

Individual usage: a corpus-based study of idiolects

Michael Barlow

University of Auckland
mi.barlow@auckland.ac.nz

The term idiolect is well-known in linguistics and is distinguished by the fact that there is probably no other linguistic term in which there is such a large gap between the familiarity of the concept and lack of empirical data on the phenomenon. We trace some of the competing ideas on the relationship between the language of the individual and the language of the group and then turn our attention to a corpus-based investigation of some aspects of lexical/syntactic idiolectal variation. In this study, we examine the speech of five White House Press Secretaries and show that individual differences are not based on the use of a few idiosyncratic phrases but involve a wide range of core grammatical constructions. The results also suggest that the idiolectal patterns are consistent and are maintained over a period of at least a year or two.

1. Introduction

The notion of idiolect is well-known in Linguistics, although a precise definition and the very existence of the phenomenon is a subject of debate. Dittmar (1996: 111) offers a definition of idiolect as:

the language of the individual, which because of the acquired habits and the stylistic features of the personality differs from that of other individuals and in different life phases shows, as a rule, different or differently weighted [communicative means].

There are both theoretical and practical reasons for the almost complete absence of research into the nature of individual grammars. The past lack of interest in idiolects derives from the difficulty in obtaining appropriate data and, on a theoretical level, it arises in some cases from a general dismissal of usage as being uninteresting and in others from an understandable focus on the general rather than the particular. Linguists are interested in language and languages and those interests naturally lead to abstractions and idealisations, which often preclude individual grammars. Many linguists have denied the existence of idiolects. Barthes (1977:21) states that “the idiolect would appear to be largely an illusion.” In a similar vein, Jakobson (1971:82) notes that the notion of idiolect “proves to be a somewhat perverse fiction”. These particular negative positions may have arisen in response to Bloch's original definition of an idiolect as “the totality of the possible utterances of one speaker at one time in using a language to interact with one other speaker” (1948).

The denial of the existence of idiolectal grammars may be due to an emphasis on the importance of the language community and from this standpoint the issue is perhaps best represented not as whether idiolects exist or not, but whether it is the language of the individual or the language of the group that is primary. The most common position in linguistics has been to view language as an abstract social construct: Saussure's *la langue* in various guises. For instance, Labov has consistently argued that the language of the group is the proper object of study. He states “language is not a property of the individual, but of the community. Any description of a language must take the speech community as its

object if it is to do justice to elegance and regularity of linguistic structure” (1989:52). The first statement can be taken to be an assumption or axiom related to sociolinguistic studies in the Labovian mode. While the second statement can be interpreted in a variety of ways, it is clear that the main idea is that regularity is only associated with the language of the group and the suggestion appears to be that individuals, as members of different intersecting groups, will exhibit seemingly random irregularities. The opposing position is taken by Hudson (1996:29) who argues that language must be in the individual rather than in the community and on this view the idiolect is the primary reality and it is groupings of idiolects that constitute different social varieties. I tend to agree with Hudson and assume that it is the idiolect that is logically prior, based on the reality of individual usage. Taking this standpoint, we can then examine the language of individuals to assess the evidence for sociolects for specific aspects of grammar. (See Brezina 2010 for an example of this approach in relation to epistemics.)

There are a variety of issues concerning idiolects and sociolects, which while following a similar conceptual path are somewhat distinct. We briefly outline these points here and return to them at the end of the paper. The different strands in the idiolect-sociolect discussion relate to (i) conventionality and communication, (ii) regularity and stability, (iii) the reality of sociolects, and (iv) variation and idiolects/sociolects .

As mentioned above, one well-established argument for a social view of language as the appropriate object of study for linguists goes back to Saussure and rests on the idea of meaning as a set of conventions supported by a social group. Closely related to this line of argument is the fact that language is not private (Jakobson 1971); it is successfully and primarily used for communication within a group. Based on this observation, the argument is made that in order for communication to occur, language must be a property associated with the group and not with the individual.

A second line of reasoning concerns the lack of regularity and stability in idiolects. The lack of regularity is taken for granted based on observable idiosyncrasies of speakers and the lack of stability is based on the fact the language of individuals changes over time. Most problematic for linguistic investigations on this topic is the well-known fact that the language of an individual changes depending on the interlocutor and the general context. It is probably fair to say that this line of argument for the primacy of the group has more to do with problems of definition and with practical problems investigating idiolects than with theoretically-grounded concerns about individual grammars.

The third argument, reflecting Labov's position, is that sociolects are the natural level or organisation for language analysis. Labov (1989:52) states that “individual behaviour can be understood only as a reflection of the grammar of the speech community.” The social group is the only level where regularities emerge and that if idiolects are examined, the analyst is confronted only by idiosyncrasies and the coherence seen in the language of groups is obscured. Wunderlich notes the priority of sociolects and defines an idiolect as an individual's share of the sociolect (1996: 110).

A more subtle issue concerns the nature of the relation between the individual and the group with respect to variation. If we take as a starting point the existence of variation for some variable within a sociolect, then we can posit two extreme views of the nature of variation in idiolects. One is that each idiolect exhibits the same form of variation as the sociolect. The second is that there is no variation within each idiolect and that it is only inter-individual variation which shows up as variation within a sociolect. Clearly, there are a variety of permutations of these patterns. Meyerhoff and Walker (2007) in a study of copula absence note: “These results reaffirm the validity of modelling variable rules in a community grammar, rather than as an aggregation of idiolectal norms.” My claim is that much more

extensive data is needed in order to get a picture of the relationship between sociolects and aggregations of idiolects.

Here we will take a cognitive sociolinguistic position and explore the issue variation in idiolects empirically from the standpoint of usage-base account of language. In this paper we examine frequencies of use of lexical and syntactic patterns and show that

- (i) individual patterns appear to be stable over a period of at least a year or two
- (ii) individual patterns of variation are maintained despite differences from general usage within the community
- (iii) inter-speaker variation is typically greater than intra-speaker variation based on samples taken over different time periods
- (iv) idiolectal variation is based on core aspects of language and not on peripheral idiosyncrasies.

Understanding the nature of idiolects is necessary in order to gain a richer understanding of a variety of issues in addition to those outlined above. For instance, knowing the ways in which language production differs from speaker to speaker will further understanding of the distinctions between comprehension and production. Thus while corpus analyses reveal the patterns of language that under a usage-based model of grammar (Langacker 1988, 2000, Bybee 1998, Kemmer and Barlow 2000) are instrumental in structuring linguistic knowledge, this knowledge relates mostly to comprehension. Since corpora are amalgamations of the speaking or writing of different people, the patterns that are extracted from corpus data tell us primarily about language understanding. It is assumed that productive patterns are some subset of comprehension patterns of language, but this notion is very vague and only with a detailed knowledge of idiolects can we understand the connection between comprehension and production in grammatical terms.

The working hypothesis for this investigation is that the variation that distinguishes individual speakers lies in the profile of the central components of lexicogrammar, and not only in some idiosyncratic peripheral phraseology. On the basis of earlier preliminary study (Barlow and Kemmer 2004), it appears to be worthwhile to follow this centrality hypothesis and so we focus on high frequency items rather than search for low frequency (but highly distinctive) markers of the speech of individuals. It is also assumed that because we are working with high frequency items, differences in individual profiles are likely to be based on preferences rather than absolutes, which translates into marked differences in the frequency of use of particular patterns. What we find is that for some linguistic item, two or more speakers might show marked preferences or dispreferences for the item, with other speakers exhibiting smaller differences in usage. Naturally, given a series of linguistic items, this will amount to idiolectal variation.

2. Method

To pursue this study, the speech patterns of five White House Press Secretaries are investigated. There are several advantages of using this spoken data. One is the fact that the transcripts, which are sufficiently accurate representations of the spoken interactions for the present study, are available and readily accessible. Second, the volume of speech transcribed is considerable and we can work with individual speech corpora between 200,000 and 1,200,000 words of running text. A third and very important advantage of working with this dataset is that the context of the discourse is held constant across the different samples and different speakers. The content changes, of course, from questions about Monica Lewinsky to the war on terror, but the overall format of press conferences does not change and virtually all the discourse involves the press secretary being questioned closely by members of the press. The identity of the press corps members is not known, but they tend to remain in their jobs longer than the press secretaries, who rarely last much more than a year or so in their post.

2.1 The Idiolect Collection

In order to reduce the likelihood of priming from one day to the next, transcripts from consecutive days were not chosen, which means that there is a gap of at least one day between the speech events collected; and, in fact, the typical interval might be two or three days and occasionally there is a longer gap.

The five press secretaries chosen for the study are: Mike McCurry (1994-98), Ari Fleischer (2001-03), Scott McClellan (2003-06), Tony Snow (2006-07), and Dana Perino (2007-08). The choice was made on the basis of the length of their tenure since one goal of this study is to work with reasonably large samples of idiolects taken over the course of several months at a minimum. The sampled data covers around a year or more of speech output for each speaker.

The press conferences occasionally contain a short statement made by the press secretary, but mostly the format involves answering questions posed by journalists. The length of each press conference varies, but they usually last 30 to 50 minutes. A small portion of a press conference is illustrated in Figure 1.

Q Just back to the National Guard for a moment. You said it's a success story, they moved in quickly. But could you have gotten equipment there any sooner were we not in Iraq? I mean, could it have been closer, would there have been units that were closer, equipment closer?

MR. SNOW: I don't -- that's a hypothetical question that I'm not sure --

Q It's not hypothetical.

MR. SNOW: Well, it is -- let me put it this way. We have no indication that people did not get what they needed as soon as they needed it.

Figure 1: Portion of the original transcript

It is clear that the quality of the transcription does not approach that of transcriptions carried out by discourse linguists: there is no indication of overlapping speech, pause length, or backchannel information. The transcriptions are produced by 4 or 5 stenographers whose goal is to provide a verbatim record of the press conference, an aim that journalists are also keen to see implemented. Corrections, if made, are most likely to be in names. False starts are recorded, as shown above, but pauses and hesitations within words are not included (Smolkin 2000). And, as can be seen in the sample above, the transcription is presented in the form of sentences and paragraphs rather than intonation units. Nevertheless, the retention of false starts within the transcript suggests that it is a reasonable rendition of the language of the press conferences and its imperfections notwithstanding, the data are sufficiently robust to be used for lexical and grammatical analysis.

The conventions used in original transcripts to indicate who is speaking vary in their implementation to some extent, but generally a question asked by a reporter is introduced by Q positioned at the beginning of a line and the response by the Press secretary is indicated by his or her name (e.g., MR. SNOW). Other administration officials sometimes take part in the press conference and in this case their contribution is labelled using their name or title.

For the purposes of this study the transcriptions were edited in order to remove all the speech not

belonging to the press secretary. The result, which perhaps cannot be described as a corpus in that it no longer represents language as a discourse, looks decidedly odd, as can be seen from the snippet displayed in Figure 2.

```

<Q>
<SNOW> I don't -- that's a hypothetical question that I'm not
sure --
<Q>
< SNOW> Well, it is -- let me put it this way. We have no
indication that people did not get what they needed as soon as they
needed it.

```

Figure 2: Transformed transcript

The outline of the structure of the discourse is retained only with respect to turn taking. The actual content of the speech of the reporters or other administration officials has been removed. Thus the data has only one use: the analysis of the utterances of each press secretary. All interactions not involving the press secretary are indicated by <Q>, even in those cases where the contribution is a statement by a White House Official and not a question. Eliminating the contributions of the interlocutors makes the ngram and concordance analysis quite simple and robust.

The speech samples were split into text files containing 200,000 words each. For two speakers there is just a single 200,000 word file, but for the other speakers there are three, four and six such files, enabling comparisons between the speech of one individual at different times and the speech of different individuals.

The files were tagged for POS using the CLAWS7 tagset and for semantic tags using the USAS tagset. The tagging was performed using the tools at the Wmatrix site, <http://ucrel.lancs.ac.uk/wmatrix/>, (Rayson 2008). The files were analysed using the concordance program MonoConc Pro and the collocation/ngram analyser Collocate.

The makeup of the idiolect collection is as follows:

Press Secretary	Number of Samples (200K words)	File names
Tony Snow	1	T1
Mike McCurry	6	M1 M2 M3 M4 M5 M6
Scott McClellan	3	S1 S2 S3
Ari Fleischer	4	A1 A2 A3 A4
Dana Perino	1	D1

Figure 3: Composition of the idiolect collection

The files were analysed initially using frequent bigrams and trigrams; frequent POS bigrams; and on the basis of these results, specific concordance searches were carried out to probe particular constructions.

3 Results and Discussion

3.1 Ranked Word Bigram Profiles

The bigrams for each sample were extracted using Collocate. Setting a minimum frequency of 7 produced around 4000 bigrams per sample, which were ranked according to decreasing frequency. Typically, the most frequent bigrams occurred over a 1000 times. While bigrams do not themselves reflect grammatical units, they can be used as indicators of grammatical patterns and, as we will see, they can also be used as a fingerprint associated with each speaker.

In the first analysis of the spoken data, the bigrams were ranked according to frequency and the top 10 bigrams from each sample were combined to create an amalgamated list of the bigrams. Naturally there are many overlaps in the bigrams retrieved from each sample -- *of the, in the, to be, the president* – and so the duplicate bigrams were removed, leaving a list of 15 unique bigrams, representing the most frequent bigrams from the different samples. This list was then used as a metric against which each sample was measured in the following way. Using the list of bigrams, we determine the ranking of each bigram in each sample. For example, the bigrams *of the, the president* and *I think* have the following rankings in the different files.

	T1	M1	M2	M3	M4	M5	M6	S1	S2	S3	A1	A2	A3	A4	D1
of the	2	1	1	1	1	1	1	2	2	2	2	2	2	2	2
the president	3	2	2	2	2	2	2	1	1	1	1	1	1	1	1
I think	7	6	10	14	4	3	3	3	3	3	4	5	6	4	3

Figure 4: Ranking of bigrams in each sample

The table in Figure 4 can be seen as a bigram profile of each file or sample. If we take the sample M1 (Mike McCurry sample 1) we see that it is characterised with the rankings 1 2 6 for the three bigrams; M2 has the values 1 2 10; and M3 has the rankings 1 2 14. It is worth repeating that the ranking of bigrams is accomplished by extracting a set of bigrams from all the files and then recording the rank of each of the selected bigrams in each file. The full ranked bigram profile is produced by adding more data points and in the first probe we use 15 very common bigrams.

The goal is to compare the speech of two different speakers with the speech of one speaker on two separate occasions (in a similar context). What might we expect to find? First, it might be thought that since the bigrams are very common and mostly represent grammatical sequences, then all the file samples will show similar patterns, with some noise or other variation thrown in. On this view the frequent bigrams are expected to follow the patterning of single words in which the same few grammatical words float to the top of general word frequency lists. Alternatively, it would also be reasonable to expect the opposite: that there would be quite a bit of variation from one sample to the next, reflecting the fact that each press conference deals with different issues, which in turn require different kinds of explanations and therefore different patterns of language use that exceed any variation attributable to individual speakers.

The ranked bigram profile for the first speech sample for Mike McCurry, Mike 1, is given in Figure 5. The list of bigrams, *the president, of the, going to, in the, I think, and the, to the, that we, on the, that the, think that, to be, I don't, continue to, and we have*, are arranged along the x axis and the rank of the bigrams, up to 120 is given on the y axis.

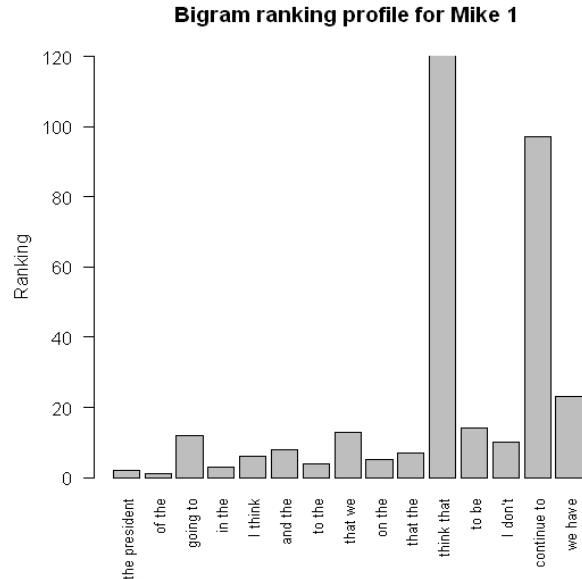


Figure 5: Ranking of 15 bigrams for one speech sample

We are now in a position to compare the rankings for different samples from the same speaker and samples from different speakers, as depicted in Figure 6. The figure contains six profiles, selected as representative of the overall results. The left hand side shows the profiles for three samples of the speech of Mike McCurry; the right hand side contains profiles of three different press secretaries: Ari Fleischer, Scott McClelland and Dana Perino. In these profiles, the y axis once again represents rank, which means that the shorter the bar, the more frequent the bigram. The bars extending all the way up represents bigrams with a ranking greater than 60.

The results, taken as overall patterning of the data, are illuminating. Concentrating first on Mike McCurry's speech output represented by the ranked bigram graphs, we can clearly see that the profile is very similar for each sample even though the bigrams selected are so common as to be unremarkable in themselves. Thus based on the relative frequency of use of the most frequent bigrams, we can say that the speech of one individual is very similar from one situation to the next, at least when the general context remains the same. By similarity in speech, we mean that the speaker prefers some patterns to others and does that rather consistently even when the content of the utterances changes. We should emphasise that the data collected in the different files covers a span of several months, not days, which means that the patterns are entrenched and are not ephemera resulting from priming.

In contrast to the consistent patterns for the ranked bigrams for Mike McCurry, the profiles on the right hand side of the figure show that different speakers exhibit distinguishable profiles. Even though we are using a very crude tool and a small selection of frequent bigrams, we are, I would claim, seeing some evidence of stable individual differences. The stability is visible in the strong similarity of the different samples of speech of Mike McCurry, while the inter-speaker variability stands out in the markedly different profiles associated with the other speakers. This contrast between intra-speaker stability and inter-speaker variation points to significant differences in production among individuals. The bigrams themselves are obviously not the units of grammar. Nevertheless, they are powerful indicators of units of grammar since differences in production of constructions and other grammatical units will surface as differences in bigram distributions.

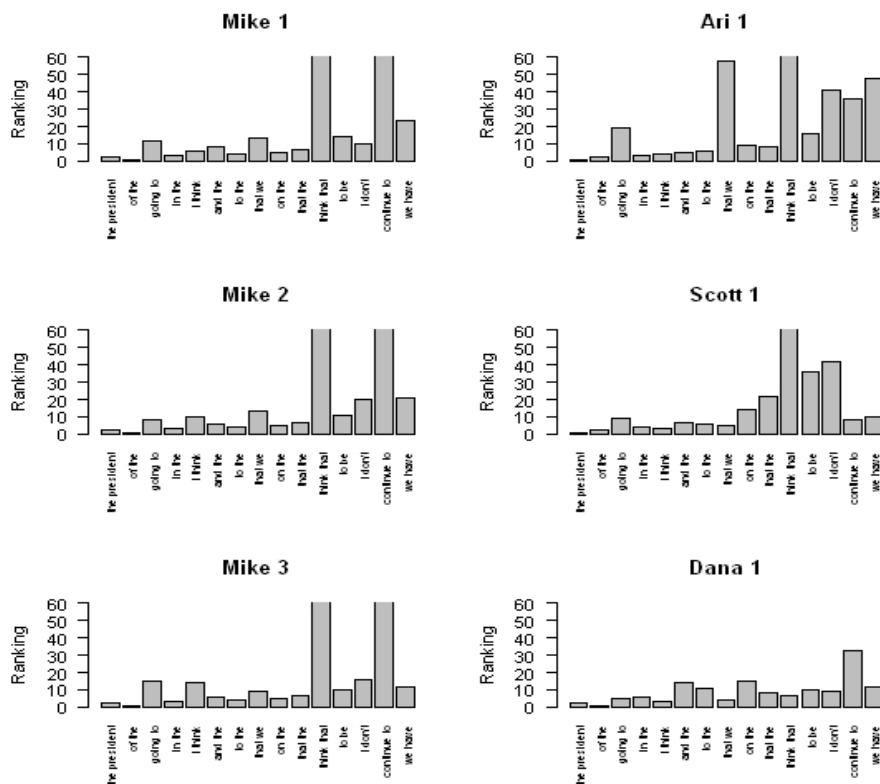


Figure 6: Six speech samples

The results show that the frequent language patterns used by the different speakers all involve frequent use of these fifteen bigrams, but the proportion in which they are used varies from one speaker to the next, but remains fairly constant for each speaker. This simple probe confirms the existence of idiolectal language patterns and provides some clues as to their nature. Most of the bigrams are typical function word sequences, suggesting that the consistent patterning is likely to result from extensive preferred grammatical sequences. The alternative explanation is that the differences observed here are the consequence of the repetition of a limited number of highly frequent lexical phrases. In other words, these results might be said to merely reflect the speakers' preferences for particular lexical phrases or idioms (*as a result, in terms of, in spite of, etc.*) and that it is a small number of lexical units such as these rather than grammatical structures that lead to the patterns observed in Figure 6. The presence of bigrams such as *the president* and *continue to* in the list suggest that there is a lexical component, but also contained within the list are *of the, to the, in the, to be, etc.* Thus the evidence points to a normal mix of lexical and grammatical structures. While it is not surprising to find that individual speakers have their own characteristic phrases and favourite expressions, the fact that the tracks of the preferred speech patterns of an individual are apparent even in the use of the most common bigrams suggests that idiolectal variation goes well beyond idiosyncratic phrases, on the one hand, and sociolinguistic groupings on the other.

Let us examine a more complex data set involving the 46 most frequent bigrams. These are: *a lot, able to, about the, and i, and that's, and the, and we, at the, by the, continue to, for the, forward on, going to, have a, have to, i don't, i think, i'm not, in the, is a, it is, move forward, not going, of the, on the, president has, some of, talked about, that the, that they, that we, the president, the world, there are, think that, this is, to be, to do, to get, to the, trying to, want to, we are, we have, will be, and with the.*

The bigrams were selected by the same method as for the 15 bigram set: by combining the most frequent bigrams from all the speech samples to come up with single set of the topmost frequent word

pairs. In Figure 7 we see the six samples of the speech of McCurry. The content of the profiles is similar to that in Figures 5 and 6. The layout, however, is somewhat different. The ranking of each bigram is indicated by a line rather than a bar and the y axis is orientated with low rank values at the top. The bigrams themselves are not shown in the graph, but are ordered alphabetically along the x axis.

There are, not surprisingly, some differences among the profiles for the six speech samples, but overall they are remarkably similar. What we see in Figure 7 is evidence of the stability of individual preferences in ways of speaking over a period of a year or more. The patterning here is a reflection of the spoken output of McCurry in his role as press secretary and is not simply a consequence of the context of press conferences.

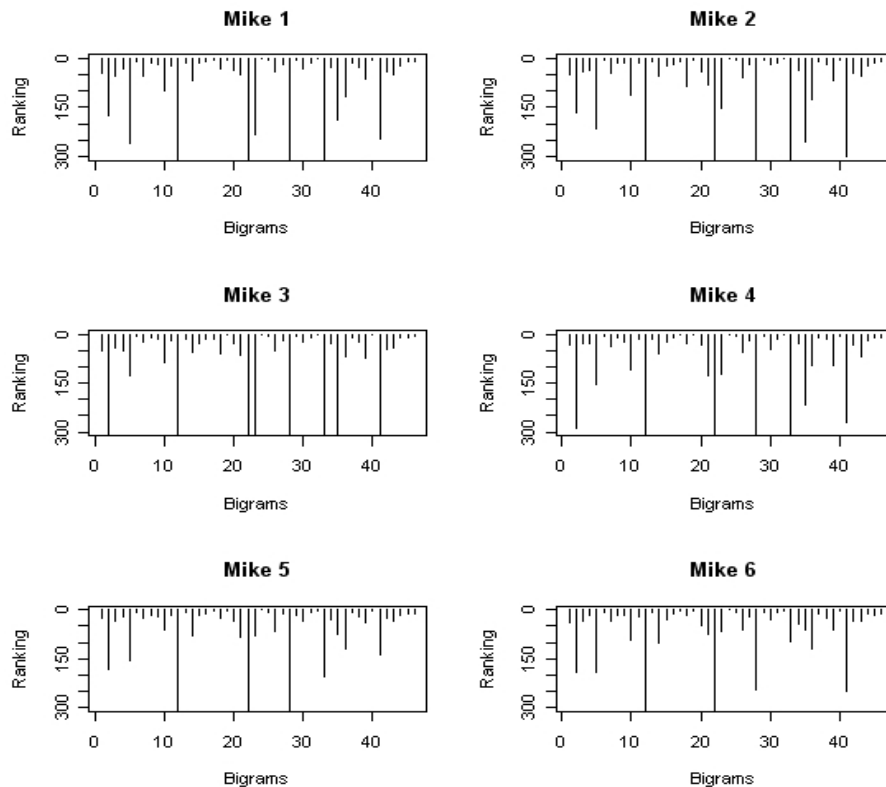


Figure 7: The ranked bigrams for six samples of the speech of McCurry

It might be argued that this ranked bigram profile merely represents the language patterns that arise naturally when dealing with the issues that arise at White House press conference, but if we compare the profiles of the three different press secretaries, we see how much they differ from each other. Figure 8 contains the ranked bigram profiles for three samples from Mike McCurry, three from Scott McClellan, and three from Ari Fleischer. Scanning vertically from one graph to the next, we find very similar patterns and scanning horizontally, on the other hand, we can plainly see the different patterns associated with the different speakers.

The ranked bigram plots in Figure 8 are simple representations, yet they confirm the initial findings based on the analysis based on 15 common bigrams and support the notion of substantial inter-speaker differences in speech production. It is to be expected that not all bigrams will be equally effective in distinguishing speakers and it can be seen that some regions of the graphs are similar across the different speakers. The bottom-up approach adopted here involves making as few assumptions as possible and using a mass of data selected solely on the basis of frequency.

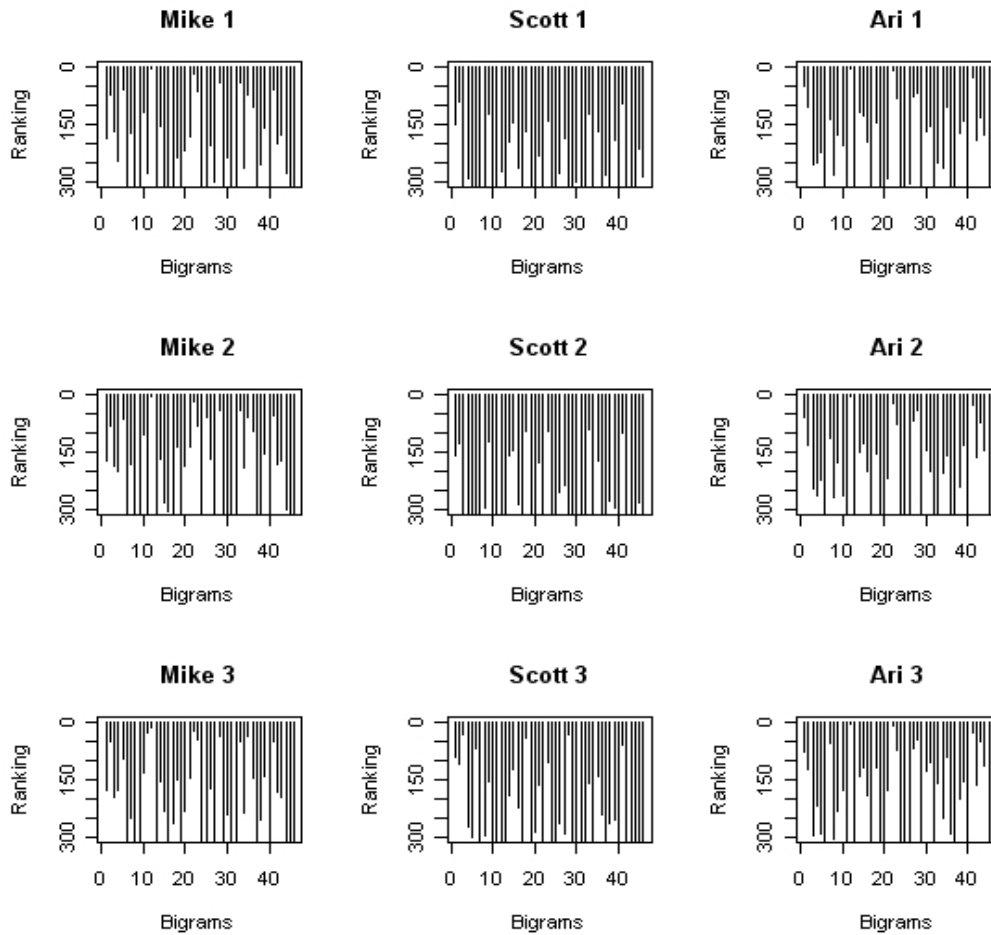


Figure 8: Ranked bigram profiles for three samples from three speakers

3.2 Bigram and trigram analysis

On the basis of these diagrams, we can confirm that the patterns of speech of individuals is recognisable. In order to confirm the visual patterning and to discover more detail on the relationship between bigrams and individual speakers, we can examine the multidimensional relationships between each sample and each set of bigram frequencies and for this probe we use all the data: 15 samples from the 5 different speakers. The data is analysed using a form of correspondence analysis based on measurements of the chi-squared distance between samples and bigram frequency (Baayen 2008). Each sample is judged in terms of its similarity to other samples based on the bigram frequency data and the bigrams are compared against each other by reference to the similarity of the data from each speaker sample. The result is a multidimensional space, which is reduced to two dimensions representing the two strongest factors, as illustrated in Figure 9. We can observe that the main factor on the horizontal axis accounts for 35% of the variance and, taken together with the vertical axis, over 60% of the variation in the data is accounted for by these two dimensions. Most importantly, the diagram shows that the sets of bigram frequencies lead to samples from the same speaker to be clustered together in the same region and samples from different speakers to be distributed in distinct locations of the two-dimensional plot.

these lexical phrases is the cause of the distinctive patterning for the different speakers. To ensure that this is not the case, we can simply delete the data for these two bigrams and perform the correspondence analysis again on the remaining 44 bigrams. The consequence of removing these two bigrams from the dataset is slight: the general configuration of the samples remains essentially as in Figure 9, but with a small spreading apart of the samples of Scott's speech. The minor change in the results is not surprising since the two matrices being analysed are very similar, a 15 X 44 matrix and a 15 X 46 matrix, but it is useful to confirm that it is the overall contribution of a set of bigrams that leads to the distinguishability of the speakers.

As mentioned above, we can locate the position of bigrams on the plot and we are able to examine the output of the correspondence analysis to determine which bigrams contributed the most to each factor. However, when looking at these bigrams we want to distinguish those for which the variation in frequency was highest from sample to sample from those more telling bigrams in which the variation in frequency was highest from speaker to speaker. In other words, we need to assess intra-speaker and inter-speaker variability for the key bigrams. To do this we go back to the corpus and check the distribution of those bigrams that have the most weight in terms of distinguishing different speech samples in order to see whether they are, in fact, distinguishing one *speaker* from another. It turns out that the bigrams do generally distinguish speaker behaviour and even selecting, *of the*, the plainest and most frequent of all bigrams, leads surprisingly to evidence of stable speaker differences. (See Figure 10. The letters indicate the speaker and the number refer to different time samples.)

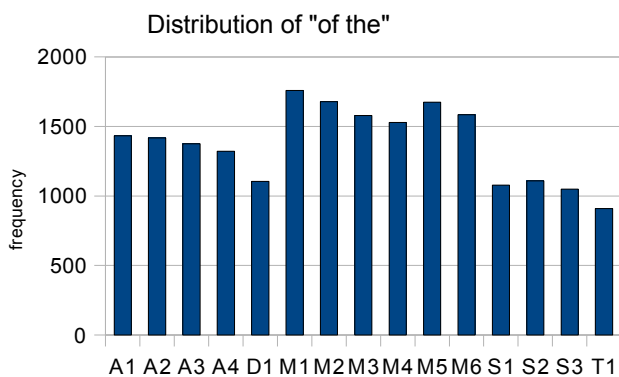


Figure 10: Frequency of *of the* in different speech samples

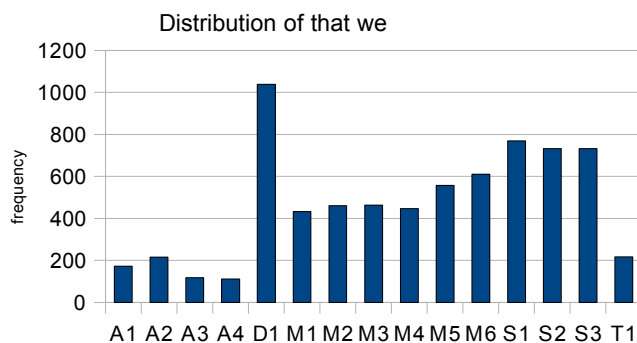


Figure 11: Frequency of *that we* in different speech samples

Figure 11 plots the frequency of the bigram *that we*, one of the bigrams contributing the most to Factor 1. There is some variation within samples of the same speaker, but the inter-speaker variation is much more pronounced. The bigram *that we* is particularly favoured by Dana; Scott also uses it quite heavily; and Ari uses it the least. In this graph and subsequent graphs we are looking for two things. One is the

consistency across the samples of the same speaker and the second is the variation across speakers. In this case we are able to see that the variation among speakers is greater than the variation for samples of a single speaker.

The frequency distribution of *we are*, *the president*, and *and that's* are illustrated in the bar graphs in Figure 12.

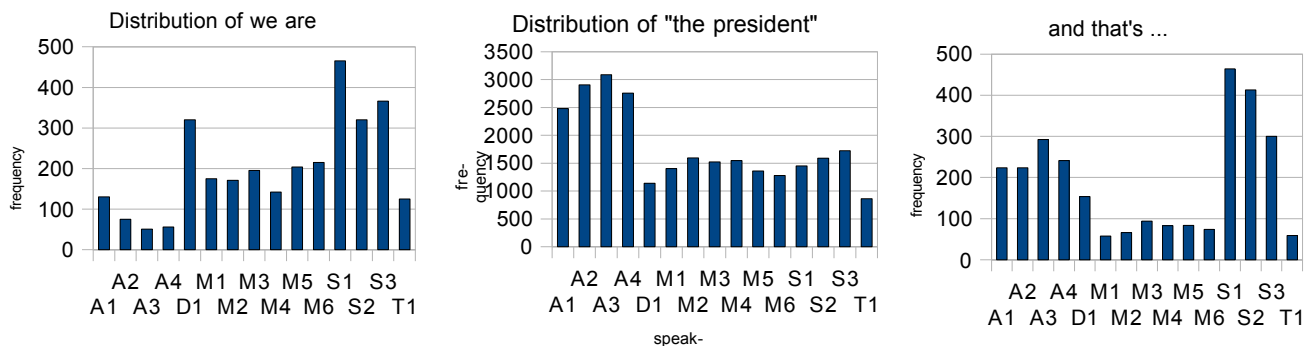


Figure 12: Frequency of *we are*, *the president*, and *and that's* in different speech samples

These bar graphs are representative of the bigram frequency distributions associated with the different speakers. Note that in the case of *we are* and *the president* the distributions are like two pieces of a puzzle and are in an almost inverse relationship. Dana and Scott favour *we are* and disfavour *the president*. Ari favours *the president* over *we are*; Mike is more neutral and doesn't show a strong preference. Tony is the exception in that he disfavors both *we are* and *the president*. The common bigram *and that's*, used to express consequences or to give an explanation, shows a fairly high degree of variation across speakers. These graphs, taken in combination, offers a good insight into the general patterns of variation that are found for the distribution of bigrams in the press conference data.

In general the patterns of trigrams mirrors those of bigrams, However, let us now turn to examine trigrams that according to the correspondence analysis play a lesser role in contributing to the variation existing across the speech samples. What characterises this group is lower frequencies of counts and the lack of extreme difference among the samples. One of these trigram phrases is *I don't know*.

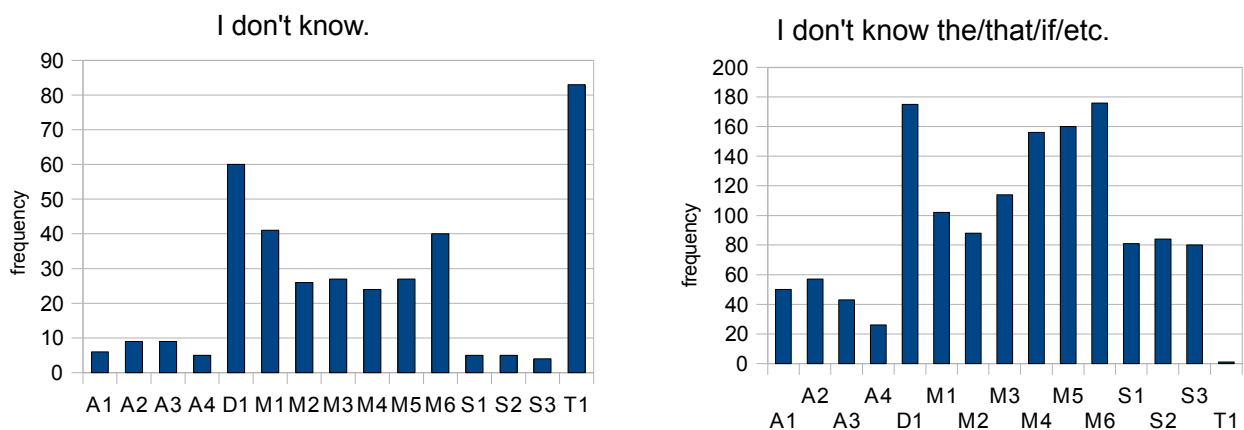


Figure 13: Distribution of *I don't know.* and *I don't know the/that/if etc.*

The plot for *I don't know* reveals some differences in frequency of use among speakers, but both the frequencies per sample are less than 10 for both Ari and Scott. Looking more closely at the corpus, we can distinguish *I don't know* period from *I don't know the/that/if etc.* The counts of *I don't know* period are rather low and it's clearly an expression to be avoided as far as press secretaries are concerned. We see that Mike McCarry uses the phrase relatively frequently in both its uses. For the other speakers,

however, there is a more marked difference in the use of the two forms with Tony exhibiting the most extreme variation. He is the most frequent user of *I don't know* – period and makes almost zero use of the pattern *I don't know whether* etc. Differences between Ari and Scott also arise in the counts of the *I don't know the/that/if* form.

3.4 POS Tag ngrams

The results obtained so far confirm that the patterns of speech of an individual are recognisable and the fact that we have probed the different samples using the most frequent bigrams and trigrams shows that that the differences in idiolects are not due to a few idiosyncratic phrases, but are due to differences in the preferences in the use of grammatical constructions in addition to differences in the use of common phrase like *in terms of* and *as a result*. The data analysed here constitute the first steps in isolating abstract functional dimensions underlying the low-level patterns depicted in these results. We have seen potential differences in preferred referential style in which the press secretaries' language exhibit preferences in the use of *I* versus *we* versus the third person NP, *the president*. Distinctions in the expressions related to reason-result and other causal expressions also appear to be present in the data but elaboration of these themes requires further detailed analysis which is beyond the scope of this paper.

The next probe is part-of-speech bigrams. The methodology remains the same: the ranking of frequent POS bigrams is determined for each speech sample and a master list of the most frequent POS bigrams is produced. These are: . cc, appge nn1, at jj, at nn1, at nn2, at1 jj, at1 nn1, ddl nn1, ddl vbz, ii at, ii ddl, ii nn1, io at, io nn1, jj nn1, jj nn2, nn1 ., nn1 cc, nn1 cst, nn1 ii, nn1 io, nn1 nn1, nn1 vbz, nn2 ., nn2 ii, np1 np1, pph1 vbz, ppis1 vv0, ppis2 vv0, ppy vv0, to vvi, vm vvi, vvgk to, vvi ii, vvn ii, xx vvi.

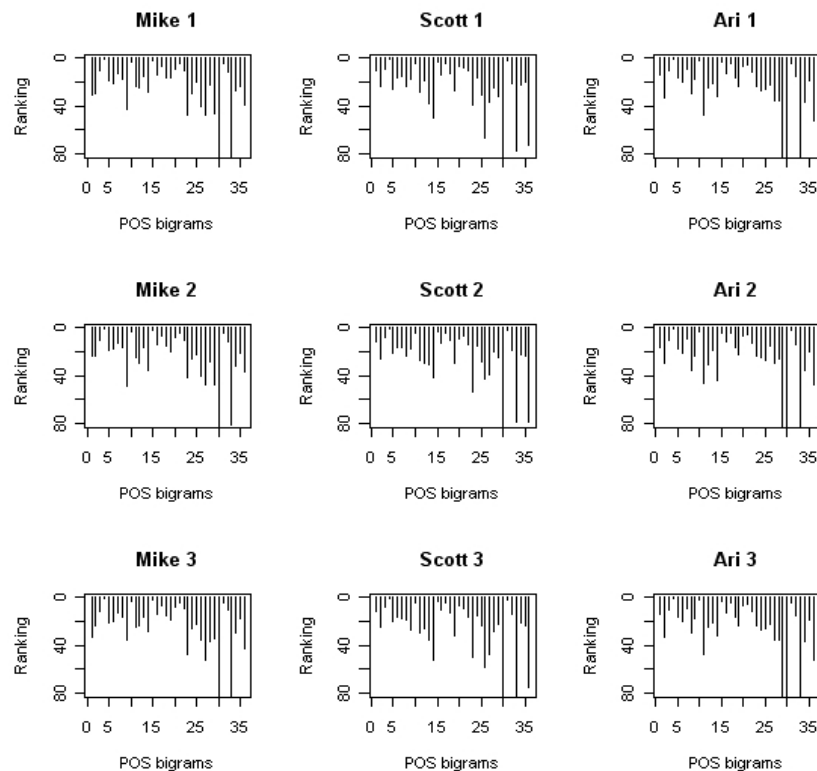


Figure 14: Ranked POS bigram profiles for three samples from three speakers

The ranking of each POS bigram is determined for each speech sample and the results are displayed in graphical form (Figure 14.) The plots in Figure 14 reveal some intra-speaker similarities and some inter-speaker differences, but the distinctions among speakers do not appear to be as clear on the basis of a cursory visual inspection as they do for the word bigrams and trigrams.

Turning to the correspondence analysis of the data, we find, however, that the POS bigrams clearly differentiate the five speakers, as detailed below. Before examining the results, we should note that in order to avoid a partial replication of the word bigram analysis, some common tag sequences were eliminated from the original list because the tag pairs were essentially word pairs. The tag sequence PPH1 VBDZ, for example, is equivalent to the lexical string *it is* and so to determine whether grammatical sequences represented by POS speech tags can distinguish individual speakers, it is necessary to remove those tags that are unambiguously tied to particular word pairs. The purpose of this particular analysis is to determine whether the more abstract level of grammatical description associated with tag sequences is able to distinguish the output of individual speakers.

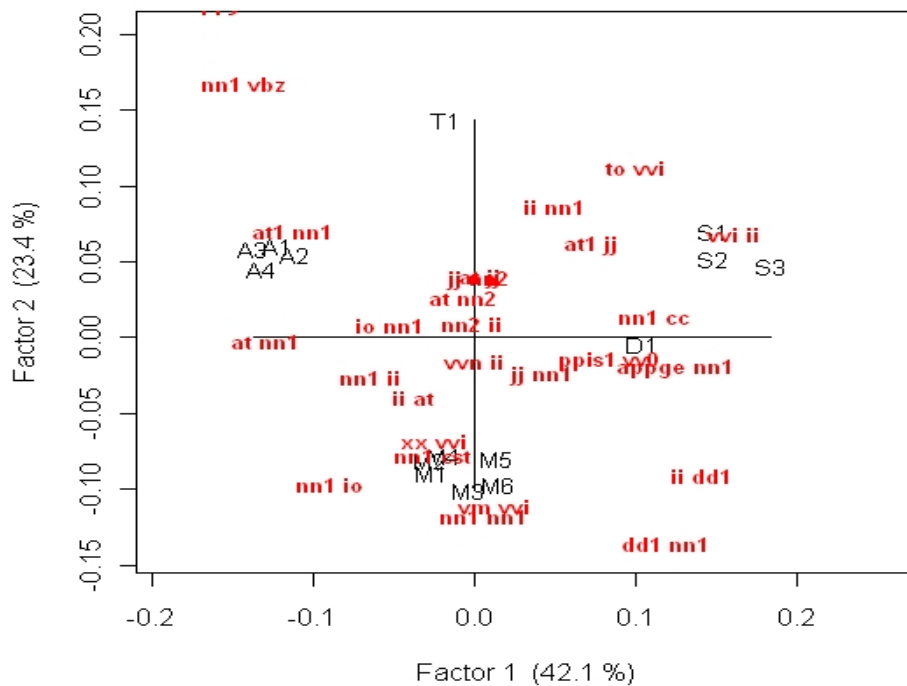


Figure 15: Correspondence analysis between POS bigrams and speech samples

The graphical representation of the results of the correspondence analysis is displayed in Figure 15 and it is evident that tag bigrams distinguish speakers just as well as word bigrams do. On the left side of the x axis we find AT NN1 and AT1 NN1 (indefinite and definite article and singular common noun); on the right side VVI II (verb plus preposition) and APPGE NN1 (possessive pronoun and noun). Tag sequences NN1 NN1 and VM VVI (modal and verb) occur at the lower end of the y axis, near to the samples for Mike. The tag sequences contributing the most to the two factors are AT NN1, PPIS2 VV0, AT1 NN1, NN1 IO, VVI II (Factor 1) and DD1 VBZ, TO VVI (Factor 2). Let us first consider AT NN1 (definite article and singular common noun). It might be thought something as general as AT NN1 would be used to the same extent by all speakers, but with a large enough sample, we find some differences emerging. We already know, however, that these differences exist because we have detailed

above how the frequency of the noun phrase *the president* varies from speaker to speaker and so we must determine the extent to which variation in AT NN1 is due to the use of the phrase *the president*. Frequent use of *the president* does not invalidate AT NN1 as a bigram that distinguishes speakers, but it is useful to know precisely what kind of variation data we are dealing with. It turns out that about a quarter of AT NN1 sequences are, in fact, *the president*. If we remove these uses from the counts for AT NN1, we obtain the results illustrated in Figure 16. The counts are quite high and so the small differences found among the speakers may well reflect a real difference in usage.

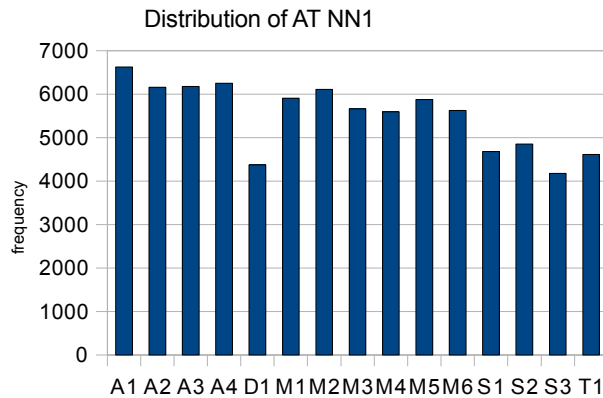


Figure 16: Distribution of AT NN1 (not including *the president*)

We find in Figure 17 a fair degree of variation among speakers for the tag sequence *ppis2 vv0* (*we + verb*). Bearing in mind the case of AT NN1, we need to investigate whether the distribution reflects a preference for the grammatical sequence based perhaps on a preference for some larger constructional unit like AT NN1 – or whether what we see here is essentially a lexical preference. If we go back to the corpus and check the lexical instantiation of the tag sequence we observe that although *we need* is a very common instantiation of the pattern it is not over dominant.

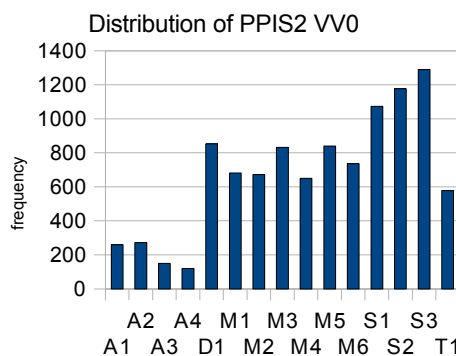


Figure 17: Distribution of PPIS2 VV0

There are some modest distinctions among speakers for the indefinite article plus noun (AT1 NN1) and for noun plus *of* (NN1 IO), and VVI IO (V plus preposition), DD1 VBZ and infinitives (TO VVI). While not dramatic, it is surprising to find any differences in the use of these common tag sequences. The illustrative examples of the distribution of AT1 NN1 and TO VVI are depicted in Figure 18.

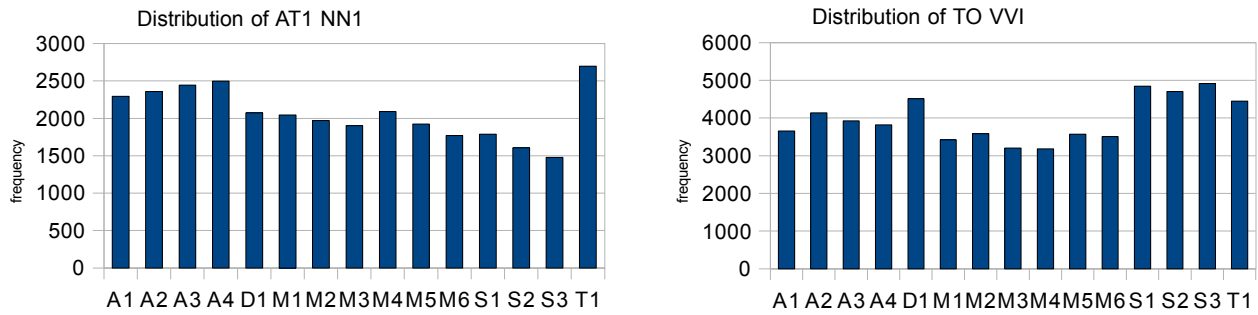


Figure 18: Distribution of AT1 NN1 and TO VVI

The differences among the speakers are not marked, but there is some consistency in the results, which suggests that future fine-grained studies will yield interesting insights into the different strategies used by the different speakers.

It would be interesting to determine the length of tag sequences at which the distinction among speakers disappears. Thus adding a third tag would be expected to increase the level of variability, as will adding a fourth tag and so on. An analysis using POS 5-gram was carried out and the results derived from a correspondence analysis contained the same general configuration as shown in Figure 15. Unfortunately, going beyond 5-gram tag sequences is not feasible with the sample size we are working with because the frequency of the sequences becomes quite low and the more frequent sequences then tend to be equivalent to lexical phrases.

3.5 Grammatical probes

In this final section we investigate idiolects by assessing idiolectal variation in terms of various grammatical properties related to the ngram analyses.

Negation

The distribution of *not* in is shown in Figure 19. Not surprisingly the frequencies are quite high. Tony and Dana use *not* the most, their usage being around twice as frequent as Scott's.

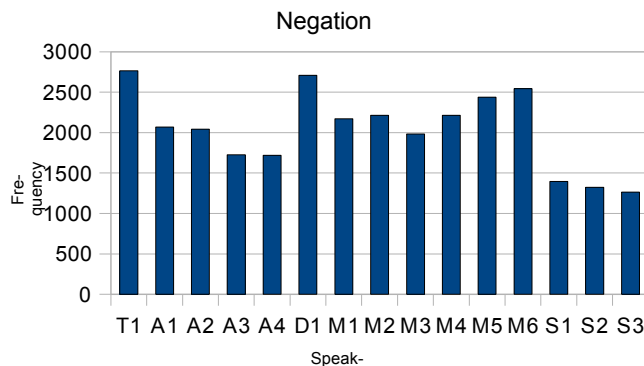


Figure 19: Distribution of *not/n't*

The most common lexical phrases involving *not* are *I don't know*, *I don't think* and *I'm not going*. These make up around 5-10% of the total uses of the negative form. Each speaker tends to use the phrases in roughly the same proportion in the different time samples, as shown in Figure 20.

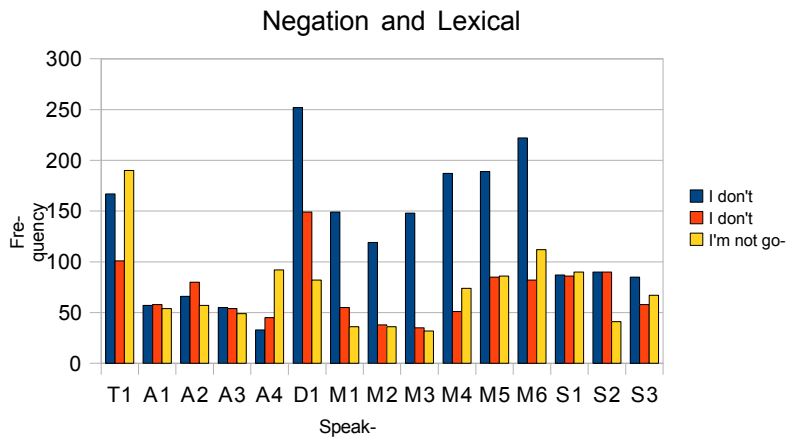


Figure 20: Distribution of common lexical phrases containing *not/n't*

Use of *will* and *be going to*

It is generally accepted that there are some differences in the future meaning expressed by *be going to* and *will* (refs), but there is also an overlap in the meaning of the two forms allowing for individual preferences to come to the fore. The data from the White House press secretaries points to some individual preferences in the choice between the future markers. Typically *going to* is used about 20% of the time and *will*, 80%, by the different speakers, with a slightly higher use of *going to* by Scott. Dana uses *going to* 30% of the time and Tony distinguishes himself by using *going to* more than *will*.

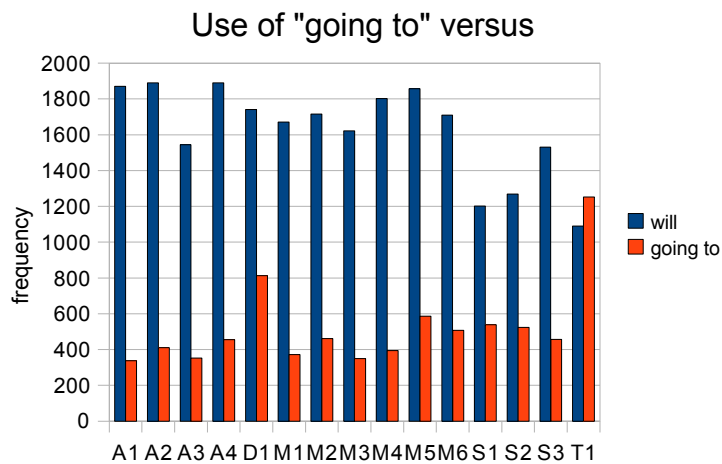


Figure 27: Distribution of *going to* and *will*

Construction: [X of the Y]

We can examine a couple of examples of the results of sequences of five POS tags and identify those tag sequences that contribute strongly to the variance in the correspondence analysis. Among these we have AT NN1 IO AT NN1, which is essentially an *the X of the Y* construction particularly favoured by Mike and Ari. If we examine the indefinite form INDEF X of the Y, we obtain a similar picture except for the use of the form by Mike, which is proportionally much less.

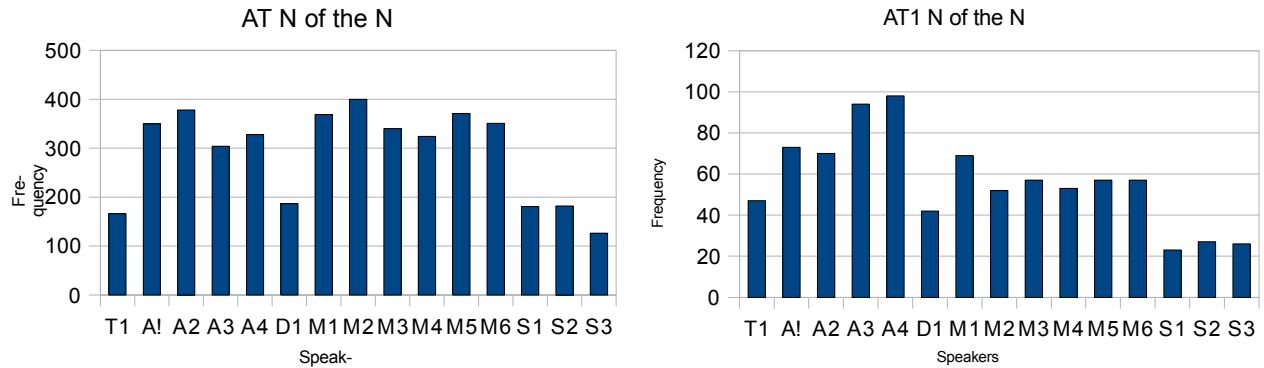


Figure 19: Distribution of *the X of the Y* and *INDEF X of the Y*

Use of passive constructions

The result of a search for the use of the passive is telling. First of all, the patterning of data for this construction is similar to what we have seen in the bigram and trigram graphs in that we find that the intra-speaker variation is less than the inter-speaker variation. If we focus first on the overall frequencies of passive use (Figure 20), we discover that Ari uses the construction about twice as often as Tony. In contrast, if we look at only those passives with a following *by* phrase, which occur around 10% of the time, we notice a shift in relative proportions of use, with Mike favouring *by*-phrases and Dana tending to avoid their use. We find similar changes in the relative proportions when comparing the choice of particular verbs (*make*, *do* and *commit*) in the passive construction, as shown in Figure 21. Thus we see that the frequency of use of the passive by Mike and Scott is rather similar (Figure 20), but if we focus on particular verbs used in the passive construction (Figure 21), significant differences emerge between the two speakers.

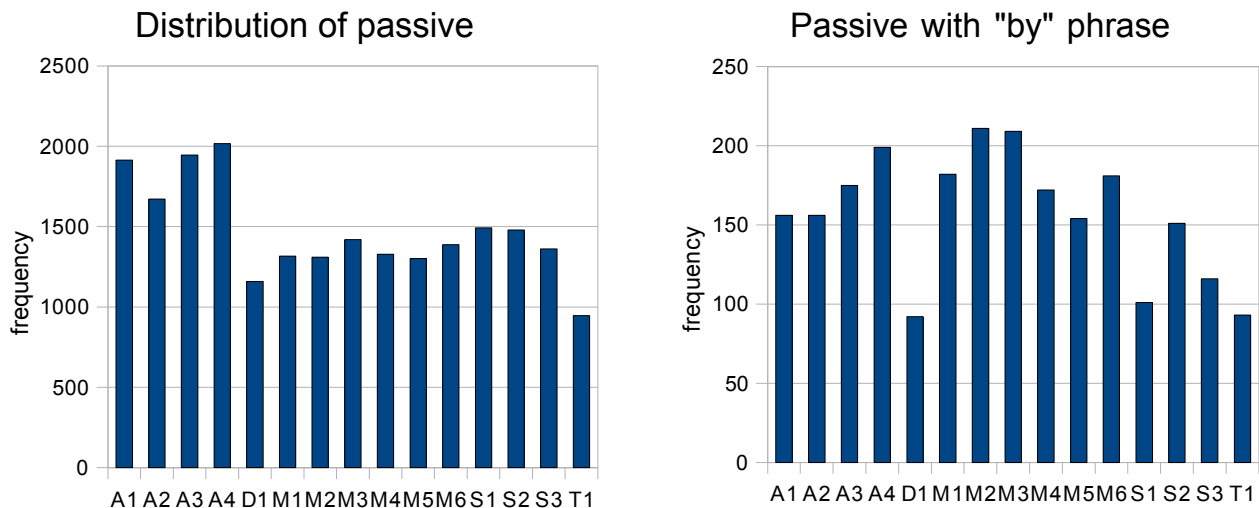


Figure 20: Distribution of use of passive (in total) and passive plus *by* phrase

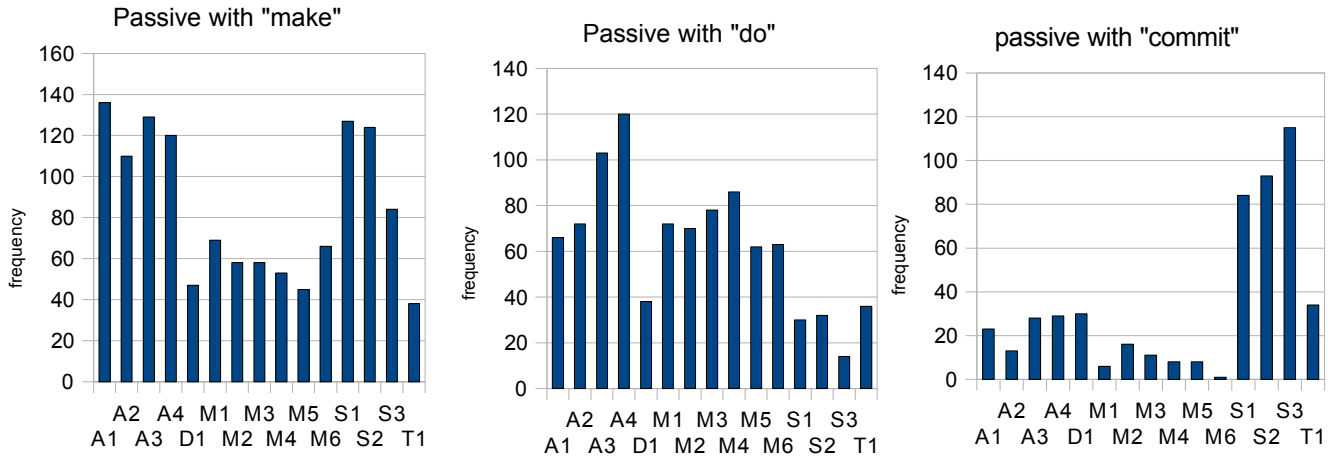


Figure 21: Frequencies of passive with *make*, *do* and *commit*

It + ADJ constructions

Here we examine two similar constructions: *it + ADJ + that* clause and *it + ADJ + to infinitive*, as shown in Figure 21.

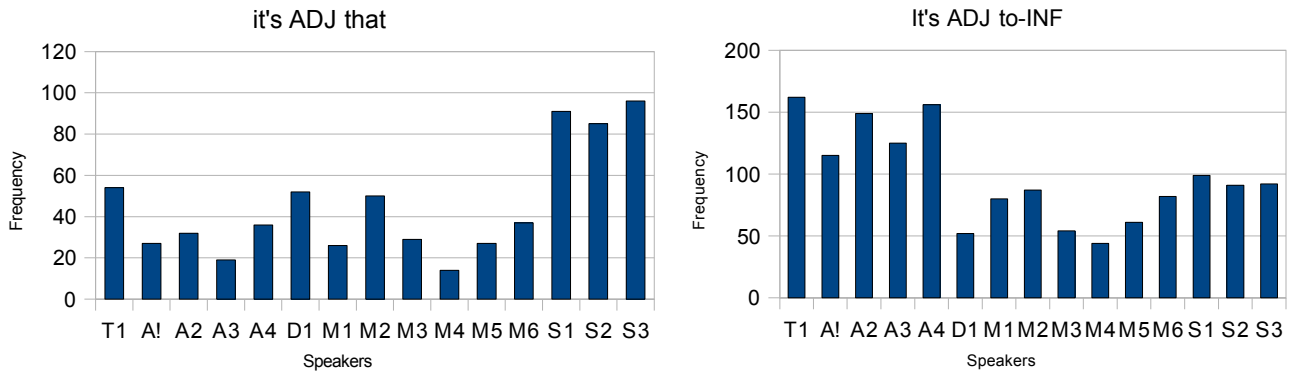


Figure 21: Frequencies of *it's ADJ that* and *it's ADJ to-INF*

The frequency of the two structures is of the same order with the *to-INF* version being somewhat more common. Mike McCurry does not favour the use of either form. Ari has a marked preference for the *to-INF*. Scott is the most frequent user of the *ADJ that* form and he uses both structures equally frequently. The graphs shown in Figure 21 along with other data presented here point to marked preferences for particular constructions. One data set by itself is not sufficient to distinguish speakers, but a consideration of just a selection of probes strongly suggests the reality of different grammars as defined in a usage-based model of language.

4. Conclusion

The patterns investigated in this paper show the speech of different individuals in their role as White House press secretaries over a period of a year or two. Given this restricted context, the data can only reflect the idiolectal variation associated with these speakers within this particular setting and cannot be said to constitute an account of individual grammars. Despite these restrictions the various probes

described here establish quite clearly that there are dramatic differences in the speech of individual speakers across a wide range of lexicogrammatical patterns. Not every probe produced results that definitively distinguished each speaker, but the analyses nevertheless yield some interesting and perhaps surprisingly strong patterns. The most striking results are (i) the stability in an individual's productions and preferences over time and (ii) the extent of the differences among speakers. It is clear that a combination of a few probes are sufficient to establish the individual preferences of each speaker within this particular context. The results also show that idiolectal variation is based on core aspects of language and not on peripheral idiosyncrasies. Even the most frequent bigram "of the" is an indicator of different preferences in the use of constructions by different speakers.

The extent and breadth of the idiolectal variation described above is quite surprising and it is not readily apparent what processes lead to the maintenance of an individual grammar underlying spoken production that is so distinct from the ambient language, i.e., the patterns of language that speakers are continually exposed to as they process the speech of their interlocutors. One way of stating this distinction is to contrast comprehension and production and note that the type and token frequencies associated with language comprehension are far different from those associated with language production. The theoretical consequences of this disparity remain to be explored.

How do we explain the idiolectal variation identified in this paper? Whatever the particular explanation turns out to be, the most promising framework is a usage-based model of language in which there is a strong link between language in use and the cognitive representation of language (Langacker, 1987) and within a usage-based theory an exemplar-based approach (Pierrehumbert 2001, Bybee 2006, Hay and Bresnan 2006) might offer a way of explaining the data. See Barlow (2010).

In conclusion, we can return to the issues raised in the introduction concerning the status of idiolects and their relation to sociolects. The arguments for a social perspective on language based on the notions of conventionality and the use of language for communication are not at odds with the data on idiolects presented here since the speakers clearly share the same grammar to the extent that they process and in many cases produce the same constructions. The differences arise with respect to style and preferred patterns of use in production. In this study there is not enough data to properly consider the issue of regularity and stability of individual grammars, although there are signs of a surprising consistency over a period of a year or so. Similarly, we cannot on the basis of these results make any suggestions as to whether the idiolects or sociolects are the primary unit of language, but the variation in individual grammars is quite extensive and it seems unlikely that the view of an idiolect as an individual's share of the sociolect can be upheld, taking into account the lexical and syntactic usage found in the analysis presented here. Labov's (2010) suggestion that "reports of idiolectal grammars have been found to be accidents of introspection rather than differences in the production and interpretation of language in its community setting." can only be upheld if differences in frequency of usage are taken to be inconsequential. However, the remarkable consistency in the patterns described here supports the view that frequency of use of constructions is an integral part of idiolectal grammars.

References

- Barlow, Michael. 2000. Usage, Blends, and Grammar. In Michael Barlow and Suzanne Kemmer (eds), *Usage-Based Models of Language*. 315-344. Stanford: CSLI
- Barlow, Michael. 2010. MS. How to distinguish individual speakers: a corpus-based investigation of idiolects.
- Barlow, Michael & Suzanne Kemmer. (eds), 2000 *Usage-Based Models of Language*. Stanford: CSLI
- Barlow, Michael & Suzanne Kemmer. 2004. Input grammar and output grammar. CSDL 7: Experimental and Empirical Methods. University of Alberta, Edmonton, Canada

- Bartes, Roland. 1977. *Elements of Semiology*. Hill and Wang: New York.
- Baayen, Harald. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge: Cambridge University Press.
- Bloch, B. 1948. A set of postulates for phonetic analysis. *Language* 24, 3-46.
- Bresnan, Joan and Jennifer Hay. 2008. Gradient grammar: An effect of animacy on the syntax of *give* in New Zealand and American English.' *Lingua* 118, 2, 245-259.
- Brezina, Vaclav. 2010. "Definitely know what I'd do": In search of epistemic sociolect/idiolect. LAUD Conference. Landau. Germany.
- Bybee, Joan. 2006. From usage to grammar: The mind's response to repetition. *Language* 82, 4,
- Bybee, Joan. & Joanne Scheibman. 1999. The effect of usage on degrees of constituency: the reduction of don't in English. *Linguistics* 37-4.575-596.
- Dittmar, Norbert. 1996. Explorations in 'Idiolects'. In Robin Sackmann and Monika Budde (eds). *Theoretical Linguistics and Grammatical Description: Papers in honour of Hans-Heinrich Lieb*. Amsterdam: Benjamins.
- Gries, Stefan T. 2005. Syntactic priming: A corpus-based approach. *Journal of Psycholinguistic Research* 34.365-399.
- Hay, Jennifer and Joan Bresnan. 2006. 'Spoken syntax: The phonetics of giving a hand in New Zealand English. *The Linguistic Review* 23: *Special Issue on Exemplar-Based Models in Linguistics*, 321-349.
- Jakobson, Roman. 1971. *Studies on Child language and Aphasia*. The Hague: Mouton.
- Kemmer, Suzanne and Michael Barlow. 2000. Introduction: A Usage-Based Conception of Language. In Barlow, Michael and Kemmer, Suzanne (eds), *Usage-Based Models of Language*. 7-28.
- Labov, William. 1989. The exact description of the speech community: Short 'a' in Philadelphia". In R. Fasold and D. Schiffrin, (eds.) *Language Change and Variation*, Washington D.C.: Georgetown University Press, 1-57.
- Labov, William. 2010. The community as the focus of social cognition. LAUD Conference. Landau. Germany.
- Langacker, Ronald. 1988. A usage-based model. In B. Rudzka-Ostyn, (ed), *Topics in Cognitive Linguistics*, 127-61. Amsterdam: Benjamins.
- Langacker, Ronald. 2000. A dynamic usage-based model. In Barlow, M and Kemmer, S. (eds), *Usage-Based Models of Language*. 1-63.
- Meyerhoff, Miriam and James A. Walker. 2007. 'The persistence of variation in individual grammars: Copula absence in 'urban sojourners' and their stay-at-home peers, Bequia (St Vincent and the Grenadines).' *Journal of Sociolinguistics* 11: 346-66.
- Pierrehumbert, Janet B. 2001. Exemplar dynamics: Word frequency, lenition and contrast, in Bybee, J. and Hopper, P. (eds.), *Frequency and the Emergence of Linguistic Structure*, John Benjamins, Amsterdam, 137-157
- Rayson, Paul. 2008. Wmatrix: a web-based corpus processing environment, Computing Department, Lancaster University. <http://ucrel.lancs.ac.uk/wmatrix/>
- Smolkin, Rachel. 2000. 'What Did He Say?' *American Journalism Review*, September. (Retrieved January 9, 2009 from http://www.ajr.org/article_printable.asp?id=2620)
- Tomasello, Michael. 2003. *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press.
- Tomasello, Michael and Daniel Stahl 2004. Sampling children's spontaneous speech: How much is enough? *Journal of Child Language*, 31, 101-121
- Weiner, E. Judith & William Labov. 1983. Constraints on the agentless passive. *Journal of Linguistics*, 19:1 29-58
- Wunderlich, Dieter. 1996. *Foundations of Language*. Cambridge: Cambridge University Press.