A Pilot Study on Gender Differences in Conversational Speech on Lexical Richness Measures

Short title: Gender-based Lexical Differences

Sameer Singh

Department of Computer Science

University of Exeter

Exeter EX4 4PT, UK

tel: +44-1392-264 053

fax:+44-1392-264 067

email: s.singh@ex.ac.uk

Abstract

Recent research on gender differences in language has mostly addressed cognitive differences. These differences have been observed on different cognitive verbal and non-verbal tasks and conclusions on the variability in language production and comprehension have been drawn from their results. In this paper, a different approach is presented. This pilot study examines lexical richness measures in conversational speech across a total of thirty subjects. All subjects were recorded and transcribed in a conversational setting. Their transcribed speech was analyzed using a set of lexical richness measures based on word-frequencies. On the basis of these measurements, statistical discriminant analysis is able to classify the two groups with 90% (74% with leave-one-out cross-validation) correct prediction rate at a statistically significant level ($p = .03$). The results are discussed in detail including correlation and principal components analysis. The paper concludes that there are interesting differences across the two groups on the measures studied and further research in this area is needed.

1. Gender differences

Research on gender differences related to language has mainly focused on cognitive functions. For example, recent papers have studied a variety of differences as for young school children (Robinson et al. 1996; Beller and Gafni, 1996), college going students (Stumpf and Stanley, 1996), adults (Halpern and Wright, 1996; Feingold, 1993, 1996; Mckeever, 1996; Hedges and Nowell, 1995; and Voyer et al., 1995) and subjects with affective or communicative disorders, e.g. Schizophrenia (Addington et al., 1996), Major Depressive Syndrome (Fava et al. 1996) and Aphasia (Hier et al., 1994). A valuable commentary on the state of current research on gender differences is available in Hyde, (1996). A number of other studies have also been published on the subject prior to 1994. Unfortunately, most of the above mentioned and several other recent studies are not directly related to language production as in conversation, and only a few studies have partially addressed gender differences in spoken and written language output. Examples include Mulac and Lundell (1994) - written discourse; Ferber (1995) - transcribed speech patterns; Freed and Greenwood (1996) - conversational analysis; and Feldstein, Dohm and Crown (1993) - speech rate analysis.

The study of gender differences in conversational speech is important for several reasons. One reason is its relevance to psycholinguistics. It has been previously established that for some language tasks, different parts of male and female brains are activated for the same task, Ojemann (1983). However, all language studies at present work with a single language production model which is being increasingly refined with modern research. If we were to establish key gender differences in language production, then such models must be studied in light of these new findings in order to be valid. Another reason why a conversational study is important, is its usefulness to linguists and psychologists in general. For example, the study of socio-linguistics and language acquisition models is assisted through the understanding of how male and female speakers use language. Also, such work has important consequences on how language disordered subjects are rehabilitated. Holme and Singh (1996) have used the same measures as in this study for quantifying lexical deficits in aphasia. One could argue that if there are gender differences in how patients are affected by stroke, as suggested by Hier et al. (1994), and if there are also gender differences in language production, whose improvement is the most important goal of rehabilitating patients, then probably therapy should be gender based.

Currently, there is only a weak support in favour of gender differences in language production. Ferber (1995) studied how easy it is to identify the gender of the speaker by reading transcribed speech. It was found that on most occasions it can be guessed correctly, but such guesses are heavily dependent on the context, for example men are generally supposed to be talking more about technical issues than women. Freed and Greenwood (1996) have analyzed conversation between four female and four male pairs, examining the use of phrases such as ‘you know’ and questions within three types of discourse. This study failed to find any conclusive differences and emphasizes the role of context in variables studied. Feldstein, Dohm and Crown (1993) however showed positive differences across the two groups using speech rate but this study was related to speech listening and shows that the rate of speech perception is different. It seems that studies in the area of gender differences as evaluated on natural language production are limited in number.

Analyzing conversation for characterising lexical richness in speech can be accomplished in a number of ways. In this paper the word-frequency measurement approach is adopted which has been previously used for analyzing aphasic conversation

(Singh, 1994; Holmes and Singh, 1996). This approach is based on the statistical finding that different speech samples (including aphasic speech) follow a set of statistical laws, e.g. Zipf's law (1932), which makes them comparable on the basis of word-frequencies and lexical richness measures, (Howes, 1964). This approach is useful since most subjects can be easily compared on the basis of their conversational performance. The effect of contextual factors on word-frequencies is better controlled in this context than it would be possible in a pragmatic analysis based on turn-taking data, since subjects are not interrupted during their speech and the output is objectively analyzed rather than a subjective analysis of subject behaviour during conversation. In order to further reduce any subjective influences, all transcripts in this study have been analyzed by the author himself, and were verified later by two independent judges. In this study, first details of the sample and method are presented, followed by a description of the linguistic and statistical analysis and final results. The final discussion will summarize findings in this paper and argue the utility of analyzing conversational language for identifying gender differences. Of course, similar work can be done in other areas where different groups of subjects are supposed to have different language organization in the brain, such as monolinguals and bilinguals, left- and right-handed subjects, etc.

2. Sample

A total of thirty normal subjects (13 male and 17 female) without any previous history of language problems were included in the analysis. We characterise those subjects as normal who do not suffer from language impairments resulting from a disease or stroke. The overall analysis takes roughly 150-180 minutes per transcript, and therefore doing this analysis on a larger scale was not immediately possible. All subjects were roughly matched for age (all above 50 years of age), and previous results on their word-frequency measurements show that their performance on the measures chosen was not affected by age (a Pearson's correlation coefficient of +.12 between subjects' age and their discriminant score, see Holmes and Singh (1996)). It was also decided to examine the effect of educational background on subjects' conversational speech. Using variables studied in this paper, subjects were analyzed using two groups: type-I subject who had a university degree or professional qualification, and type-II subjects who left school early and were now working as artisans. Discriminant analysis showed minor differences between type-I and II subjects on lexical richnesss measures which were not statistically significant at the 5% significance level (p = .072) (Holmes and Singh, 1996). Hence, it was decided to combine the two groups as one, and study gender differences on this 'normal' group.

All subjects were recorded on a one-to-one basis. The recording was done in a quiet room with a clip-microphone. Subjects were asked about their hobbies, life experiences, current activities, and any other topic they were interested in. This recording of free and spontaneous conversation was then used for transcription and at least 1000 words were extracted for further analysis from the answers given. This word-limit was consistently observed across all transcripts and all transcripts varied between 1000-1200 words. It was considered important to match the text-length since variable measurements may vary with text-length. Since no pragmatic turn-taking analysis was performed, only the subject interviewed was transcribed in a simple form. On the transcript, different turns were however transcribed on different lines in order to keep it structured and to help correct mistakes whilst listening to the original recordings. Words such as "you know", "to be honest" and other stereotyped expressions were eliminated from the analysis. Also, 'yes' and 'no' responses were eliminated from the analysis so that lexical richness measures were not affected by these elements of empty speech.

3. Method

All transcripts were initially analyzed using the Oxford Concordance Program (OCP) (Hockey and Martin, 1988). The program lists word-frequencies for different words used in the text and also provides other useful statistics including the type-token ratio, text-length and total vocabulary. The OCP output and the original transcript were used for computing lexical richness measures which will be described next.

3.1. Lexical richness measures

A total of eight measures, which have been used previously (Singh, 1994) were selected for this study. These are described below:

1. Noun Rate per 100 Words
2. Pronoun Rate per 100 Words
3. Adjective Rate per 100 Words
4. Verb Rate per 100 Words
5. Type-Token Ratio
6. Clause-Like Semantic Units (CSU) per 100 Words. A CSU is the minimum number of words in a grammatically cohesive string with semantic meaning.
7. W - Brunet's Index (Brunet, 1978), given by $N^{V^{-.165}}$. A lower value represents a lexically rich language. In most cases, values lie in between 10 and 20.
8. R - Honoré Statistic (Honoré, 1979) given by 100*log (N)/ (1 - V_1/ V). A higher value indicates a lexically rich language. In most cases, values lie in between 1000 and 2000.

Here, N is the text-length, V is the total vocabulary and V_1 is the number of words used only once. In our study, none of the measures were found to correlate with text length.

The first four measures are the based on lexical items: open class (nouns, adjectives and verbs) and closed class (pronouns). Type-token ratio is the ratio of the total vocabulary used to the overall text -length and is a measure of lexical richness. Samples with many different words used with small text-lengths will have a high value which represents a lexically richer speech (typical values range from .25 to .35 in a text-length of 1000 words). Clause-like semantic units is a measure of how well subjects group words into phrase like structures, and a set of rules have been devised for text segmentation in Singh and Bookless (1997). Most CSUs have boundaries at conjunctions and at best a CSU can be a complete sentence. For example, “I went to the market | and I bought magazines”, has two CSUs of length 5 and 4. Singh and Bookless (1997) has also compared CSU to other conventionally used procedures such as Morphological Lexical Unit (MLU) and argued that CSU measure is more reliable and easier to compute. CSU rate per 100 words was computed. A low rate (implying longer CSUs) denotes better sentence making ability which may be, in turn, related to the ease with which subjects form noun and verb phrases. The last two measures, W and R, have been widely used in authorship-attribution studies and a detailed treatment can be found in Holmes (1992, 1994). These measures although working with text-length, do not correlate significantly with it. W is a measure of lexical richness which relates how varied the vocabulary is for a given piece of text and the final measure R has been explained by Holmes (1992) as: “ It directly tests the propensity of an author {in our case a speaker} to choose between the alternatives of employing a word used previously or employing a new word ... When comparing texts, therefore, the higher the R-value is, the richer the vocabulary in the sense that a greater number of words appear infrequently”, (p. 93).

The above measures were computed using the OCP output. All grammatical items were individually marked on the transcript manually and later counted. CSUs were computed using the transcript itself and a plot showing CSU frequencies and the number of words in them was also plotted. However, this plot is only useful for showing individual differences and was not used for this study. All measurements were recorded for further statistical analysis.

3.2. Descriptive analysis

Table 1 shows descriptive statistics for male and female subjects. At first glance, there seems to be considerable variability across individual performances. On the noun-rate measure, male subjects achieve a higher average value, however, the opposite is true for pronoun rate. This is supported by the correlation analysis results shown later which confirm that the noun and pronoun rates are inversely related. The variability amongst female speakers on noun rate is higher compared to males, however, both groups are nearly similar in variability on the pronoun rate. Male subjects have a higher average adjective usage but there is also higher variability in performances, for example adjective rate of 11.12 (maximum) seems very much like an outlier. On the verb rate measure, female subjects use them in higher numbers as minimum, maximum and mean for this group are higher than the male group. There are minor differences on the TTR measure, which supports the argument that vocabulary of a speaker is not a function of his/her gender. The ability to form long phrases is certainly different across the two groups as indicated by the CSU statistics. Males score higher on this measure as indicated by lower counts, however their performance is more varied than the female group. Again, the W measure is less affected by gender but on the other hand, R is significantly variable across the two groups. Male subjects have higher values, which supports the argument that they are less repetitive in their selection of words, i.e. most of the words used by male subjects had been used only once. On the other hand, female subjects use less number of words only once and tend to repeat them and their performance seems less varied compared to the male group.

Table 1.

Descriptive analysis presented here is useful for summarizing data in a meaningful manner. It allows us to compare variable differences across the groups studies and comment on unusual and important observations. Statistical analysis results are presented next which offer a broader view to understand data obtained.

4. Statistical data analysis

The data obtained from the language analysis of transcripts was further subjected to other statistical tests. The results will be presented here for correlation analysis, discriminant analysis and principal components analysis. The purpose of the first analysis is to identify significant relationships between different variables and observe key gender differences. Discriminant analysis will be used thereafter to classify male and female performance patterns. This analysis identifies the relative importance of variables in such a classification process and provides a final discriminant score which can be used to rank individual performances. A stepwise analysis will also be used to give us a simpler model of discriminant function. Finally, principal components analysis will be performed to identify a smaller set of uncorrelated variables which can be used to better explain data used here.

4.1. Correlation analysis

A number of interesting observations can be made from Table 2. Pearson's correlation coefficient differs across male and female subjects across almost all variables. In some cases, although the relationship is similar in direction (positive or negative), their magnitudes are considerably different, e.g. N-rate & P-rate, A-rate & P-rate, A-rate & V-rate, P-rate & TTR, A-rate and W, and R & CSU (all negative), A-rate & TTR, P-rate & W (all positive). On other occasions, the coefficients are either similar, or in some cases so different that they have different signs, e.g. N-rate & A-rate (males positive, females negative), N-rate & CSU-rate (Males negative, females positive) and W & CSU (males positive, females negative). In Table 2, all strong relationships (coefficients > 0.7) have been highlighted. Some of these are worthy of close observation.

N-rate and P-rate are significantly related in the male population. This relationship exists in the same direction, i.e. inverse, but is not considerably strong in the female group. It seems that female subjects use nouns and pronouns interchangeably, a phenomenon not so much observed in the male group. Pronouns are also inversely related, to a significant extent, to adjectives in the male group. Once again, the relationship exists inversely in the female group too, but it is less intensive. Pronouns and Verbs are positively related in each group, which implies that when subjects use more pronouns, they need more verbs to form phrases. Surprisingly, nouns are inversely related with verbs, which implies that the usage of nouns inhibits the use of verbs in sentences, an unusual finding. Verbs and adjectives are strongly related in the male group, negatively. Lexical richness measures W and R are inversely correlated, which is expected since a reduction in W and an increase in R both imply lexically richer speech. Both these measures are also strongly related to the verb rate, R positive and W negatively. This implies that an increase in verb rate, contributes to an increase in the number of words used only once and thereby increases R. Other strong relationships include TTR & P for males (ρ = -.72) and W & P for males (ρ = +.70).

Table 2.

4.2. Discriminant analysis

Discriminant analysis was performed in order to classify performance patterns based on pre-assigned groups (male and female). The data was also used differentially, for generating a discriminant function and for testing it. Cross-validation was performed using the jackknife method, also popularly known as the leave-one-out method. Using this approach, for a data set of size *n*, a discriminant function is generated using *n*-1 patterns and it is tested on the remaining pattern. This method iterates a total of *n* times with different sets used for function generation and function testing.

A classification rate of 90% was obtained initially, and 27 out of 30 cases were correctly predicted to belong to their respective groups. With cross-validation, a classification rate of 74% was obtained where 22 out of 30 cases were correctly predicted. The results achieved were significant at the 5% significance level (χ^2 = 16.89, p = 0.03, d.f. 8). Inspection of misclassified cases revealed close competition amongst probability assignments. Figure 1 below shows the discriminant plot.

Figure 1

In the above figure, four 1's in the plot represent a female subject and four 2's represent a male subject. The subjects have been plotted on the X-axis which represents the final discriminant score on discriminant function 1, the axis of maximum variability. These scores can also be used to judge relative individual performances within the two groups. The plot clearly shows that male and female subjects separate out on the set of eight variables proposed. Also, it shows more variability in male performances than female performances as 1's cluster more closely compared to 2's. A classification summary is presented in Table 3.

Table 3.

In addition to a classification summary, discriminant analysis also provides us with details of the role played by individual variables in the analysis. This information is available as the degree of correlation between the generated discriminant function scores and individual variable scores. The results are shown in Table 4.

Table 4.

These results show that R and CSU-rate are the most significant discriminators across male and female groups. Variables such as TTR and A-rate are considerably less varied across the two groups. A stepwise analysis was further performed which also considers variable R as most important but then includes variable W in the model, and on the basis of these two variables alone, classifies nearly 83% of the cases correctly.

4.3 Principal Components Analysis

Principal components analysis was performed to compute a smaller set of independent variables which could explain the variance in data (Jollife, 1986). The results of this analysis are presented in Table 5.

Table 5

A total of three principal components were extracted with eigenvalues greater than 1. The first two principal components explain nearly 72% variance in data, and the first three taken together explain nearly 85% variance. The first principal component therefore explains greater variance than successive principal components. It mainly contrasts lexical richness measures (TTR, W and R) with sentence making ability measures (CSU-rate, P-rate and V-rate). A higher score is achieved on this component when TTR, W and R scores are high and CSU-rate, P-rate and V-rate scores are low. The second principal component is dominated by the CSU-rate and N-rate, and seems to represent the sentence making ability as affected by the N-rate. High scores on this component are achieved for high scores on N-rate, P-rate, V-rate, TTR. CSU-rate and R, and low scores on A-rate and W. A plot of PC1 and PC2 is shown in Figure 2.

Figure 2

Figure 2 shows that male subjects (B) are more varied on the PC1 axis whereas female subjects (A) are more varied on the PC2 axis. This means that male performances vary considerably on lexical richness and sentence making ability factors, and female performances vary mostly on their ability to form phrases based on nouns.

5. Conclusion

From the above discussed analyses, a number of gender differences in conversational speech have emerged. Some of these differences are minor, whereas others are important and should prompt further research. On the whole, it appears that language is used in different forms in conversation by male and female groups. Male speech is lexically richer and phrases used tend to be longer. Female speech on the other hand contains shorter sentence structures, is more repetitive in its use of lexical items, uses nouns and pronouns interchangeably, and is dynamic using more verbs. However, we must not forget that the above comment should be taken with care, especially when Table 1 shows wide ranging performances in both groups and that we have used a limited number of subjects for the analysis.

Differences on individual variables need closer observation. It is here that further effort should be made to identify more subtle differences. Although all variable measurements differ across the two groups, some differences are more important than others. This information is not readily available by observing standard deviation figures alone since different variables have been measured on different scales. Hence, Table 4 shows us a summary of the relative importance of variables in discriminating across male and female speech patterns. R and CSU rate are most varied across these two groups, whereas differences in TTR and A-rate are minor. This information is important for refining sentence production models. It could be that different lexical items are accessed in different ways from different lexical stores in men and women, and that their retrieval procedures can be identified as separate using conversational data. Another source of valuable information for this is Table 2 which lists correlation coefficients across different variables. Male subjects on average have higher correlations, both positive and negative, across most variables compared to female subjects. This evidently implies that language is being processed differently across the two groups, however, the data presented in this study needs to be taken into account with other gender based language studies to identify models which incorporate key differences.

It appears that lexical richness analysis is also helpful in classifying different forms of speech. A closer inspection of misclassified subjects may reveal exciting possibilities. In this study, this is not possible since biographical and personal information about subjects tested is limited. Misclassified subjects (males predicted as females and vice-versa) could actually be different than other members of the group, and these differences will be useful to study. PCA analysis in this study mutually supports the discriminant analysis classification results. It confirms that the two groups differ on lexical richness and sentence making factors, which separates them in two different groups (as shown in Figure 1).

The aim of the current study was to identify possible gender differences in conversational speech on a set of word-frequency measures. The study has yielded interesting results which should encourage further work in this area. We need to identify whether differences in the use of various grammatical items are also manifested in subjects' behaviour. For this, the same set of tape-recordings can now be analyzed for turn-taking data, and also for speech-rate variables. Then we can further discuss the relationship between language behaviour and speech-rate, or turn-taking behaviour, for instance, is it that low CSU rates (longer phrases) in male subjects implies less number of turns and higher speech-rates with less pauses ?

The question of generalisation of these findings is an important one. How well can we generalise the results. In our work we have tried to keep the subjects matched for age and interviewing style, etc. It is not unreasonable to question whether the variability in questioning style, age of the subject, length of sample, etc. may affect the quality of measures extracted. We can briefly comment on these issues. The questioning style has been kept fairly straightforward in our study and previous experience with recording the same subject over a period of time shows that the measures used in our study do not vary significantly if the content of the discourse is different. Second, the topic of investigating whether younger subjects have different language structure and lexical richness than older subjects is a topic of research in itself. By keeping the age range fixed, we assume that we are comparing like with like. It has been shown that as people grow old, they complain of word finding problems and make simpler syntactical structures in their language production even though the lexical or syntactic knowledge itself is not affected (its accessibility may be affected; Singh (2000)). We have kept the length of discourse to more than 1000 words and the appropriateness of this number has been discussed already in our previous publications. Finally, we have only used 30 subjects. So any results must be used with caution and we recommend a larger scale study on similar principles now that some areas of investigation have been highlighted by this pilot study.

Summary

In this paper, thirty subjects (17 females and 13 males) were tape-recorded in a conversational setting and were analyzed for word-frequency dependent lexical richness measures. The sample size was limited since it takes roughly between 150-180 minutes to analyze one subject. This data was analyzed using multivariate statistics. Results show that there are key differences on several variables studied, and that the two groups can be easily classified at a statistically significant level. Male and female subjects use lexical items differently in conversation, and the relationship between lexical richness and sentence making abilities in conversation are different across the two groups. The paper concludes that there is need for further research in this area.

References

Addington, D., Addington, J. and Patten, S. (1996) Gender and affect in schizophrenia. *Canadian Journal of Psychiatry - Revue Canadienne de Psychiatrie*, 41 (5): 265-268.

Beller, M. and Gafni, N. (1996) The 1991 international assessment of educational progress in mathematics and sciences - the gender differences perspective. *Journal of educational psychology*, 88 (2): 365-377.

Brunet, E. (1978). Le Vocabulaire de Jean Giraudoux. *Structure et Evolution*. Slatkine, Genéve.

Fava, M., Abraham, M., Alpert, J., Nierenberg, A. A., Pava, J. A. and Rosenbaum, J. F. (1996) Gender differences in axis I comorbidity among depressed outpatients. *Journal of Affective Disorders*, 38 (2-3): 129-133.

Feingold, A. (1993) Cognitive gender differences-a developmental perspective. *Sex Roles*, 29(1-2) 91-112.

Feingold, A. (1996) Cognitive gender differences - where are they and why are they there, *Learning & Individual Differences*, 8 (1): 25-32.

Feldstein, S., Dohm, F.A., and Crown, C. L. (1993) Gender as a mediator in the perception of speech rate. *Bulletin of the Psychonomic Society*, 31 (6): 521-524.

Ferber, R. (1995) Is speakers gender discernible in transcribed speech ? *Sex Roles*, 32 (3-4): 209-223.

Freed, A. F. and Greenwood, A. (1996) "Women, Men and type of talk - what makes the difference", *Language in Society*, 25 (1): 1-26.

Halpern, D. F. and Wright, T. M. (1996) A process-oriented model of cognitive sex differences. *Learning and Individual Differences*, 8 (1): 3-24.

Hedges, L. V. and Nowell, A. (1995) Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science*, 269 (5220): 41-45.

Hier, D. B., Yoon, W. B., Mohr, J. P., Price, T. R. And Wolf, P. A. (1994) Gender and aphasia in the stroke bank. *Brain and Language*, 47 (1): 155-167, 1994.

Hockey, S. and Martin, J. (1988) *OCP - Users' Manual*. Oxford University Computing Service.

Holmes, D. I. (1992). A Stylometric Analysis of Mormon Scripture and Related Texts. *Journal of the Royal Statistical Society (A)*, vol. 155, Part 1, 91-120.

Holmes, D. I. (1994). Authorship Attribution. *Computers and the Humanities*, vol. 28, No. 2, 87-106.

Holmes, D. I. and Singh, S. (1996) A Stylometric Analysis of Conversational Speech of Aphasic Patients. *Literary and Linguistic Computing,* 11(3), 133-140.

Honoré, A. (1979). Some Simple Measures of Richness of Vocabulary, *Association of Literary and Linguistic Computing Bulletin*, 7:172-177.

Howes, D. H. (1964) Application of the Word-Frequency Concept to Aphasia", in A. V. S. de Rueck and M. O'Connor (Eds.) *Disorders of Language*, Ciba Foundation Symposium, London: Churchill.

Hyde, J. S. (1996) Gender and cognition - a commentary on current research. *Language & Individual Differences*, 8 (1): 33-38.

Jollife, I.T. (1986) *Principal Components Analysis*. New York: Springer-Verlag.

Mckeever, W. F. (1996) Cognitive gender differences - where are they and why are they there - considering the interrelationship between biological and environmental factors - response. *Learning & Individual Differences*, 8 (1): 69-72.

Mulac, A. and Lundell, T. L. (1994) Effects of gender-linked language differences in adults written discourse - multivariate tests of language effects. *Language & Communication*, 14 (3): 299-309.

Ojemann, G. (1983) Brain Organization for Language from the Perspective of Electrical Stimulation Mapping. *Behavioral and Brain Sciences*, 6, 189-230.

Robinson, N. M., Abbot, R. D., Berninger, V. W. and Busse, J. (1996) The structure of abilities in math-precocious young children - gender similarities and differences. *Journal of Educational Psychology,* 88 (2): 341-352.

S. Singh and T. Bookless. Analyzing Spontaneous Speech in Dysphasic Adults, The *International Journal of Applied Linguistics,* vol. 7.2, no. 2,

pp. 165-182, 1997.

Singh, S. Designing Intelligent Interfaces for Users with Memory and Language Limitations, *Aphasiology,* 2000, 14(2):157-177, 2000.

Singh, S. (1994) *C*omputational Linguistics in Speech and Language Therapy, *Proc*. *5th International Conference on Speech Science and Technology,* Perth, Australia, vol. 2, pp. 486-491.

Stumpf, H. and Stanley, J. C. (1996) Gender-related differences on the college boards advanced placement and achievement tests, 1982-1992. *Journal of Educational Psychology*, 88 (2): 353-364.

Voyer, D., Voyer, S. and Bryden, M. P. (1995) Magnitude of sex differences in spatial abilities - a meta analysis and consideration of critical variables. *Psychological Bulletin*, 117 (2): 250-270.

Zipf, G. K. (1932) *Selected Studies of the Principle of Relative Frequency in Language.* Cambridge, Massachusetts: Harvard University Press.

Table 1. Data description

Variable	Minimum	Maximum	Mean	STDEV
Noun-rate				
male	12.65	19.11	15.79	1.91
female	11.12	18.09	14.47	2.18
Pronoun-rate				
male	13.19	18.83	15.90	1.89
female	13.60	19.03	16.68	1.84
Adjective-rate				
male	4.82	11.28	6.98	1.66
female	5.75	8.60	6.79	.77
Verb-rate				
male	16.63	24.52	20.97	2.15
female	18.13	27.36	22.05	2.42
TTR				
male	.27	.35	.31	.02
female	.24	.34	.30	.02
CSU-rate				
male	10.97	16.68	14.47	2.04
female	13.17	18.49	16.14	1.58
W				
male	13.75	15.29	14.50	.40
female	13.95	16.21	14.69	.55
R				
male	1402.55	1863.65	1632.59	135.37
female	1212.06	1669.50	1474.12	119.38

Table 2. Correlation analysis

	N-rate	P-rate	A-rate	V-rate	TTR	CSU	W
P-rate							
male	**-.80**						
female	-.31						
A-rate							
male	+.49	**-.74**					
female	-.31	-.26					
V-rate							
male	-.58	**+.82**	**-.71**				
female	-.32	**+.77**	-.06				
TTR							
male	+.57	**-.72**	+.69	-.56			
female	+.44	-.45	+.19	-.19			
CSU-rate							
male	-.17	+.52	-.55	+.51	-.36		
female	+.65	+.23	-.54	+.13	+.07		
W							
male	-.60	**+.70**	-.68	+.52	**-.97**	+.31	
female	-.44	+.35	-.19	+.11	**-.97**	-.04	
R							
male	+.40	-.68	+.58	-.54	**+.70**	-.67	**-.75**
female	+.30	-.29	+.22	-.23	**+.84**	-.08	**-.86**

*All correlation coefficients greater than 0.7 have been highlighted.

Table 3. Classification summary

Actual Group	*No. of Cases*	*Predicted Group Membership* *1* (female)	*2* (male)
Group 1 (female)	17	16 (94.1%)	1 (5.9%)
		*14 (82.3%)	*3 (17.7%)
Group 2 (male)	13	2 (15.4%)	11 (84.6%)
		*5 (38.4%)	*8 (61.6%)

Overall correct classification rate = 90% (*73.3% with cross-validation)

*results with leave on out method of cross-validation

Table 4. Pooled-within group correlation between discriminating variables and canonical discriminant function

R	.63592
CSU-rate	-.47091
N-rate	.32559
V-rate	-.23810
P-rate	-.21324
W	-.19897
TTR	.17461
A-rate	.07659

Table 5. Principal Components Analysis

Eigenvalue	4.324	1.392	1.062
Proportion	0.541	0.174	0.133
Cumulative	0.541	0.715	0.847
Variable	PC1	PC2	PC3
N-rate	0.299	0.410	-0.434
P-rate	-0.399	0.163	0.380
A-rate	0.311	-0.366	0.144
V-rate	-0.332	0.245	0.554
TTR	0.413	0.306	0.251
C-rate	-0.235	0.626	-0.274
W	-0.397	-0.354	-0.323
R	0.401	0.043	0.311

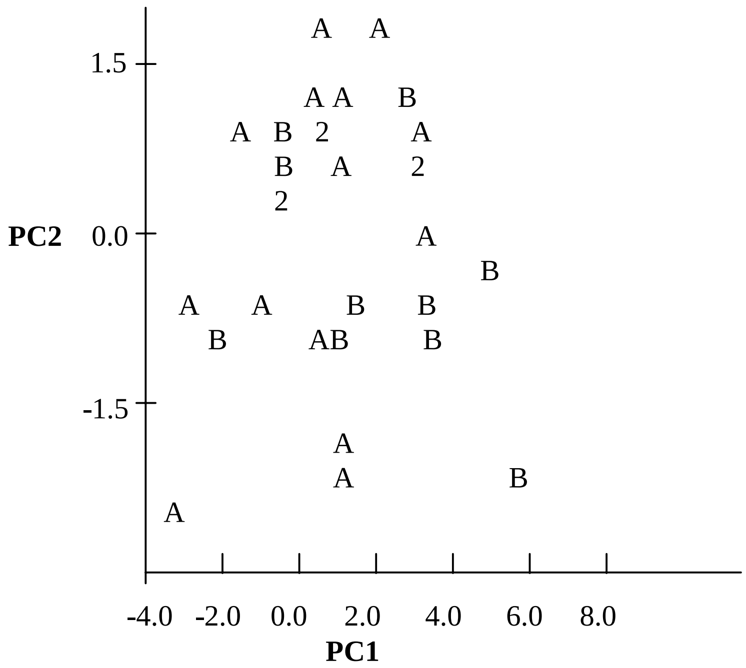
Figure 1 All-groups Stacked Histogram

Canonical Discriminant Function 1

```
        4 +                    2            2
          |                    2            2
          |                    2            2
F         |                    2            2
r       3 +           1        112          2
e         |           1        112          2
q         |           1        112          2
u         |           1        112          2
e       2 +           1        111      2  22                       2
n         |           1        111      2  22                       2
c         |           1        111      2  22                       2
y         |           1        111      2  22                       2
        1 +           111      111111 2  21     2         2         2
          |           111      111111 2  21     2         2         2
          |           111      111111 2  21     2         2         2
          |           111      111111 2  21     2         2         2
          +-------------------------------------------------------------------
X
            out    -4.0        -2.0           .0        2.0         4.0     Function  1
score
```

note: 4 sets of 1's denote one female subject and 4 sets of 2's denote one female subject

Figure 2. PCA plot (PC1 vs. PC2)



note: Female subjects are represented by ‘A’ and male subjects by ‘B’. 2 represents one male subject together with a female subject