos(\vec{x},\vec{y}) = \frac{\sum_{i=1}^{n} x_i * y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 * \sum_{i=1}^{n} y_i^2}} \quad 3)$$

The range of difference between the two vectors (similarity($\vec{x},\vec{y}$)) is between 1.0 (=*cos*(0°)) for two vectors pointing in the same direction – e.g. speech samples which are identical – and 0.0 (=*cos*(90°)) for two orthogonal vectors – two speech samples which are completely different, because weights are by their definition not negative.

3 Speaker Classification Tests

The performance of speaker classification was assessed on the basis of the *probability distribution functions* (PDFs) of the difference (E) of the paired speech samples between two contrastive hypotheses. One is the hypothesis that two speech samples were uttered by the same speaker (SS hypothesis) and the other is that two speech samples were uttered by different speakers (DS hypothesis). These probabilities can be formulated as $P(E|H_{ss})$ and $P(E|H_{ds})$ respectively, where E is the difference between two speech samples in comparison, H_{ss} is the SS hypothesis and H_{ds} is the DS hypothesis. In this study, the PDF of the difference assuming the SS hypothesis is true is called the SS PDF (PDF_{ss}), and the PDF of the difference assuming the DS hypothesis is true the DS PDF (PDF_{ds}). Each PDF was modelled using the kernel density function (KernSmooth library of R statistical package), which is a non-parametric way of estimating PDF. Examples of PDF_{ss} and PDF_{ds} are given in Figure 1.

The PDF_{ss} and PDF_{ds} of Figure 1 do not conform to a normal distribution. This is the motivation for the use of the kernel density function. PDF_{ss} and PDF_{ds} are not always monotonic, and may result in more than a single crossing point, particularly when the dimension of a vector is less than 5. Thus, the performance of the system with a vector length less than 5 is not given. These two PDFs also show the accuracy of this particular speaker classification system. If the crossing point (θ) of the PDF_{ss} and the PDF_{ds} is set as the threshold, we can estimate the performance of this particular speaker classification system from these PDFs. Area 1 in Figure 1 – the area bound by the grey line (PDF_{ss}), the vertical dotted line of $x = \theta$ and the line of $y = 0$ – is the predicted error for the SS comparisons. Area 2 of Figure 1 – the area bound by the black line (PDF_{ds}), the vertical dotted line of $x = \theta$, and the line of $y = 0$ – is the predicted error for the DS comparisons. The accuracy/error rate of a speaker classification system (both in SS and DS comparisons) was estimated by calculating Areas 1 and 2.

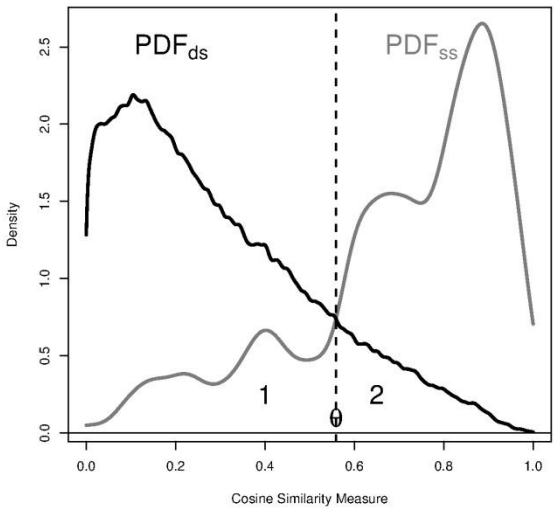

Figure 1: Example of PDF_{ss} (grey curve) and PDF_{ds} (black curve). The vertical dotted line (θ) is the crossing point of PDF_{ss} and PDF_{ds}. Probability density = Y-axis; Cosine similarity measure = X-axis.

4 Test Results and Discussion

In Figure 2, the same speaker (SS) and different speaker (DS) comparisons classification accuracies and the average accuracy between them are plotted separately for the case particles and fillers as a function of the number of vector dimensions.

For the fillers, according to Figure 2, the performance of the SS and DS comparisons are comparable until 20 dimensions, after which the DS comparisons perform better than the SS comparisons. For the case particles, the DS comparisons consistently outperform the SS comparisons. This underperformance of the SS comparisons may be due to the fact that the sample number for estimating the PDF_{ss} (208) is far fewer than that for estimating the PDF_{ds} (86112).

Figure 2 indicates that the average speaker classification accuracy reaches as high as 69.8%

for the case particles with 35 dimensions and 82.7% for the fillers with 25 dimensions, insofar as the performance of speaker classification is consistently better for fillers than case particles. This indicates that fillers carry more speaker-specific information than case particles.

Communication has been traditionally viewed as an intentional act of transferring information. However, whatever the mode of communication (spoken or written), along with the linguistic information about the symbolic content of the intended message, paralinguistic or extralinguistic information about the speaker/writer, such as age, sex, social background, psychological state, health, etc. (Nolan 1983) is also conveyed.

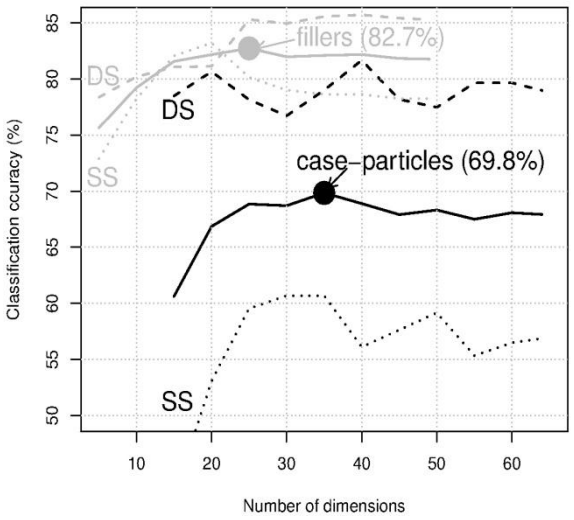

Figure 2: The SS (dotted lines) and DS (dashed lines) comparison accuracies, and their average accuracies are plotted separately for case particles (black) and fillers (grey). The circles indicate the best average accuracy for each type.

Fillers transfer more than the linguistic information encoded in messages, and thus we propose that they are more closely related to paralinguistic and extralinguistic information. This can also be understood from the fact that fillers do not conventionally appear in written texts. It has also been argued based on empirical data that fillers manifest the cognitive process that the speaker is undergoing (Sadanobu and Takubo 1995) and also reflect a speaker's difficulty in conceptual planning and linguistic encoding (Watanabe et al. 2008). The cognitive process is a well-known source of individual differences (Cooper 2002). On the other hand, case particles are key players for linguistic information, such as the syntactic relationship between a noun phrase of a sentence and the predicate of the sentence, the logical relationship between two clauses, etc., which are more directly important for accurately transferring the content information encoded in messages than fillers. Since case particles serve as the dominant carrier of the information directly connected to the propositions of the messages, it is likely that case particles do not have much more capacity to further carry idiosyncratic information of individual speakers. One of the reviewers argues that fillers carry more individualising information mainly because they are relatively free from grammar, which more rigidly controls the use of case particles.

Speaker classification accuracy drastically improves from 15 dimensions (60.6%) to 25 dimensions (69.8%) for case particles. The same increase in accuracy can be observed with fewer dimensions (from 5 dimensions: 75.6% to 15 dimensions: 81.5%) for fillers. This observation that more dimensions need to be included for the case particles in order to reach the same optimal performance level as the fillers is likely due to the fact that the first 15-20 most frequently used case particles are so ubiquitous in the utterances. Hence, the added function of bearing the individualising information of a speaker is too great for case particles. Also note that the curve of the case particles in Figure 2 starts with 15 dimensions because the PDF_{ss} and the PDF_{ds} with less than 15 dimensions become non-monotonic having multiple crossing points between them, and thus sensible results could not be obtained with less than 15 dimensions.

5 Conclusions

It has been demonstrated that Japanese case particles and fillers carry speaker idiosyncratic information to the extent that the average speaker classification accuracy is ca. 69.8% and 82.7%, respectively. We discussed the argument that fillers are more endowed with the idiosyncratic information of speakers than case particles because of the different levels of information with which they operate. Namely, case particles mainly handle a linguistically lower level of structural information, which is directly relevant to the content of messages, whereas fillers assume the task of conveying paralinguistic and extralinguistic information, which have a stronger relevance to a speaker's cognitive processes and are highly diverse at the individual speaker level.

Acknowledgments

The author thanks for the valuable comments and suggestions from the three anonymous reviewers.

References

Baayen, H., Van Halteren, H., and Tweedie, F. (1996) Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing 11*(3): 121-132.

Binongo, J. N. G. (2003) Who wrote the 15th book of Oz? An application of multivariate analysis to authorship attribution. *Chance 16*(2): 9-17.

Burrows, J. F. (1987) Word-patterns and story-shapes: The statistical analysis of narrative style. *Literary and Linguistic Computing 2*(2): 61-70.

Cooper, C. (2002) *Individual Differences* (2nd ed.). London: Arnold; New York: Oxford University Press.

Doddington, G. (2001) Speaker recognition based on idiolectal differences between speakers. *Proceedings of 2001 Eurospeech*: 2521-2524.

Halliday, M. A. K., Macintosh, A., and Strevens, P. D. (1964) *The Linguistic Sciences and Language Teaching*. London: Longmans.

Holmes, D. I. (1992) A stylometric analysis of Mormon scripture and related texts. *Journal of the Royal Statistical Society Series a-Statistics in Society 155*: 91-120.

Holmes, D. I., Robertson, M., and Paez, R. (2001) Stephen crane and the New-York Tribune: A case study in traditional and non-traditional authorship attribution. *Computers and the Humanities 35*(3): 315-331.

Ishihara, S., and Kinoshita, Y. (2010) Filler words as a speaker classification feature. *Proceedings of the 13th Australasian International Conference on Speech Science and Technology*: 34-37.

Juola, P., and Baayen, R. H. (2005) A controlled-corpus experiment in authorship identification by cross-entropy. *Literary and Linguistic Computing 20*(Suppl): 59.

Maekawa, K., Koiso, H., Furui, S., and Isahara, H. (2000) Spontaneous speech corpus of Japanese. *Proceedings of the 2nd International Conference of Language Resources and Evaluation*: 947-952.

Manning, C. D., and Schütze, H. (2000) *Foundations of Statistical Natural Language Processing* (2nd ed.). Cambridge, Mass.: MIT Press.

Martin, S. E. (2004) *A reference grammar of Japanese*. Honolulu: University of Hawai'i Press.

Nolan, F. (1983) *The Phonetic Bases of Speaker Recognition*. Cambridge: Cambridge University Press.

Sadanobu, T., and Takubo, Y. (1995) The monitoring devices of mental operations in discourse: A case of 'eeto' and 'ano (o)'. *Gengo kenkyu [Language Studies]*(108): 74-93.

Watanabe, M., Hirose, K., Den, Y., and Minematsu, N. (2008) Filled pauses as cues to the complexity of upcoming phrases for native and non-native listeners. *Speech Communication 50*(2): 81-94.

Weber, F., Manganaro, L., Peskin, B., and Shriberg, E. (2002) Using prosodic and lexical information for speaker identification. *Proceedings of the 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 1*: 141-144.