LINGUISTIC COMPUTING IN SPEECH AND LANGUAGE DISORDERS

Sameer Singh

Department of Mathematical Sciences
University of the West of England
Bristol BS16 1QY
email: s2-singh@csd.uwe.ac.uk

ABSTRACT

This paper describes the methodology of using linguistic measures to test the quality of spontaneous speech for patients with speech and language disorders. The discussion involves agrammatic patients with language problems. The paper discusses the various measures, their usefulness and the actual method of analysis.

1.0 INTRODUCTION

Speech and language are closely related for any kind of oral communication. It is usual to find those who suffer speech and language deficits, also display a variety of other impairments in their reading, writing and listening abilities. The classification of such disorders, both in theory and practice, has been quite controversial. Speech disorders are more often a result of the loss of proper muscular control or the incoordination of the speech mechanism. The language disorders on the other hand arise out of damage to the brain hemispheres and syntactically or semantically impair the speech. In the past, it has been nearly impossible to classify or quantify such disorders, due to their complex and overlapping nature. Wertz(1985) lists speech and language deficits under the umbrella of five neuro-pathologies, namely, Aphasia, Language of Confusion, Language of Generalised Intellectual Impairment, Apraxia of Speech and Dysarthrias. The problem of quantifying speech and language impairments can be easily appreciated by the recognition of the fact that there exist more than 111 different kinds of aphasias only. Despite the magnitude of such a problem, there are some elementary rules which help in roughly suggesting the kind of disorder a patient may have. These include considering the localization of lesions, cause of illness and performance on tests.

Speech and language therapy aims to improve the performance of patients in oral communication, if possible, otherwise helps them to acquire compensatory skills of communication like writing, gesturing, etc. In the past, a number of tests have been designed to guage the performance of patients under clinical conditions. Although such tests provide valuable statistics to the therapists, it does little to reflect the ability of a patient to perform adequately in their natural environment. Hence it is not uncommon to find patients who do very well on tests, but completely fail to communicate either at their home or workplace, or vice-versa. Speech and language therapy progresses in more than one step as described below.

i) *Localization:* The general localization of lesions is established in order to confirm the kind of disorder. It has been mostly found that damage to the left hemisphere often results in syntactically

impaired speech and damage to the right hemisphere in semantically impaired speech.

ii) *Cause:* The majority of patients suffer C.V.A. Other causes include tumour, trauma, hemorrhage, parkinson's disease, etc.

iii) *Appraisal:* This would involve gathering the biographical, medical and behavioral data. The test scores are included.

iv) *Diagnosis:* It is a comparision of the appraisal data with the already formed opinion of the therapist.

v) *Prognosis:* A statement is made about the patient's potential to improve highlighting the areas of treatment.

vi) *Focussing: Therapy* Once the area of work has been determined, the method of treatment is chosen and the treatment is initiated.

In the next section, we will discuss and familiarise ourselves with some conventional methods of gathering medical data.

2.0 ANALYSING SPEECH AND LANGUAGE

Medical data is an essential pre-requisite for diagnosis, prognosis and focussing therapy. A number of specialised tests concentrate on specific aspects of speech and language. The general language ability can be tested by Boston Diagnostic Aphasia Examination, Western Aphasia Battery, Comprehensive Examination for Aphasia, Minnesotta Test for Differential Diagnosis, PICA and others. Apart from these, a number of other tests evaluate the functional language ability, motor speech, orientation, nonverbal intelligence, verbal fluency and comprehension. They are useful in generating a clinical profile of the speech and language disorders. If the patients are tested on language, motor speech, intelligence and orientation, their different collective score patterns reflect different kinds of disorders themselves. The tests are necessary for understanding both the kind and extent of disorder. However it is unfortunate that a single test does not qualify as a comprehensive testing procedure and the diversity of the various tests prevents them from being combined in a meaningful way. Also conventional testing procedures are fraught with problems. The most important ability of normal spontaneous speech is hardly tested. Some old statistical methods like regression analysis are still used for predicting speech recovery, which does not prove very useful in most of the cases.

3.0 LEXICAL RICHNESS MEASURES

We will now address the topic of quantifying spontaneous speech. Such a procedure will utilise lexical richness measures, which can statistically analyse normal conversation with patients. We are conducting research at UWE, Bristol, which aims to establish an optimum set of measures for agrammatic patients. Some further discussion on agrammatism can be found in Kean(1977). Lexical richness theory assumes that as a patient recovers, his improved language will manifest itself in the form of enhanced vocabulary and better scores on certain linguistic parameters. The

vital step is to establish the right set of measures in such a way, that their ease of use is instrumental in their acceptance by the therapists for daily usage. Some of these measures have been derived from the authorship attribution studies in literature as discussed by Holmes(1992). A number of linguistic aspects need considertion. Such indices will work on a given sample of normal speech and look at the usage of verbs, nouns, pronouns, adjectives, clauses, discourse markers, time and number expressions, etc. It is important for them to remain independent of the text length and the topic of conversation. Sometimes it may be impossible to ensure this in the absolute sense. The following measures are supposed to be good contenders for the evaluation of normal spontaneous speech.

i) *Rate of Usage of Nouns:* This may be measured as the rate of nouns per 100 words.

ii) *Rate of Usage of Pronouns:* This is measured as rate of usage per 100 words. A very high rate of using the word I has been observed.

iii) *Rate of Usage of Adjectives:* This is also measured as rate per 100 words. A higher value should indicate a better quality speech. They are important in qualifying the vivid imagination of the topic under consideration.

iv) *Rate of Usage of Verbs:* Their rate per 100 words is a good indicator of ability to describe actions or processes. It has been conventionally assumed that dysphasic adults have a reduced rate of verb usage, which has been proved wrong according to our studies on a set of 20 patients. In general, a higher rate of usage should imply a better organised and articulate speech.

v) *Clauses/ Semantic Units:* In agrammatic speech, clauses are difficult to mark out in the strict sense, so most of our work concentrates on analysing the clause like, flexible semantic units. It can be appreciated that a lower rate of semantic units per 100 words will imply individual units containing more words. This is an indication of the ability to join more words together in the form of a meaningful semantic unit. Hence a lower rate means a better quality speech and vice-versa. The ability of patients to hold words together in a buffer and to access them to form longer sentences has always been questionable and hopefully this parameter will help in measuring this ability.

vi) *Words within Units:*

A simple bar/line graph can be drawn to illustrate the frequency of using N words in a semantic unit as shown in figure 1. If the peak occurs at a higher value of N, it represents a better speech. Some credit should also be given to patients who have maximum number of words in a clause like semantic unit.

vii) *Brunet's Index:*

This is a measure of vocabulary richness and is also termed as the 'W' index. It is defined by

$$W = N^{V^{-0.165}}$$

The lower the W value, the richer the vocabulary. Typically it varies between 10 and 20. Here V_i is the number of words spoken only i times and N is the total text length. V is the total vocabulary.

viii) *Honore's Statistic:*

This a measure which tests the propensity of an individual to choose the alternatives of employing a word used previously or employing a new word. It is defined as

$$R = \frac{100 \log N}{1 - V_1 / V}$$

A higher R- value represents a richer vocabulary in the sense that a greater number of words appear infrequently.

ix) *Yules K:*

This is a measure of richness of vocabulary, based on the assumption that the occurrence of a given word is based on chance and can be regarded as a poisson distribution. K is defined as

$$K = 10^4 (\Sigma i^2 V_i - N) / N^2$$

where $(i = 1, 2, ...)$

Yules K can be calculated for all the words in the text or only for nouns.A lower value represents a better quality language.

x) *Type Token Ratio:*

This is the ratio of the total vocabulary to the text length and a higher value should indicate the ability of using new words, therefore a better

quality speech. On its own, this parameter can sometimes be deceiving, since it may happen patients and normal subjects come up with nearly the same type token ratios.

Once these statistical linguistic measures have been computed and the values documented into tables, we are able to serve two main objectives. It becomes possible to compare these to average normal scores to find out any differences. Also a new set of scores can be compared to the old ones after a stipulated period of therapy. In order for us to accomplish the above differences, we need to know what exactly are the average scores for normal adults and how these can be combined in a useful way to give us an overall picture of the speech quality. Some discussion on combining scores on different sections of a test is discussed by Kertesz and Poole(1974).

5.0 METHOD OF ANALYSIS

The overall process of analysing dysphasic speech involves i) *Data Collection*: Although this can be the most time consuming phase for a group study, it should be easier for studying single patients. The recording will take 20-30 minutes for a 1000 word sample. ii) *Data Transcription:* Simple data transcription should serve the purpose. iii) *Textual Analysis:* Oxford Concordance Program is run on the mainframe system. It gives a textual analysis including text length, vocabulary, frequencies of using n words, type token ratio and other valuable statistics. iv) *Manual Analysis*: The nouns, pronouns, adjectives, verbs, semantic units, time and number expressions, etc. are counted. Most of the softwares available for doing this are still not very efficient. v) *Graphical Representation*: For the purposes of comparision in a group study, line/ area graphs are drawn.

There are a number of good reasons why comparing lexical scores within a group may be important. In the event of a large group, it may be interesting to note whether patients having similar age, background, sex, handedness, cause of illness, therapies, years onset, etc., also have similar scores on some of the linguistic parameters. This could be helpful in turn for the speech and language therapists in directing their efforts in a more organised way.

6.0 FUTURE DEVELOPMENTS

It seems likely in the future, more effort will be directed towards improving spontaneous speech of a patient. According to our studies, we shall be able to assign a specific pattern to an individual patient, comprising his scores on our linguistic measures, depending on his quality of speech at any given time. It would be reasonable to attempt classification based on the overall combined scores, where the patients are classified according to their state of recovery rather than the variety of disorder. Such classification could also be acheived using Neural Networks which work with unsupervised algorithms. Any clustering effect will be quite interesting to study. Due to the wide variety of scores with high variance within individual measures, rule based systems are not ideal contenders for classification as they have been used in other branches of medicine. Neural networks can also help in prediction of future scores using previous score patterns and will hopefully perform more efficiently than regression analysis. Neural Networks have been used in linguistic

computing for attributing authors of disputed works, Tweedie(1994). Although it is an accepted practice in the speech and language therapy profession, to diagnose and treat everyone on an individual basis, progress on a wider scale can only be made by better understanding of the generic aspects of language disorders. It conclusion, what we need is a more agreed platform for treating speech and language disorders.

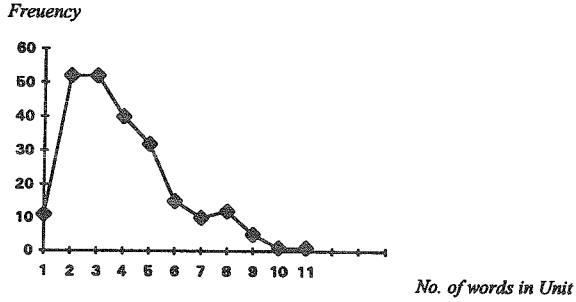

Fig. 1 (Patient EP8)

Total No. of Words Sampled	= 1125
Total Vocabulary	= 302
Total No. of Sentences	= 117
Total No. of Units	= 231
Max. No. of Words in Units	= 11
Rate of Clauses per 100 words	= 20.53
Type Token Ratio	= 0.2684

REFERENCES

Wertz, Robert T. (1985) *"Neuropathologies of Speech and Language: An Introduction to Patient Management"*, in Clinical Management of Neurogenic Communicative Disorders, Little, Brown and Company, pp. 2-17, 1985.

Kean, Mary-Louise (1977) *"The Linguistic Interpretation of Aphasia Syndromes: Agrammatism in Broca's Aphasia, An Example"*, Cognition, 5, pp 9-46.

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

Kertesz, A. and Poole, E. (1974) "*The Aphasia Quotient: The Taxonomic Approach to Measurement of Aphasic Disability*", The Canadian Journal of Neurological Sciences, pp. 7-15.

Tweedie, F., Singh, S. and Holmes, D.I. (1994) "*Neural Network Applications in Stylometry: The Federalist Papers*", presented and published in the proceedings of the 3rd International Conference on the Cognitive Science of Natural language Processing, Dublin.