The effect of voice disorders on lexical tone variation: Exploratory study in an African language

Gail Jones, Anita van der Merwe, *Lynda Olinger, Mia le Roux, Jeannie van der Linde

Department of Speech-Language Pathology and Audiology, University of Pretoria, South Africa.

*IQVIA, The Human Data Science Company™, South Africa.

Corresponding author: Prof Anita van der Merwe. Email: anita.vandermerwe05@gmail.com

Running head: Lexical tone variation in voice disorders

Keywords: Lexical tone variation, Voice disorders, Setswana, Tonal minimal pairs

Abstract

Purpose. The aim was to determine if the presence of a voice disorder in speakers of Setswana, an African tone language, will negatively impact the accuracy of identification by typical first language judges of words belonging to tonal minimal pairs.

Method. A quasi-experimental between-group comparison and individual case studies were conducted. Five participants with different types and degrees of voice disorders and nine control participants produced 10 tonal minimal word pairs. Five judges had to identify which of a pair was produced.

Result. The mean scores of the control and experimental speakers as groups differed, but the difference was not statistically significant. Control participants scored between 19.6/20 and 14.2/20 words correctly identified. Individual data revealed that four of the nine control participants attained at least one perfect score across judges and six had mean scores of 18.0/20 and higher. The highest scoring experimental participant, presenting with a mild voice disorder, attained a mean of 18.0/20. The lowest scoring participant, presenting with the most severe dysphonia, had a mean of 12.2/20 words correctly identified.

Conclusion. These preliminary results appear to suggest that a severe voice disorder could compromise lexical tone variation and by implication the intelligibility of a message.

Keywords: Lexical tone variation, Voice disorders, Setswana, Tonal minimal pairs

Introduction

Seventy percent of the world's languages are tone languages and are spoken natively in Africa, Asia and the Americas (Wong, Perrachione, Gunasekera & Chandrasekaran, 2009; Yip, 2002). In tone languages, voice pitch variation at the word and syllable level is implemented phonologically to convey and distinguish lexical meaning. The nature of tone variation on syllables of a word could differentiate the meanings of two phonologically similar structures (Cole, 1992; Yip, 2002). Tone variation in tone languages is referred to as lexical or syllabic tone. Lexical tone variation should be distinguished from intonation. Both depend on changes in voice pitch, but intonation is a prosodic feature that is realized at the sentence level (McCabe & Altman, 2017; Zerbian & Barnard, 2008). Speech, language and hearing disorders could negatively impact the ability to produce or perceive lexical tone variation and, as a result, the communication abilities of an individual (Van der Merwe & Le Roux, 2014a; 2014b; Wong et al., 2009). In view of the high number of tone language speakers across the world, greater emphasis should be placed on understanding the communication difficulties of these individuals in order to address assessment and remediation more effectively (Wong et al., 2009). In the current study the focus is on the effect of a voice disorder on lexical tone variation. Voice disorders could potentially interfere with tone variation due to compromised vocal fold movement and configuration of the glottis (Nguyen & Kenny, 2009). The inability to vary voice pitch adequately may impact the ability of an individual to control lexical tone and therefore to convey an intelligible message.

Research regarding the impact of speech pathology on tone variation has focused mostly on Asian languages. Studies involving Cantonese and Mandarin speakers with cerebral palsy and dysarthria (Ciocca, Whitehill, & Ma, 2004; Jeng, Weismer & Kent, 2006),

Chinese alaryngeal speakers (Yiu, Van Hasselt, Williams, & Woo, 1994) and Cantonese speakers with Parkinson's disease (Whitehill & Wong, 2007) suggest that tone is prone to disruption in tone language speakers who present with speech disorders. The possible impact of muscle tension dysphonia on tone variation was explored in a study involving Vietnamese speaking teachers (Nguyen & Kenny, 2009). Vietnamese implements pitch variation and phonation type (laryngealization and breathiness) to form six tonal distinctions. Nguyen and Kenny (2009) found that muscle tension dysphonia interferes with tone phonation by lowering tonal fundamental frequency (F0) in high tones and tones with extensive F0 variation. They also included a perceptual task performed by listeners who had to subjectively assess two parameters, identification and intelligibility, of tone-bearing syllables produced in isolation. These results indicated that perception was compromised for tones with extensive F0 variation and with no typical phonation type. The results reported by Nguyen and Kenny propose that muscle tension dysphonia will negatively impact tone production and perception in Vietnamese. However, the type of voice disorder and the tone characteristics of a specific language could determine the potential effect on tone variation.

The current study involved speakers of Setswana, a two tone language. Setswana (also referred to as Tswana) belongs to the Bantu language family and is spoken in South Africa. All Bantu languages are two-tone languages and all distinguish between high and low tones. These are, however, relative terms and are not absolute values. Syllabic tone is high or low relative to the tone of an adjacent syllable in a word (Zerbian & Barnard, 2008). Syllables may consist of a single consonant (C), a single vowel (V) or a CV shape. Tone is present across all vowels and syllabic consonants. The tone variation pattern within a word consists of a specific sequence of tone heights (high or low tone). For example, a Setswana

word with a CVCV structure may have four potential patterns of tone variation, namely: high-high (HH); high-low (HL); low-low (LL); or low-high (LH). Tone plays a role in distinguishing meaning between two phonologically similar words (Cole, 1992: Snyman, 1989). Such words are described as tonal minimal pairs. For example, the high-low pattern in *bó.nà* (see) distinguishes the word from *bò.ná* (they) that has a low-high pattern (Van der Merwe & Le Roux, 2014a; 2014b). Tone is indicated in these examples, but is not present in ordinary orthography. A relatively small change in tone production can change the meaning of a word or render it unintelligible (Yip, 2002).

To study the effect of voice disorders on lexical tone variation, the identification of words from a tonal minimal pair, produced by speakers with a voice disorder and judged by typical first language (L1) speakers, was selected as research strategy by the authors (Jones, 2016; Jones, Van der Merwe, Olinger, Le Roux & Van der Linde, 2018). Tonal minimal pairs present the opportunity to isolate and study tone variation while other word features remain constant. The potential effect of sentence intonation is also eliminated. By assessing the accuracy of identification by a listener of a word from a tonal minimal pair, inferences can be made regarding the effectiveness of tone variation by the speaker. A preliminary list of tonal minimal word pairs that could serve as research stimuli was developed. The list was refined during a series of studies (see detail in the section on Stimuli) (Jones, 2016; Jones, et al., 2018). Familiarity with the words in the list had to be ensured as words unfamiliar to a participant would impact validity of the data. Multilingualism and differences in language background, which are prevalent in Bantu-language speakers, determine the vocabulary of an individual. The current study reports data gathered with the final 10-pair (20 words) word list that resulted from the validation process.

Purpose of the research

The aim of the study was to determine if the presence of a voice disorder in a speaker influences the accuracy of identification of words belonging to a tonal minimal pair as perceived by typical L1 Setswana-speaking judges (typical listeners). The effect of a voice disorder was assessed by comparing the scores of typical (control) L1 Setswana speakers without voice- or other communication disorders to those of L1 Setswana-speaking individuals with voice disorders (experimental participants).

The specific objectives were to compare the performance of typical speakers and speakers with voice disorders and determine: (1) if there was a significant difference between the mean number of words correctly identified by judges; (2) the number of words perceived as unintelligible; and (3) whether there were any perfect scores (100% correct identification by judges). An additional objective was to compare the performance of the individual experimental speakers who presented with different voice disorders.

Method

Research design

A mixed-method approach was followed (Bryman, 2006). First, a quasi-experimental between-group comparison was performed (Mitchell & Jolley, 2010). The primary dependent variable was number of words correctly identified by the judges. Second, individual case studies were conducted to corroborate the results of the quantitative analysis and provide further insight into the comparison of results. The case studies were exploratory in nature and entailed gathering information regarding the voice disorder with which each experimental participant presented.

Participants

All participants met the following inclusion criteria: L1 speaker of Setswana who used the language daily; lived in an urban area of Gauteng to control exposure to vocabulary as far as

possible; proficient in English or Afrikaans to ensure effective communication with the researcher (English/Afrikaans-speaking first author); schooled by medium of Setswana and English and could read these languages fairly well; hearing thresholds at 500Hz, 1000 Hz and 2000Hz not exceeding 25dB on a hearing screening test; no history of psychological or cognitive conditions (for example, depression or dementia); no language disorder (anomia and/or syntactic or comprehension problems) as judged by the reseacher and a Setswana-speaking speech pathology assistant during an interview; between 18 and 65 years of age to limit the potential effect of pitch breaks associated with male puberty and age-related changes to the larynx or auditory system.

Experimental participants. Five participants (P1 – P5) with voice disorders were in the experimental arm of the study. Purposive sampling took place at out-patient clinics of public sector hospitals in urban areas of Gauteng (province in South Africa). Participants met the following inclusion criteria: presence of a voice disorder as diagnosed by an ear, nose and throat (ENT) specialist using a flexible laryngeal endoscope at the time of the study; a mild to severe dysphonic voice as judged in consensus by two speech-language pathologists (first author and a colleague) with three years' experience in treating voice disorders in a hospital context and using the GRBASI (G=grade, R=roughness, B=breathiness, A=asthenia, S=strain, I=instability) rating scale (De Bodt, Wuyts & Van de Heyning 1997); no other speech disorder (dysfluency or articulation errors) or oral structural (abnormal occlusion, use of dentures, tongue thrust, abnormality of the hard or soft palate, asymmetrical position of the lips, tongue, jaw, velum or face during rest and movement) impairments as judged by the first author during an interview**.** Experimental participant information is summarised in Table I.

Place Table I approximately here

Control group. Nine typical speakers (C1 – C9) were in the control arm of the study. Participants met the following inclusion criteria: A typical voice; no language, speech or oral structural impairment. Five of these participants were nurses and four were college tutors teaching Setswana. Their ages ranged between 18 and 60 years: C1 = 46; C2 = 41; C3 = 60; C4 = 57; C5 = 55; C6 = 21; C7 = 18; C8 = 21; C9 = 20 years.

Judges. Five individuals (J1 to J5) participated in the perceptual analysis of recorded data. Judge 1 was the wife of P2, was 60 years old and had completed 10 years of schooling. She grew up in an urban area of Gauteng and lived in the North-West Province (where Setswana is indigenous) during the study. J2 to J5 were college tutors and their ages varied between 18 and 22 years. They constitute a homogeneous group of typical L1 Setswana speakers. They were of similar age and all grew up in the North-West Province, which determines geographical dialect exposure. They had all completed 12 years of schooling and were all employed by the same institution in an urban area of Gauteng.

Ethical considerations. Permission from appropriate authorities and ethical clearance from a Faculty Research Ethics Committee had been obtained prior to commencement of the research. Before participation in the study, all control and experimental participants read an information letter and completed informed consent documentation, which was available in both English and Setswana. Participants were informed that they will be required to produce a list of Setwana words or judge the meaning of words, but not that the focus was on differential tone variation and the effect of a voice disorder.

Stimuli

Experimental stimuli consisted of 10 Setswana minimal pairs (20 words) that are phonologically similar but, which differ according to tone pattern (see Table II). Both words

in a pair had to be either a verb or a noun. Verbs were preceded by the Setswana infinitive prefix *go*.

The word list was the product of a four-phase validation process (Jones, 2016; Jones et al., 2018). During Phase 1 a preliminary list with 45 pairs was compiled from dictionaries. During Phase 2 familiarity with the words in the list was assessed by a group of nine L1 speakers. To control for exposure to vocabulary in a multilingual society, speakers who lived and worked in the urban areas of Gauteng were selected. Based on these results which showed that not all assessors were familiar with the words, the list was narrowed down to 20 pairs. During Phase 3 this list, a picture that illustrates each word and a descriptive sentence in both Setswana and English were validated by ten other L1 speakers. The sentences were included to ensure that the participant would understand the meaning of the word. Four pairs were not consistently familiar and the list was further narrowed down to 16 pairs. Some pictures were also changed. During Phase 4 the 16-pair experimental list was further validated by nine other L1 speakers and five judges. Data were also collected from individuals with voice disorders. Their data were not utilized to validate the list but were extracted for the purpose of the current study. Based on the word-specific results of Phase 4, the list was narrowed down to a final list with 10 pairs. The six lowest scoring words (mean between 42% and 64%) and their respective pairs were omitted. In the current study the data collected with the final list of 10 pairs during Phase 4 are reported.

The twenty words were subsumed in four lists (A, B, C, and D). Each contained the same words, but they appeared in a different order. Lists were randomised across participants to prevent the judges from becoming familiar with the order in which words appear. Words from a pair were also randomized across a list. For each list a manual was prepared in advance. In the manual, each word appeared on a separate page together with

a picture, a Setswana and an English descriptive sentence. To conform with ordinary orthography, tone was not indicated. Participants produced only the target word.

Place Table II about here

Recording procedures

Every participant was seen individually by the reseacher in a quiet room away from any noise that may have interfered with the audio recordings. All instructions and procedures were explained to the participant prior to data collection. These instructions were also written down in a step-by-step format and were made available in both English and Setswana. Recordings were made with an Acer Aspire E15 laptop on which PRAAT (Boersma & Weenink, 2005) audio recording and play back software was installed. The sampling rate for all recordings was 44100 Hz. A Logitech USB headset with microphone (Logitech H330 USB 2.0 Stereo Gaming headphones w/Boom microphone) was used. The microphone frequency response rate was 100Hz -10000Hz.

The participants were first shown the stimulus manual (either A, B, C or D) and they were allowed sufficient time to familiarise themselves with the words they were to produce. Before the audio recording commenced, the participants were given an opportunity to practise with two unrelated words, which were not included in the list. The participant was told to say the word naturally but clearly so that it conveyed the intended meaning according to the picture and Setswana sentence. Once the researcher was satisfied that the word production was representative of naturalistic speech, she prompted the participants to say each of the target words by using the prompt “please say the next word”. At least five seconds delay after each production was imposed to allow speakers sufficient time to prepare for the next production, and to allow judges to (later) judge the production. Data collection and recording took place in this manner for all participants.

Perceptual task procedures

The recordings of both control and experimental participants were scored by all judges during a single session. Audio recordings of the control and experimental participants were randomised when played back to the judges. To reduce listener anticipation and bias the judges were blind to the fact that control and experimental participants were involved or that any had a communication disorder. The judges sat in a quiet room away from any environmental noise. The room size was 7 x 7 meters and the judges sat to the left (3) and right (2) of the researcher, 3 to 4 meters away from the speaker system. The Acer Aspire E15 with PRAAT software (Boersma & Weenink, 2005) was used to play back the recordings. External speakers (JVC: UX-P3) were plugged into the laptop to control the intensity and sound quality of the audio presentations. Intensity was adjusted until it suited all judges (75dB). Each was given a page with written instructions in English and Setswana. The first author went through the instructions step-by-step. The judges were shown a stimulus manual and were given sufficient time to familiarise themselves with the stimuli.

Judges were each given a listener score sheet (either list A, B, C, or D) which correlated with the stimulus list that the particular participant had used. Words from a tonal pair appeared in two columns on the list together with an English translation for each in brackets to convey the meaning of the target word (Example**:** *pàpá* [father] – *pápà* [porridge])**.** Tone was indicated to aid word recognition. Though this is not used in ordinary orthography, L1 speakers are familiar with the concept. The judges listened to each participant's audio recording, played back to them as a group, and were requested to indicate on a score sheet which word of a pair they heard. In a third column there was the option to indicate that the word was unintelligible or not clearly recognisable as one of the two words. They were encouraged to tick this column if they were not sure. The options

were clearly defined and the judges had to tick one of the three options for each of the words they heard. Each stimulus was presented once. They were given four minutes to rest in between judging each participant's recording.

To determine intra-rater reliability, five of the word productions by a speaker were replayed to the judges. These were taken randomly from the list and varied across lists. The judges were unaware that some words were presented a second time. Only one word from a pair was repeated. During Phase 4 of the stimuli validation process, the judges listened to 14 speakers (9 + 5) each producing 37 (32 of the 16-pair experimental list + 5 repeated) words. For the current study, the data of 280 words (14 individuals x 20 words) are reported.

Data analysis

The total number of utterances correctly identified by each judge per participant was determined. These scores were then used to calculate a mean percentage score for each participant.

Statistical analyses. All data were analysed with a statistical software package, Stata Release 12 (StataCorp, 2011). The statistical aim was to determine whether a significant difference existed between the scores of the control and experimental groups. Normality of the data was assessed by means of histograms and quantile plots (Ghasemi & Zahediasl, 2012).

A Wilcoxon Rank test, which tests whether two independent sample groups of non-normally distributed data show a difference that is statistically significant (Vassarstats, 2012), was used. In addition to testing for a significant difference between the control group scores and the experimental group scores, the effect size difference was also determined. The effect size measures the magnitude of the difference between the two groups' results (Tavakoli, 2012). As the data was not normally distributed the non-

parametric estimator for common language effect size was calculated (Ching-Hong, 2016). The effect size determines the probability of a randomly drawn control group participant scoring higher than a randomly drawn experimental group participant. The probability is expressed as a percentage, and is determined by comparing the frequency of scores obtained by control group participants to the total number of scores, as obtained by participants in both groups.

Descriptive statistics. To explore the individual performance of control and experimental participants across judges, descriptive statistics were implemented. The mean percentage of words correctly identified by the judges per participant, the number of words perceived as unintelligible per participant, and the number of perfect scores (100% of words correctly identified) per judge were determined. The range of scores across judges for each participant was also determined. The descriptive statistics could collectively provide indications of the impact of a voice disorder on tone production, taking into consideration the range of performance of typical speakers and the possible effect of the tone identification ability of an individual judge in the context of single word recognition.

Reliability

Intra-rater reliability. A point-to-point intra-rater (judge) comparison was carried out on the data collected during Phase 4. If the judge demonstrated consistency for a given word, a score of 1 was recorded. If the judgements differed, a score of 0 was assigned. A judge could obtain a perfect score of 45 (100%) for the control participants (5 words x 9 participants) and 25 (100%) for experimental participants (5 words x 5 participants). For control participants, four of the five judges obtained 86% or higher (95% highest score) and Judge 4 obtained 71%. For experimental participants the scores of four of the judges varied between 84% and 92% and Judge 4 again had the lowest score of 64%.

Inter-rater reliability. To determine if the results of all judges could be included in the study, two measures were used to calculate inter-rater reliability. These were the Kappa Statistic (Landis & Koch, 1977), which aims to provide a quantitative measure of agreement between judges, and the Spearman's Rank Coefficient, which indicates the linear relationship between two variables (Vassarstats, 2012). The correlation coefficients were determined by calculating the number of times a judge correctly perceived each word for each of the participants and comparing the number of correct responses to the number of correct responses of the other four judges.

The Kappa score was 0.44 for the control group and 0.41 for the experimental group, indicating a moderate level of agreement between the judges. The correlation coefficients, adjusted for multiple comparisons, showed a strong positive linear relationship and a highly significant correlation between the judges in most of their assessments. Most values were closer to 1.0 than to zero (see Table III). Correlations were not significant between Judges 1 and 2 for the control group ($p=0.093$), and between Judges 1 and 4 ($p=0.438$) and Judges 2 and 5 ($p=0.567$) for the experimental participants. All other correlations were statistically significant ($p < 0.05$). The positive correlation between most of the judges' scores implies that judges were more likely to score in agreement. Judge 4, who displayed the lowest intra-rater reliability, showed a positive correlation in most instances. Analyses were run without the data of Judge 4, and there was no significant difference to the results. The Kappa statistic remained 0.41 (experimental group) for Judges 1, 2, 3, and 5 and increased slightly to 0.45 (from 0.44) for the control group. The results of Judge 1 who was the wife of P2, which could have introduced bias, showed a positive correlation with the performance of Judges 2, 3 and 5. Also, she was not the highest performing judge for any of the participants.

High levels of agreement between judges is an indication that the data are reliable. Based on these outcomes the data of all judges were utilized.

Place Table III about here

Results

Words correctly identified

The mean percentage of words correctly identified across judges for both control and experimental participants are presented in Table IV. As a group the control participants attained a mean of 87.6% (SD=9.3) (17.6/20 words) correct while the experimental participants attained 78.6% (SD=11.0) (15.7/20 words)**.** The difference between the two mean scores is 9.0%. The Wilcoxon Rank test indicated no significant difference in scores between the two groups (p=0.109, z=1.604).

Although no significant difference was found, the effect size of the difference between the two groups was 76.7%. An effect size inferring 76.7% probability corresponds to a 'medium-sized' effect (Cohen 1988). This indicates that there may well be a clinical significance (Cooper, Wears, & Schriger, 2003) and that a voice disorder could negatively impact tone variation. Clinical significance assesses the magnitude of the difference between the two groups by means of effect sizes and is not affected by sample size.

Place Table IV about here

Number of words perceived as unintelligible

Table V displays the total number of words, per participant, which were perceived by a judge as unintelligible. None of the control participants produced a word that was perceived as unintelligible. Of the experimental participants, P1 and P2 each had one unintelligible word as judged by one Judge each. Their grade ratings on the GRBASI scale were G_1 and G_2

respectively. P3, who was the only participant with a G_3 rated dysphonia, had three words that were unintelligible to one or more judges.

Place Table V about here

Number of perfect scores

The number of perfect scores (100% correct identification of all words in the list) obtained across participants and across judges is summarised in Table VI. All judges correctly identified all words for at least one control participant, while no judge was able to identify all words for any of the experimental participants. Both Judges 2 and 4, who perceived more words than the other judges as unintelligible (see Table V), did display 100% correct identification for one control participant each.

Place Table VI about here

The range of scores of individual control and experimental participants

In Figures 1 and 2 an overview of each participant's performance is displayed. These data complement the information in Table IV. The figures provide each participant's range of scores from the highest to the lowest score that was obtained across the five judges. The mean score, which is an average of all five judge's scores for a participant, is also indicated.

Place Figures 1 and 2 approximately here

Four of the nine control participants (C1, C4, C5, C9) attained at least one highest score of 20/20 words correctly identified across judges and six had mean scores of 18.0/20 and higher. Five of them had mean scores higher than any of the experimental participants. C7 and C8 displayed lower mean scores than the other seven control participants. Their lowest score across judges was 14/20.

Experimental participants P1 ($G_1R_1B_0A_0S_0I_1$), P2 ($G_2R_2B_2A_1S_1I_1$) , P4 ($G_2R_1B_1A_1S_2I_1$), and P5 ($G_2R_2B_1A_1S_0I_0$) had mean scores that were within the range of the control

participants (see Table IV and Figure 2). Of the experimental participants, P1 attained the highest mean score and a highest score of 19/20 across judges (See Figure 2). This participant, who was the only experimental participant with a rating of '1' for Grade (G), presented with a mild voice disorder due to a vocal fold nodule. The "grade" ratings on the GRBASI scale of the other three participants, mentioned above, were G2. Their highest scores were 17/20, 18/20 and 18/20 and mean scores across judges were 15.4, 17.0 and 16.0 respectively. P2 attained slightly lower scores than P4 and P5. P4 was dysphonic since a thyroidectomy and P5 had a unilateral vocal fold paralysis. P2 presented with gastroesophageal reflux disease**,** a chronic post-nasal drip, oedema of the posterior commissure, and had a polyp that was excised one month before data collection took place.

The poorest performance was from P3, who obtained a mean score of 12.2/20 (61%). The overall 'grade' (G) of his dysphonia was rated as '3' ($G_3R_3B_0A_1S_1I_0$). The range of his scores across judges was between 12/20 and 13/20 words correctly identified.

Discussion

The current study explored the possible impact of a voice disorder on high-low lexical tone variation during the production of Setswana tonal minimal word pairs. No statistically significant difference was found between the control and experimental groups. The outcome of the group comparison was due to the wide range of performance of typical speakers and listeners. Two control speakers (C7 and C8) had notably lower mean numbers of words correctly identified by the judges than the other control speakers. However, they did not produce any words perceived as unintelligible. This finding suggests that some tone variation did occur, but it was not in all instances sufficient to allow for consistently accurate discrimination by all listeners. In addition, judges were not equally able to correctly identify a word. The auditory perception of tone by Judges 2 and 4 was inferior to that of the other

judges. However, both these judges were able to correctly identify all words for at least one control speaker. One possible explanation is that these judges did not remain focused during the identification task. Collectively the performance of the control participants and judges appears to reflect the ability of typical speakers and listeners to produce and perceive tone variation in a tonal minimal pair word production and perception task. The fact that words were not produced and judged in a sentence could explain these results. The context of a sentence could aid word identification.

Though the group comparison did not show a significant difference, the effect size statistical analysis did point to a 76.7% probability that a control participant will perform better than a speaker with a voice disorder. Supporting this outcome are the findings that only speakers with voice disorders produced words that were unintelligible to judges and that 100% correct identification was only attained for control speakers. Furthermore, an analysis of the individual data of the participants who presented with voice disorders suggests that the presence of a severe voice disorder could impact lexical tone variation negatively. P3, who had the lowest mean number of words correctly identified, was diagnosed with laryngeal tuberculosis (TB), also called tuberculous laryngitis, and received medical treatment for the condition at the time of the study. This condition causes laryngeal and supraglottic lesions which include mucosal hyperemia, thickening, granulomas and ulcerations (Durand, Joseph & Baker, 1998). P2, who had the second lowest mean number of words correctly identified, also presented with conditions which could have affected the histology of the vocal folds. Histological changes of the vocal folds could potentially cause frequency perturbation and aperiodicity of vibration. Structural lesions affect the mass and stiffness characteristics of the vocal folds and could interfere with phonation by compromising the mucosal wave and vocal fold approximation (Colton, Casper & Leonard,

2011; Ferrand, 2012). Increased mass or decreased elasticity of vocal folds would inhibit pitch variation and have an effect on tone variation.

It is important to note that a limited number of individuals who present with a voice disorder took part in the current study. Further research that focuses on the impact of specific voice disorders should be undertaken, involving larger numbers of participants with different degrees of dysphonia. Acoustic analysis of change in fundamental frequency of voice across syllables will further augment the data gathered from listener judgement. For perceptual analysis more listeners than was used in the current study, should be involved. A larger number of typical speakers should also be included in comparative studies.

Clinicians treating individuals who speak tone languages should be aware of the potential negative impact of a voice disorder on intelligibility. In tonal languages, voice assessment and management should address pitch variation ability and the implementation in word production. A list of tonal minimal word pairs could be a valuable clinical tool.

Conclusion

No significant difference was found between the two groups of speakers with and without voice disorders regarding the ability to vary tone appropriately on Setswana minimal word pairs. However, individual data provide preliminary indications that a severe voice disorder could compromise lexical tone variation and by implication the intelligibility of a message. The strongest evidence of the negative impact of a voice disorder on lexical tone variation was found in the case of a participant with a severe organic condition. This exploratory study should be augmented by further research.

Declaration of interest: The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

REFERENCES

Boersma P, Weenink D. (2005). PRAAT: Doing phonetics by computer (Version 4.3.01): Computer program. http://www.praat.org/ accessed 5th January 2015.

Bryman, Y. (2006). Integrating quantitative and qualitative research: How is it done? *Qualitative Research, 6(1)*, 97-113.

Ching-Hong, C. (2016). Effect size measures in a two-independent samples case with nonnormal and nonhomogeous data. *Behaviour Research Matters, 48(4)*, 1560-1574.

Ciocca, V., Whitehill, T. L., Ma, J. K. Y. (2004). The impact of cerebral palsy on the intelligibility of pitch-based linguistic contrasts. *Journal of Physiological Antropology and Applied Human Science, 23*, 283-287.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences*. (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cole, D. T. (1992). *An introduction to Tswana grammar*. Johannesburg: Longman Penguin SA.

Colton, R. H., Casper, J. K., & Leonard, R. (2011). *Understanding voice problems: A physiological perspective for diagnosis and treatment*. (4th ed.). Philadelphia, PA: Lippincott Williams & Wilkins.

Cooper, R. J., Wears, R. L. & Schriger, D. L. (2003). Reporting research results: Recommendations for improving communication. *Annals of Emergency Medicine, 41*, 561-564.

De Bodt, M. S., Wuyts, F. L., & Van de Heyning, P. H. (1997). Test-retest study of the GRBAS scale: Influence of experience and professional background on perceptual rating of voice quality. *Journal of Voice*. 11, 74-80.

Durand, M., Joseph, M., Baker, A. S. (1998). Infections of the upper respiratory tract. In

A. S. Fauci, E. Braunwald, K. J. Isselbacher, J. D. Wilson, J. B. Martin, D. L. Kasper, et al. (Eds.). *Harrison's Principles of Internal Medicine* (p. 183). (14th ed.). New York, NY: McGraw-Hill.

Ferrand, C. T. (2012). *Voice disorders: Scope of theory and practice*. Boston, MY: Pearson Education, Inc.

Ghasemi A., Zahediasl S. (2012). Normality Tests for Statistical Analysis: A Guide for Non-Statisticians. International Journal Endocrinology and Metabolism. 10(2):486-489.

Jeng, J., Weismer, G., Kent, R. D. (2006). Production and perception of Mandarin tone in adults with cerebral palsy. *Clinical Linguistics and Phonetics, 20(1)*, 67-87.

Jones, G. L. (2016). *Tone variation in Tswana-speaking individuals: The effect of voice disorders*. University of Pretoria, Unpublished Master's dissertation.

Jones, G., Van der Merwe, A., Van der Linde, J., Le Roux, M. (2018). Development of a Setswana tonal minimal pair word list as research tool. *South African Journal of African Languages, 38(2)*, 127-135.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement and categorical data. *Biometrics, 33*, 159-174.

Mitchell, M. L., & Jolley, J. M. (2010). *Research design explained*. (7th ed.). Boston, MY: Wadsworth Cengage Learning.

McCabe, D. J., & Altman, K., W. (2017). Prosody: An overview and applications to voice therapy. Global Journal of Otolaryngology, 7(4), 1-8, MS ID 555719.

Nguyen, D. D., & Kenny, D. T. (2009). Effects of muscle tension dysphonia on tone phonation: Acoustic and perceptual studies in Vietnamese female teachers. *Journal of Voice, 23*, 446-459.

Snyman, J. W. (1989). An introduction to Tswana phonetics. Hout Bay: Marius Lubbe.

StataCorp. (2011). *Stata Statistical Software: Release 12*. College Station, Texas: StataCorp.

Tavakoli, H. (2012). *A dictionary of research methodology and statistics in applied linguistics*. Tehran: Rahnama Press.

Van der Merwe, A., & Le Roux, M (2014a). Dysarthria and apraxia of speech in selected African languages: Zulu and Tshwana. In N. Miller & A. Lowitt (Eds.), *Motor speech disorders: A cross-language perspective* (pp. 125-142). Bristol: Multilingual Matters.

Van der Merwe, A., & Le Roux, M. (2014b). Idiosyncratic sound systems of the South African Bantu languages: Research and clinical implications for speech-language pathologists and audiologists. *South African Journal of Communication Disorders, 61*, 22-29.

Vassarstats (2012). Concepts and Applications of Inferential Statistics. http://vassarstats.net/textbook/index.html/, accessed 22nd October 2015.

Whitehill, T. L. & Wong, L. L-N. (2007). Effect of intensive voice treatment on tone-language speakers with Parkinson's disease. *Clinical Linguistics and Phonetics, 21*, 919-925.

Wong, P. C. M., Perrachione, T. K., Gunasekera, G., & Chandrasekaran, B. (2009). Communication disorders in speakers of tone languages: Etiological bases and clinical considerations. *Seminars in Speech and Language 30(3)*, 162-173.

Yip, M. (2002). *Tone*. Cambridge: Cambridge University Press.

Yiu, E. M-L., Van Hasselt C. A., Williams, S. R., & Woo, J. K. S., (1994). Speech intelligibility in tone language (Chinese) laryngectomy speakers. *International Journal of Language and Communication, 29(4)*, 339-347.

Zerbian, S. & Barnard, E. (2008). Phonetics of intonation in South African Bantu languages. *Southern African Linguistics and Applied Language Studies*, 26, 235-254.

Table I. Description of experimental participants.

Participant	Age (years)	Gender	Type of voice disorder	GRBASI Score	Onset of voice disorder
P1	30	Male	Right vocal fold nodule	$G_1R_1B_0A_0S_0I_1$ (mild dysphonia based on the Grade rating of 1)	History of voice problems for 2 years; Diagnosis made 4 months before study and voice treatment up to time of study.
P2	51	Male	Excised polyp. Gastroesophageal reflux disease and chronic post-nasal drip present	$G_2R_2B_2A_1S_1I_1$ (moderate dysphonia based on the Grade rating of 2)	Polyp diagnosed 7 months earlier and excised one month before study. Slight oedema of posterior commissure.
P3	63	Male	TB Larynx	$G_3R_3B_0A_1S_1I_0$ (severe dysphonia based on the Grade rating of 3)	Tuberculosis (TB) of the larynx diagnosis made 2 years before study.
P4	53	Female	Dysphonic since thyroidectomy. Vocal folds mobile bilaterally	$G_2R_1B_1A_1S_2I_1$ (moderate dysphonia based on the Grade rating of 2)	Surgery 3 years before study.
P5	60	Male	Right recurrent nerve vocal fold paresis. No known cause.	$G_2R_2B_1A_1S_0I_0$ (moderate dysphonia based on the Grade rating of 2)	Onset 4 weeks before study.

Table II: Word list with 10 tonal minimal pairs (20 words)*

	Setswana word	**English translation**
1.	gò sèlwà (gò **sêlwa)	to pick up, to find
2.	gò sélwà (gò **sêlwa)	to oversleep, to wake late
3.	pàpá	father
4.	pápà	porridge
5.	màbòkó (**mabôkô)	brains
6.	màbòkò (**mabôkô)	praise poems
7.	màfátlhà	twins
8.	màfàtlhà	lungs, breasts
9.	màfùlò (**mafulô)	pastures
10.	màfúlò (**mafulô)	foam, froth
11.	mòlàlà	neck of a mammal
12.	mòlálá	leftover food
13.	gò bákà	to bake bread
14.	gò bàkà	to praise in song or word
15.	gò dùmà	to roar, e.g. a lion
16.	gò dúmà	to spray with insecticide
17.	gò bálélá (gò *balêla)	to cause to choke
18.	gò bàlèlà (gò **balêla)	to count for
19.	gò fìtlhà	to arrive
20.	gò fítlhà	to hide

*Reproduced from *South African Journal of African Languages* (2018) 38(2): 127-135 with permission

**These words contain a circumflex on a specific vowel in ordinary orthography. Note that tone is not indicated in ordinary orthography.

Table III. Correlation coefficients between the judges' scores of both control and experimental participants. Statistical significance (p) is also indicated.

Control Group	Judge 1 r(*p*)	Judge 2 r(*p*)	Judge 3 r(*p*)	Judge 4 r(*p*)	Judge 5 r(*p*)
Judge 1	1.00				
Judge 2	0.57 (0.093)	1.00			
Judge 3	0.68 (0.010)	0.75 (0.001)	1.00		
Judge 4	0.79 (<0.001)	0.61 (0.045)	0.70 (0.007)	1.00	
Judge 5	0.71 (0.005)	0.65 (0.018)	0.87 (<0.001)	0.89 (<0.001)	1.00
Experimental Group					
Judge 1	1.00				
Judge 2	0.61 (0.042)	1.00			
Judge 3	0.68 (0.009)	0.77 (<0.001)	1.00		
Judge 4	0.46 (0.438)	0.67 (0.011)	0.73 (0.002)	1.00	
Judge 5	0.78 (<0.001)	0.43 (0.567)	0.72 (0.004)	0.63 (0.031)	1.00

Table IV. Mean percentage of words and mean number of words correctly identified by the judges for control (C) and experimental participants (P).

Control participants	Mean percentage of words and mean number of words (n=20) correctly identified across five judges	Experimental participants	Mean percentage of words and mean number of words (n=20) correctly identified across five judges	Difference between experimental and control groups (p value)
C1	92 (18.4)			
C2	90 (18.0)			
C3	93 (18.6)	P1	90 (18.0)	
C4	98 (19.6)	P2	77 (15.4)	
C5	95 (19.0)	P3	61 (12.2)	
C6	84 (16.8)	P4	85 (17.0)	
C7	71 (14.2)	P5	80 (16.0)	
C8	73 (14.6)			
C9	95 (19.0)			
Group mean	87.6% (17.6/20)	Group mean	78.6% (15.7/20)	9.0% (p = 0.109) z=1.604

Table V. Total number of words, as produced by control and experimental participants, that were perceived by the judges to be unintelligible, and the GRBASI score of each experimental participant.

Participant	Number of words that were perceived as unintelligible	GRBASI score
C1 to C9	0	Not applicable
P1	1 (by Judge 2)	$G_1R_1B_0A_0S_0I_1$
P2	1 (by Judge 4)	$G_2R_2B_2A_1S_1I_1$
P3	3 (by Judges 2, 3 and/or 4)	$G_3R_3B_0A_1S_1I_0$
P4	0	$G_2R_1B_1A_1S_2I_1$
P5	0	$G_2R_2B_1A_1S_0I_0$

Table VI. Number of perfect scores (100% correct identification of all words in the list) obtained by judges across the control (n=9) and the experimental participants (n=5).

Judges	Number of perfect scores for control participants (n=9)	Number of perfect scores for the experimental participants (n=5)
Judge 1	3	0
Judge 2	1	0
Judge 3	2	0
Judge 4	1	0
Judge 5	2	0