Extracting speaker-specific functional expressions from political speeches using random forests in order to investigate speakers’ individual political styles

Takafumi Suzuki
University of Tokyo

In this study we extracted speaker-specific functional expressions from political speeches using random forests in order to investigate speakers’ individual political styles. Along with methodological development, stylistics has expanded its scope into new areas of application such as authorship profiling and sentiment analysis in addition to conventional areas such as authorship attribution and genre-based text classification. Among these, computational sociolinguistics, which aims at providing a systematic and solid basis for sociolinguistic analysis using machine learning and linguistically-motivated features, is a potentially important area. In this study we show the effectiveness of the random forests classifier for such tasks by applying it to Japanese prime ministers’ Diet speeches. The results demonstrate that our method successfully extracted the speaker-specific expressions of two Japanese prime ministers, and enabled us to investigate their individual political styles in a systematic manner. The method can be applied to sociolinguistic analysis of various other types of texts, and in this way, this study will contribute to developing the area of computational sociolinguistics in the field of stylistics.

1 Introduction

In this study we extracted speaker-specific expressions from political speeches using random forests in order to investigate speakers’ individual political styles. Stylistic text classification has been conventionally used for detecting the authors, registers, genres or chronological variations of texts (Kenny, 1982; Jin and Murakami, 2003). Many content-independent features have been used for these tasks, and these features have been merged in order to improve the classification performance (Grieve, 2007).

Recently, various new types of texts have been produced along with the growth of the Web (Aitchison and Lewis, 2003), and methodological development in computational linguistics has also enabled us to extract more detailed information from texts, thus stylistics has expanded its scope into new areas of application like authorship profiling (Estival et al., 2007) and sentiment analysis (Pang and Lee, 2004). Among these new areas for stylistics, computational sociolinguistics, which was first proposed by Argamon et al. (2007c), is a potentially important area. Using feature extraction carried out by machine learning with linguistically-motivated features, this area aims at providing a systematic and solid basis for sociolinguistic analyses (Argamon et al., 2007c).

Turning our eyes to the area of political text analysis, many studies have focused on the content of these texts in order to analyze the policies, attitudes or thoughts of politicians (Neuendorf, 2001). Recently, however, as politicians’ rhetoric has increased in importance along with the development of television politics or media politics (Maarek and Wolfsfeld, 2003), the style of speeches, as well as their content, has attracted more attention from scholars, especially from sociolinguistic viewpoints (Ahrens, 2005; Azuma, 2006; Suzuki and Kageura, 2008).

One of the main purposes of these sociolinguistic studies of political speeches is to investigate the individual political styles of politicians from their speech styles. However, a common problem of many of these studies and also many other sociolinguistic studies is that the expressions that are analyzed are selected solely on the basis of the researcher’s interests or preferences, which can sometimes lead to contradictory interpretations. In other words, it is difficult to determine whether these kinds of analyses have
in fact correctly identified the individual language use of politicians, and, on this basis, correctly characterised their individual political styles. Also, if their styles are characterised by the infrequent use of specific expressions, this has rarely been focused on.

Against this background, this study extracts speaker-specific functional expressions from political speeches using random forests in order to investigate speakers’ individual political styles. By classifying the texts of a speaker according to their style and extracting the variables contributing to classification, we can clarify the expressions specific to the speaker. These expressions enable us to investigate his/her individual political style on the basis of his/her speech style. Expressions that are infrequently used should also be identified.

Through our analyses of political speeches, we show the effectiveness of feature extraction using the random forests classifier (Breiman, 2001) for such tasks, and we investigate the features useful for sociolinguistic interpretation as well as those resulting in high classification performance. Additionally, we compare the classification performance of microscopic differences of features, i.e., orders of parts-of-speech and stemming, which could affect the results. Our method can be used for sociolinguistic analyses of various other types of texts, and in this way, this study will contribute to developing the area of computational sociolinguistics in the field of stylistics. The rest of this paper is organized as follows. In Section 2, we review previous work, and in Section 3, explain our experimental setup. In Section 4, we discuss our results and in Section 5, we make concluding remarks.

2 Previous work

In the field of stylistics, statistical text classification has been used for many tasks. Multivariate analyses have been the most popular approach in this field since Burrows (1989), but machine learning has currently become another popular approach (Sebastiani, 2002). Among machine-learning methods, SVM has been more widely used than any other method because of its robustness, while the random forests classifier (Breiman, 2001) we use in this study has rarely been applied in this field, though its high performance has been demonstrated in other fields such as bio-informatics (Díaz-Uriarte and Alvarez de Andrés, 2006). Recently, however, a study reported that it can perform better than SVM for authorship attribution in Japanese (Jin and Murakami, 2007), and it is therefore important to confirm the performance of random forests for stylistic text classification. Moreover, this method has more advantages for our task, which we explain in Section 3.4. One of our purposes is therefore to test the effectiveness of this method for stylistics, especially computational sociolinguistics.

Computational sociolinguistics was first suggested as a potentially important area of stylistics by Argamon et al. (2007c). Though few studies developed their points of view, Argamon et al. (2007a) applied stylistic text classification using SVM in order to investigate the stylistic differences between male and female texts in 18th and 19th century French literature, and Argamon et al. (2007b) applied similar methods for investigating differences attributable gender, race and nationality in Black drama. In an analogous study, Kim and Daelemans (2008) tried to predict author personality from a Dutch corpus using syntactic features. Along with these studies, we try to develop this area by showing the effectiveness of this approach for investigating the individual political styles of two Japanese prime ministers.

Turning our eyes to the area of political text analysis, there have been many studies analyzing their content in English (Tulis, 1987; Lim, 2002) as well as Japanese (Watanabe, 1974; Tanaka, 1994). In addition to these studies, stylistic analysis, especially from a sociolinguistic viewpoint, has recently been paid more attention by scholars in English (Ahrens, 2005) and also in Japanese (Azuma, 2006), though these studies focused on a few specific expressions. This study, by applying feature extraction using random forests, will provide a more solid basis for sociolinguistic analysis of political speeches.

3 Experimental setup

3.1 Overview

We focused on the Diet addresses of two Japanese prime ministers, namely Nakasone and Koizumi. We performed two classification experiments: one classifying the Diet addresses of Nakasone in relation to the addresses made by three of his contemporaries from 1980 to 1989 (the...
Table 2: Basic data of the corpora

<table>
<thead>
<tr>
<th></th>
<th>functional nouns</th>
<th>conjunctions</th>
<th>adnominals</th>
<th>particles*</th>
<th>auxiliary verbs**</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>tokens types</td>
<td>tokens types</td>
<td>tokens types</td>
<td>tokens types</td>
<td>tokens types</td>
</tr>
<tr>
<td>Nakasone 2052</td>
<td>2052</td>
<td>43</td>
<td>389</td>
<td>18</td>
<td>505</td>
</tr>
<tr>
<td>1980-1989 1572</td>
<td>1572</td>
<td>45</td>
<td>274</td>
<td>22</td>
<td>379</td>
</tr>
<tr>
<td>Koizumi 1211</td>
<td>1211</td>
<td>42</td>
<td>131</td>
<td>18</td>
<td>184</td>
</tr>
<tr>
<td>1989-2006 5855</td>
<td>5855</td>
<td>57</td>
<td>938</td>
<td>26</td>
<td>1305</td>
</tr>
</tbody>
</table>

* 2nd order parts-of-speech
** with stemming

Table 1: Basic data of the corpora

<table>
<thead>
<tr>
<th></th>
<th>addresses</th>
<th>all words</th>
<th>function words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nakasone 10</td>
<td>10</td>
<td>47419</td>
<td>3390</td>
</tr>
<tr>
<td>1980-1989 9</td>
<td>9</td>
<td>35979</td>
<td>2862</td>
</tr>
<tr>
<td>Koizumi 11</td>
<td>11</td>
<td>46979</td>
<td>3908</td>
</tr>
<tr>
<td>1989-2006 31</td>
<td>31</td>
<td>149234</td>
<td>5774</td>
</tr>
</tbody>
</table>

We downloaded the addresses from the on-line database Sekai to Nihon (World and Japan)\(^2\) and applied morphological analysis to the addresses using ChaSen (Matsumoto et al., 2003), a Japanese morphological analysis system. We assigned part-of-speech tags using ChaSen, and manually united the notational differences that are distinguished only by kanji and kana in Japanese.

The drafts of the speeches were reportedly written by bureaucrats as well as by the prime ministers themselves (Shinoda, 1994),\(^3\) but the texts we use in this study are those written down by secretaries in shorthand, instead of the drafts themselves, thus the individual speech styles of the politicians can be seen from the texts of the addresses. Other possible factors affecting their styles apart from the individual differences between the speakers are relatively controlled, as Diet speech is a relatively restricted register, but changing Japanese language usage in general was shown to affect their styles (Suzuki and Kageura, 2008). We thus classified the addresses from those of their respective contemporaries.

3.2 Corpora

As features, we used the bag-of-words of the relative frequencies of function words, i.e., functional nouns (nouns-dependent and nouns-pronominals assigned by ChaSen),

Our main purpose in this paper is, however, to extract and analyze features rather than to compare the classification performance as such, so this problem does not affect the validity of the selection process described above.

\(^2\)www.ioc.u-tokyo.ac.jp/~worldjpn

\(^3\)This assertion is based on interviews with related people; the drafts of the speeches themselves have never been obtained so far.
and extracted random subsets of √M the original data matrix classifier proposed by Breiman (2001). We first replicated
as a classification method, we used the random forests
methodology and readability of the texts (c.f., Tuldava, 1993; Otsuka et al., 2007), while some functional nouns can represent
abstrusity or self-assertiveness. Some of these features
were in fact used for sociolinguistic analyses of political
speeches (Azuma, 2006). The detailed interpretation of
these features will be given when we set out our results in
Section 4.2.

It is better for our purpose to use deeper order part-
of-speech tags of particles\(^5\) and the stemming version of
auxiliary verbs as they facilitate more meaningful inter-
pretation,\(^6\) but as these microscopic differences in features
(order of parts-of-speech and stemming) could affect the
classification performance, we compared the results using
first order and second order part-of-speech tags of parti-
cles, and auxiliary verbs with and without stemming. Ta-
ble 1 represents the number of addresses, and the total
number of tokens and types with all words and function
words in four classification categories, while Table 2 sets
out the respective number of tokens and types for each
part-of-speech.

3.4 Method
As a classification method, we used the random forests
classifier proposed by Breiman (2001). We first replicated
the original data matrix \(M_{i,j}\) 1000 times with replacement,
and extracted random subsets of \(\sqrt{M}\) variables from each
replicated data. We constructed an unpruned decision tree
for each sample using the Gini index formalized as fol-

\[
GI = 1 - \sum_{c=1}^{n} (p(c | x))^2
\]

where \(c\) represents the class and \(p(c | x)\) is the probability
that the divided individuals (the texts) belong to the
class in constructing a tree (Jin, 2007). We constructed a
new classifier by a majority vote of the set of trees. Two-
thirds of the bootstrap samples were used for constructing
the model and the other third were left out for testing the
model (out-of-bag test).

We calculated the variable importance using the following
formula (Breiman, 2001):

\[
VI_{\text{ach}} = \frac{\text{mean}(C_{\text{oob}} - C_{\text{per}})}{s.e.}
\]

\(C_{\text{oob}}\): number of votes cast for the correct class in the
out-of-bag data
\(C_{\text{per}}\): number of votes cast for the correct class when
\(m\) variables are randomly permuted in the out-of-bag
data
\(s.e.\): standard error

The mean value of subtractions for all trees formulated
above represents the variable importance for a permuted
variable. It represents the degree to which a class loses
its specific character when a type of word changes into
another type of word.\(^7\)

The purpose of this study is to analyze the styles of two
prime ministers rather than to examine styles in the
actual ten or eleven addresses given by the two prime min-
isters. In other words, our interest is in the styles of the
person, rather than the styles of the actual ten or eleven
texts. The styles of the person could have been embodied
in other states but we have obtained no texts other than
these actual ten or eleven. A thousand bootstrap samples
represent a thousand possible states in which the styles
could have been embodied, thus bootstrapping is appro-
appropriate for our purpose.\(^8\) Also, extracting a random subset

\(^4\) We manually excluded content-dependent adnominals.
\(^5\) Case particles’ or ‘conjunctive particles’ etc.
\(^6\) A particle can have different meanings when used in different sec-
ond order parts-of-speech, while the different forms of an auxiliary verb
have the same meaning in Japanese.
from each bootstrapping sample enhances the diversity of a set of a thousand bootstrapping samples, and enables us to shed light on more diverse aspects of the styles of the person. Additionally, we can also take into account the correlation of variables using this method. By using the variable importance above, we extracted the speaker-specific expressions characterising the individual speech styles of the person in a systematic manner, which should help shed light on his/her personality, character, and, as in this case, political style.

4 Results and discussion

4.1 Experimental results

Table 3 describes the precision, recall rates, and $F_1$ values in each classification experiment, using functional nouns, conjunctions, adnominals, particles (first and second order part-of-speech tags), auxiliary verbs (with and without stemming), particles and auxiliary verbs (four patterns)\(^9\) and all function words (four patterns).

The results confirmed the effectiveness of random forests for stylistic text classification, as $F_1$ values on Nakasone’s classification experiment attained more than 75%, and those on Koizumi’s, more than 90%, using all the function words. That is evaluated high according to the criteria (more than 70%) used in an analogous study (Argamon et al., 2007b).\(^10\) Also, the results using all function words demonstrated that Koizumi may have a more distinctive style in his era than Nakasone had in his era,\(^11\) as the experiment in relation to the former showed higher classification performance.\(^12\)

The results also showed that the classification performance of each part-of-speech differed significantly between the two prime ministers; conjunctions, adnominals and auxiliary verbs showed better performance in Koizumi’s classification while functional nouns and particles showed better performance in Nakasone’s classification. These differences represent differences in speech styles, and could also represent differences in political style between the two prime ministers, a point we investigate in Section 4.2.\(^13\)

Also, the results showed that the deeper order of part-of-speech of particles and the stemming version of auxiliary verbs can lead the equal or better performance in most cases, thus we can use these features for interpretation.

4.2 Feature analysis

Table 4 and Table 5 show the top ten variables in the classification experiments using all the function words (second order part-of-speech of particles and the stemming version of auxiliary verbs), with their translations in English, their parts-of-speech, variable importance ($V_{\text{var}}$), and the notation that these variables were frequently used (+) or infrequently used (−) in the two prime ministers’ Diet addresses compared to their respective contemporaries. These variables significantly contribute to the classification, thus they represent the individual language use of the two prime ministers.

The results showed that the classification performance given by each part-of-speech and extracted variables corresponded with each other; the important variables in Nakasone’s classification include functional nouns, particles and auxiliary verbs while those in Koizumi’s classification include an adnominal and a conjunction as well as some functional nouns and auxiliary verbs. Also, the results showed that we can extract infrequently used expressions as well as frequently used ones.

Nakasone and Koizumi are regarded as two of the strongest prime ministers in Japanese politics; they aimed

---

\(^9\)This combination was used in Suzuki and Kageura (2008) in order to make an exploratory analysis of the factors affecting the style of Japanese prime ministers’ speeches.

\(^10\)Incidentally, we also carried out the classification experiments using SVM, which, in fact, showed better classification performance than random forests for our data, but as classification performance depends on the data structure (Yu and Unsworth, 2007), investigating alternative methods is in itself important for stylistic text classification studies.

\(^11\)As we classified two prime ministers’ Diet addresses in relation to their respective contemporaries, the results imply that (a) Koizumi had a more individualistic style than Nakasone, or (b) Koizumi’s contemporaries had more monotonous styles than Nakasone’s contemporaries. We would like to confirm this point in a further study.

\(^12\)As we mentioned in footnote 1, the number of tokens could affect the classification performance, so we would like to examine this point in a future study.

\(^13\)In a study of authorship attribution in Japanese, the distribution of the first order particles was shown to lead to high performance (Jin, 1997), but, in our experiments, some other parts-of-speech like functional nouns and auxiliary verbs showed equal or better performance than particles in the classification experiments of both prime ministers. This could reflect the fact that we used spoken corpora (speeches), as opposed to the fact that the written corpora (diaries) was used by Jin (1997).
Table 3: Precision, recall rates, and $F_1$-values

<table>
<thead>
<tr>
<th>Category</th>
<th>Nakasone</th>
<th>Koizumi</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>precision</td>
<td>recall rate</td>
</tr>
<tr>
<td>functional nouns</td>
<td>83.3</td>
<td>100.0</td>
</tr>
<tr>
<td>conjunctions</td>
<td>41.7</td>
<td>50.0</td>
</tr>
<tr>
<td>adnominals</td>
<td>35.7</td>
<td>50.0</td>
</tr>
<tr>
<td>particles (1st order)</td>
<td>70.0</td>
<td>70.0</td>
</tr>
<tr>
<td>particles (2nd order)</td>
<td>70.0</td>
<td>70.0</td>
</tr>
<tr>
<td>auxiliary verbs (without stemming)</td>
<td>66.7</td>
<td>60.0</td>
</tr>
<tr>
<td>auxiliary verbs (with stemming)</td>
<td>77.8</td>
<td>70.0</td>
</tr>
<tr>
<td>particles + aux. verbs (1st order, without stemming)</td>
<td>54.5</td>
<td>60.0</td>
</tr>
<tr>
<td>particles + aux. verbs (2nd order, without stemming)</td>
<td>61.5</td>
<td>80.0</td>
</tr>
<tr>
<td>particles + aux. verbs (1st order, with stemming)</td>
<td>70.0</td>
<td>70.0</td>
</tr>
<tr>
<td>particles + aux. verbs (2nd order, with stemming)</td>
<td>77.8</td>
<td>70.0</td>
</tr>
<tr>
<td>function words (1st, without stemming)</td>
<td>72.7</td>
<td>80.0</td>
</tr>
<tr>
<td>function words (2nd, without stemming)</td>
<td>75.0</td>
<td>90.0</td>
</tr>
<tr>
<td>function words (1st, with stemming)</td>
<td>70.0</td>
<td>70.0</td>
</tr>
<tr>
<td>function words (2nd, with stemming)</td>
<td>77.8</td>
<td>70.0</td>
</tr>
</tbody>
</table>

$F_0$-values = $\frac{B^2 + \beta}{B^2 + 1} \times P \times R$ (P: precision, R: recall rates)

Table 4: Nakasone’s top ten variables with high variable importance

<table>
<thead>
<tr>
<th>Variables</th>
<th>Translation</th>
<th>Part-of-Speech</th>
<th>$V_{max}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>masu -</td>
<td>auxiliary verb</td>
<td>0.0151</td>
</tr>
<tr>
<td>2.</td>
<td>wareware we</td>
<td>noun-pronoun</td>
<td>0.0146 +</td>
</tr>
<tr>
<td>3.</td>
<td>de with, by</td>
<td>particle-case</td>
<td>0.0120 -</td>
</tr>
<tr>
<td>4.</td>
<td>i jou more</td>
<td>noun-dependent</td>
<td>0.0089 -</td>
</tr>
<tr>
<td>5.</td>
<td>watashi I</td>
<td>noun-pronoun</td>
<td>0.0070 -</td>
</tr>
<tr>
<td>6.</td>
<td>beshi should</td>
<td>auxiliary verb</td>
<td>0.0069 +</td>
</tr>
<tr>
<td>7.</td>
<td>niokeru in</td>
<td>particle-case</td>
<td>0.0042 +</td>
</tr>
<tr>
<td>8.</td>
<td>u would, shall</td>
<td>auxiliary verb</td>
<td>0.0039 +</td>
</tr>
<tr>
<td>9.</td>
<td>niote in</td>
<td>particle-case</td>
<td>0.0038 +</td>
</tr>
<tr>
<td>10.</td>
<td>soko there</td>
<td>noun-pronoun</td>
<td>0.0035 +</td>
</tr>
</tbody>
</table>

Table 5: Koizumi’s top ten variables with high variable importance

<table>
<thead>
<tr>
<th>Variables</th>
<th>Translation</th>
<th>Part-of-Speech</th>
<th>$V_{max}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>sono its, the</td>
<td>adnominals</td>
<td>0.0274 -</td>
</tr>
<tr>
<td>2.</td>
<td>ara -</td>
<td>auxiliary verbs</td>
<td>0.0238 -</td>
</tr>
<tr>
<td>3.</td>
<td>da be</td>
<td>auxiliary verbs</td>
<td>0.0183 -</td>
</tr>
<tr>
<td>4.</td>
<td>mata and, in addition</td>
<td>conjunctions</td>
<td>0.0161 -</td>
</tr>
<tr>
<td>5.</td>
<td>wagakuni my country</td>
<td>nouns</td>
<td>0.0139 -</td>
</tr>
<tr>
<td>6.</td>
<td>koto thing</td>
<td>nouns-independent</td>
<td>0.0099 -</td>
</tr>
<tr>
<td>7.</td>
<td>desu be</td>
<td>auxiliary verbs</td>
<td>0.0088 +</td>
</tr>
<tr>
<td>8.</td>
<td>beshi should</td>
<td>auxiliary verbs</td>
<td>0.0078 -</td>
</tr>
<tr>
<td>9.</td>
<td>ga -</td>
<td>particles-conjunctive</td>
<td>0.0068 -</td>
</tr>
<tr>
<td>10.</td>
<td>tai would like to</td>
<td>auxiliary verbs</td>
<td>0.0043 -</td>
</tr>
</tbody>
</table>

The style of Nakasone

Pronouns

Nakasone’s top ten variables with high variable importance include two first-person pronouns, namely ‘wareware (we)’ and ‘watashi (I)’. We plotted the transitions of these words in the addresses from 1945 to 2006 in Figure 1. The figure shows that Nakasone’s speech style was characterised by frequent use of the first-person plural pronoun ‘wareware’ and infrequent use of the first person singular pronoun ‘watashi’, compared to his contempo-
‘Watashi’ clearly represents prime minister himself, and most of his uses of ‘wareware’ in the addresses represent generic meaning; namely I + all Japanese people or people in the world. These results could indicate that he preferred mentioning people rather than himself, and in that sense that he was very conscious of people in his addresses.

Modality

Nakasone’s top ten variables with high variable importance include two modality expressions, namely ‘beshi (should)’ and ‘u (would, shall)’. We plotted the transitions of these words in the addresses from 1945 to 2006 in Figure 2. The figure shows that Nakasone’s speech style was characterised by frequent use of ‘beshi’ and ‘u’, compared to his contemporaries. ‘Beshi’ is a normative modality, representing duty or obligation in Japanese. The figure shows that he used ‘beshi’ especially often in his first, second and third addresses, and his special uses of ‘beshi’ namely preaching people’s or ‘our’ duty in relation to ethical, humanitarian, or social issues were included in these addresses, but rarely appeared in the addresses of his later tenure or of other prime ministers. This result could indicate that he preferred to mention people’s duty, and in that sense preferred to speak from top to bottom in a paternalistic way.

Figure 2 shows that he used ‘u’ more frequently in his second and fourth addresses, many of his uses of this word in these two addresses were in the collocation ‘shou-ka’, which rarely appeared in the addresses of his later tenure or of other prime ministers. This collocation is used for inquiring in Japanese, and in fact in his addresses, this collocation was frequently used for questioning people about the current and future social situation in the country, and on what ‘we’ should do in regards to the future, thus his uses of this collocation could also indicate a kind of paternalistic manner of his speech. The following are passages from his addresses:

Masa ni gendai ni ikiru seijika to kokumin ga, wareware wa, kojin no soncho to iu kiso no iu tachitsutsu, mou ichido, kojin wo hagukun de kururu katei, komyuniti, kuni ni tsuite omoi wo itasu beki dewa nai de syo-u-ka.

(We should mention that politicians and people living in the contemporary world have a serious responsibility towards our descendants and the (future) history of our country.)

Nakasone was known to be a conservative politician as well as a powerful prime minister with a strong leadership style, and he knew the importance of the role of speeches as well as of the media for leaders (Shinoda, 1994; Suzuki and Kageura, 2007; Takase, 2005). The results we obtained indicated that he preferred to show consciousness of people, and to question and lecture people about their duties. These results could be interpreted as represent his political style; he tried to show strong leadership guiding people in a paternalistic way, and tried to position himself as a powerful top-down leader of the country. From these analyses, we conclude that his political style can be characterised as a people-preaching political style.

The style of Koizumi

Connecting expressions

Koizumi’s top ten variables with high variable importance include two connecting expressions, namely adnominal ‘sono (its, the)’ and conjunction ‘mata (and, in addition)’. We plotted the transitions of these words in

14 It was based on concepts articulated by Íñigo Mora (2004); in her analysis of the speeches in the British Parliament, she divided the use of ‘we’ into four different types of references, namely (a) exclusive (I + my group), (b) inclusive (I + you), (c) parliamentary community (I + parliamentary community) and (d) generic (I + all British people).

15 Four of ten uses in his first address, four of eighteen uses of his second address, and four of twelve uses in his third address were used in that way, in other cases including his other addresses, ‘beshi’ was used for mentioning, for example, governmental policy agendas.

16 Five of nine uses in his second address, and three of eight uses in his fourth address were used in this collocation.

17 Cited from the addresses given by Nakasone at 10 Sep. 1983.

18 Cited from the addresses given by Nakasone at 24 Jan. 1983.
the addresses from 1945 to 2006 in Figure 3. The figure shows that Koizumi’s speech style was characterised by infrequent use of adnominal ‘sono’ and conjunction ‘mata’, compared to his contemporaries. ‘Sono’ is an expression assigning the former sentences or former words, and ‘mata’ is an expression connecting sentence after sentence additively. Both of these expressions make sentences more complicated in Japanese. These results correlate with the finding by a previous study that Koizumi used shorter sentence lengths than any other Japanese prime minister after World War II (Suzuki and Kageura, 2008), as infrequent use of these expressions results in shorter sentence lengths. These results could indicate that he preferred to use simple and easy-to-understand speech styles rather than complicated ones.

**End-of-sentence expressions**

Koizumi’s top ten variables with high variable importance include ‘aru’, ‘da’ and desu’. We plotted the transitions of these words in the addresses from 1945 to 2006 in Figure 4. Combining ‘da’ and ‘aru’ with ‘masu’ creates the collocation ‘de (da)-ari-masu’, a formal end-of-sentence expression, and in fact this collocation is a typical expression in Japanese Diet speeches (Azuma, 2006). The results showed that Koizumi, instead of this expression, frequently used the end-of-sentence expression ‘desu’ in his addresses, a neutral expression that is used in ordinary conversation in Japanese.19 These results could indicate that Koizumi preferred informal speech style even in Diet speeches. The following are passages from his addresses.

Kodomo wa syakai no takara, kuni no takara desu.20

(Children are precious for society and the nation.)

Watashi wa, kokumin no shijii wo haikai ni, seisiki naki kozo kaikaku wo susume masu. Honban wa korekara desu.21

(I urge structural reform with nothing sacred, supported by the people. It is time to urge full-scale reform.)

Koizumi is regarded as having changed Japanese politics and succeeded politically partly because he was able to attract people by his way of speaking (Azuma, 2006; Takase, 2005; Otake, 2006; Kabashima and Steel, 2007; Suzuki and Kageura, 2007). The results we obtained indicate that he preferred speaking in a simple, understandable and informal manner in Diet speeches. These results could be interpreted as representing his political style; he tried to inspire people’s interest in politics by his easy, understandable and informal speeches, and tried to attract people in order to obtain their support. From these analyses, we conclude that his political style can be characterised as a people-talking style.22

In this section we provided some sociolinguistic interpretation of extracted features. These features include functional expressions frequently discussed in sociolinguistic analyses, and characterise well the individual

---

19This point has already been mentioned by Azuma (2006), though he examined smaller samples.
20Taken from an address given by Koizumi on 21 Jan. 2005.
21Taken from an address given by Koizumi on 27 Sep. 2001.
22It should be noted that it is widely known that Koizumi was more successful than Nakasone in his media strategy including his use of speeches (Azuma, 2006; Takase, 2005; Otake, 2006; Suzuki and Kageura, 2007). This could have imply that Koizumi’s people-talking style could attract people better than Nakasone’s people-preaching style.
speech styles of the two prime ministers, as well as shedding light on their individual political styles. We conclude that our method, complemented with qualitative analyses, is useful for sociolinguistic analyses of political speeches.

5 Conclusion

In this study we extracted speaker-specific functional expressions from political speeches using random forests in order to investigate individual political styles of two prime ministers. By extracting the important variables using random forests, we extracted the expressions specific to the two prime ministers, which clarified their individual political styles. Our method can be used for sociolinguistic analysis of various other types of texts, and in this way this study will contribute to developing the area of computational sociolinguistics in the field of stylistics.

In this study we used the uni-gram of Japanese morphemes for our analyses, as we thought it was a basic and appropriate unit for our purposes. In some cases such as end-of-sentence expressions, however, our results indicated that a longer unit (collocation) could be more useful for sociolinguistic interpretation, thus we would like to compare the results of this study to those using longer units of functional expressions in a further study. We will also develop the method to decide thresholds of extracted variables for more systematic analyses.

6 Acknowledgement

This study was supported by a Suntory Foundation Research Grant, 2007-2008. We would like to express our gratitude for this support. An earlier version of this study was presented at the 14th Annual Meeting of Japanese Natural Language Processing (NLP2008) at the University of Tokyo and Digital Humanities 2008 (DH2008) at the University of Oulu. We would like to thank the Association for Literary and Linguistic Computing for the Bursary Award at the latter conference. Ms. Jean Anderson for her help, and participants in both conferences for their useful comments. We would also like to thank Professor Akiko Tanaka at the University of Tokyo and Professor Kyo Kageura at the University of Tokyo for their useful comments.

References


in authors and their characters’. Digital Humanities 2007 pp. 8–10.


