USING AN AUTOMATIC WORD-TAGGER TO ANALYSE THE SPOKEN LANGUAGE OF CHILDREN WITH IMPAIRED HEARING

P.J. Blamey*, M.L. Grogan† and M.B. Shields†

*Department of Otolaryngology
University of Melbourne

†Australian Bionic Ear and Hearing Research Institute

ABSTRACT - The grammatical analysis and description of spoken language of children with impaired hearing is time-consuming, but has important implications for their habilitation and educational management. Word-tagging programs have achieved high levels of accuracy with text and adult spoken language. This paper investigates the accuracy of one automatic word tagger (AUTASYS 3.0 developed for the International Corpus of English project, ICE) on a small corpus of spoken language samples from children using a cochlear implant. The accuracy of the tagging and the usefulness of the results in comparison with more conventional analyses are discussed.

INTRODUCTION

The quantitative description of the speech and language of children with impaired hearing is an important part of their habilitation and educational management. In particular, it can be used to assess their immediate needs, and to evaluate the effects of educational programs or prosthetic devices such as hearing aids and cochlear implants. The analysis of representative conversational spoken language samples is possibly the most valid method of obtaining this type of information, however, such samples are difficult to obtain in a controlled manner, and analysis by hand is both time consuming and subjective in nature (Lund and Duchan, 1993).

In an attempt to make conversational analysis less onerous for clinicians and researchers, a suite of computer programs is being developed at the Cooperative Research Centre for Cochlear Implant, Speech, and Hearing Research. The programs are known as CASALA, for Computer Aided Speech and Language Assessment. The present study is concerned only with the syntactic component. The aims of the syntactic component of CASALA are to produce an automatic, objective, reliable description of the grammatical structures used in conversation, and to express the description in a form that is accessible and useful to clinicians and teachers working directly with the children. Quick and convenient access to this level of information is likely to help teachers and clinicians to use their time most effectively in encouraging the development of appropriate syntactic structures in the children's spoken language, and in assessing the effectiveness of their methods. The input to the syntactic analysis will be an English transcription containing the utterances of the child and the conversational partner(s) as well as any explanatory notes inserted by the transcriber. The goal is to produce a summary sheet derived from an existing procedure that is commonly used by clinicians and teachers, for example the Language Analysis, Remediation, and Screening Procedure (LARSP) of Crystal, Fletcher, and Garman (1989) or the Developmental Sentence Score (DSS) of Lee (1974).

Automatic parsers and word-taggers have been developed elsewhere as components of automatic speech recognition systems, and for documentation of English language databases. Three special considerations must be taken into account before applying these programs to the present goals. Firstly, the conversational utterances of children with impaired hearing are likely to contain a larger proportion of deviant or incomplete syntactic structures and may therefore be parsed or tagged less accurately than conversations of adults with normal hearing. On the other hand, very complex structures may occur less frequently in children's spoken language than in adults' so the analyser may be able to make some simplifying assumptions. The second (related) consideration is that it is not always possible for a skilled person to identify uniquely the functions of words in children's spoken language, leading to an inherent degree of ambiguity in any comparison of word-tagging analyses. Finally, the existing programs do not provide information in a form that is directly useful to teachers or

clinicians. As an initial approach to the first consideration, we have obtained a word tagging program and evaluated its accuracy on a small corpus of conversations from children using cochlear implants.

METHOD

Conversational samples

The conversational samples used in this study were collected as part of the speech and language evaluation of children involved in the cochlear implant program at the University of Melbourne/Royal Victorian Eye and Ear Hospital Cochlear Implant Clinic. The pre- and post-operative protocols call for evaluation at 6, 12 or 24 month intervals for pre-primary, primary, and secondary-school age children respectively. At each evaluation, a videotape was made of a conversation between the child and a normally-hearing clinician who was familiar with the child. The topic of conversation was controlled (family or food) and often a picture description (of a flood scene) was used as well as the conversation.

Each sample was transcribed by hand to give an English gloss, preferably by the clinician who was the conversational partner for the conversation, and preferably within a few days of collection of the sample. These conditions minimise the number of ambiguities or misinterpretations of the child's utterances that can arise from the poor intelligibility of their speech. The preferred method of transcription is to include both the child's and the partner's utterances, however only the child's utterances have been transcribed for most samples. Eighty transcribed language samples were available, from twenty children aged two to eighteen years, all with profound hearing impairments.

For control purposes, two spoken language samples (a conversation and a picture description) were obtained for an adult with normal hearing. These samples were transcribed and analysed in exactly the same way as the other language samples.

Word tagging

The automatic word tagger used in this study is AUTASYS 3.0, purchased from Prof. S. Greenbaum at University College London. The tagger runs on an IBM compatible PC (50 MHz 486) and is capable of tagging 15,000 items in about 5 minutes. The tagger accepts English sentences as input in a text file and uses a two-stage analysis to produce an output text file in which each word and punctuation mark is tagged with one of 256 possible labels. The first stage of analysis which is called "assignment" finds all possible tags for each word or punctuation mark. The second stage of analysis chooses between the multiple tags for words and punctuation marks that do not already have a unique tag. The tags used were developed for the International Corpus of English (ICE) and are described fully by Greenbaum (1992) and Yibin (1993). There are 22 basic word-classes (see Table 1), and subclasses differentiate specialised characteristics within each main class. Subclass distinctions include common/proper nouns, singular/plural nouns, transitivity of verbs, cases of verbs and auxiliary verbs, definite/indefinite articles, and many others too numerous to mention here.

adjective	conjunction	nominal adjective	pronoun
adverb	connective	noun	prop it
anticipatory it	existential there	numeral	reaction signal
article	formulaic expression	particle	verb
auxiliary	genitive marker	preposition	
cleft it	interjection	proform	

Table 1. Main word-classes in the ICE tagset.

Analyses

Twenty seven language samples were chosen from the eighty available transcriptions. The samples were chosen to include pre- and postoperative samples from ten children covering a wide range of ages. Both conversational samples and picture descriptions were included. Text files containing only the child's utterances, and excluding all unintelligible words were produced from the transcriptions.

Each file was analysed automatically using AUTASYS, and by hand. The hand transcriptions used the ICE tagset so that the two analyses would be directly comparable. The results below compare the automatic and hand analyses for words only (excluding tags for punctuation marks) at the word-class level and the subclass level. The word tags in the hand analyses were arrived at by consensus of two people, a speech pathologist experienced in the analysis of children's spoken language, and a speech scientist. As mentioned in the introduction, it is not always possible to assign a unique tag to each word in a child's spoken language. In these cases, the analysts made use of their knowledge of the videotaped language sample to select the word tag that was most likely to represent the intended function of each word. For example, the single-word utterance "wait" is more likely to be an imperative verb than a noun if the child is distracted from the task in hand. Similarly, "open" is more likely to be a verb than an adjective if the object referred to is actually closed at the time of the utterance. Obviously, this type of information is not available to the AUTASYS program so that the disagreements between the automatic and hand analyses should not all be considered as errors.

Child	Age (yrs & mths)	Implant Stage (a)	Sample Type (b)	No of utterances	No of words	MLU (words) (c)	Word-class disagreem't (%)	Subclass disagreem't (%)
1	4y10m	pre	conv	38	42	1.11	66.7	4.8
1	6y11m	post	conv	73	149	2.04	13.4	6.7
2	5y1m	post	conv	88	120	1.36	19.2	1.7
3	5y0m	pre	conv	32	68	2.13	10.3	2.9
3	7y3m	post	pict	50	359	7.18	7.8	1.4
4	5y4m	pre	conv	45	86	1.91	14.0	1.2
4	9y7m	post	conv	152	549	3.61	8.0	2.0
5	6y5m	pre	conv	88	139	1.58	13.7	0.7
5	11y10m	post	pict	37	177	4.78	7.9	1.7
5	11y10m	post	conv	77	361	4.69	8.0	4.4
6	7y11m	pre	conv	46	114	2.48	7.0	4.4
6	7y11m	pre	pict	46	187	4.07	18.2	3.7
6	10y2m	post	conv	84	187	2.23	4.3	6.4
7	8y6m	pre	pict	61	207	3.39	14.0	13.5
7	10y4m	post	conv	62	161	2.60	8.7	10.6
8	14y8m	pre	pict	16	98	6.13	6.1	1.0
8	14y8m	pre	conv	28	128	4.57	14.1	0.8
8	17y1m	post	pict	31	297	9.58	4.0	1.0
8	17y1m	post	conv	43	187	4.35	10.7	2.1
9	16y11m	pre	pict	55	214	3.89	15.0	4.7
9	16y11m	pre	conv	75	193	2.57	10.4	3.6
9	18y8m	post	pict	61	271	4.44	9.2	3.7
9	18y8m	post	conv	123	458	3.72	9.4	5.5
10	19y7m	pre	pict	12	73	6.08	5.5	4.1
10	19y7m	pre	conv	30	107	3.57	11.2	4.7
10	20y11m	post	pict	33	209	6.33	10.5	6.7
10	20y11m	post	conv	44	187	4.25	7.0	4.3
adult	43y7m	-	pict	45	975	21.7	6.3	2.2
adult	43y7m	-	conv	46	932	20.3	7.2	1.1

(a) "pre" indicates a pre-operative sample and "post" indicates a post-operative language sample.
(b) "conv" indicates a conversational sample and "pict" indicates a picture description.
(c) MLU indicates mean length of utterance (No of words / No of utterances).

Table 2. Results from the comparison of automatic and hand word tagging analyses. The last two columns show disagreements as a percentage of tagged words.

RESULTS

Table 2 shows the results of the comparison for the 27 language samples from the children with impaired hearing and the two control language samples from an adult with normal hearing. The average length of the children's samples was 57 utterances or 197 words, with an overall mean length of utterance (MLU) of 3.9 words per utterance. Note that the MLU used here is based on the number of words per utterance rather than the more usual number of morphemes. The overall percentage of words with different word-classes in the automatic and hand analyses was 10.2% and the overall percentage of words with the same word-class but different sub-class was 4.0%.

DISCUSSION

Word-class results

The percentages of disagreement for word-class tags cover a very wide range from 4.0% to 66.7% in Table 2, but with only one sample having a disagreement greater than 20%. This sample had an MLU of 1.11, and therefore included mostly single-word utterances that may have been ambiguous in their function. In this case, pragmatic analysis would probably be a more appropriate technique than syntactic analysis. It seems unlikely that automatic word-tagging will be a useful tool for language samples having such a low MLU, so this sample has been excluded from the discussion in the remainder of this paper. Even with this extreme case excluded, there is still a significant correlation between MLU and percentage disagreement on word-classes. The Pearson correlation coefficient, r is -0.49, ($p < 0.01$). This suggests that more complete, or possibly more complex utterances are easier to tag than shorter ones.

A chi-squared analysis showed that there were significantly more word-class disagreements for the children's samples than for the control adult samples ($\chi^2 = 15.99$, $p < 0.001$). Many of the word-class disagreements were consequences of common grammatical errors in the children's spoken language. For example, the sentence "They are stand in the house" resulted in "are" being tagged automatically as a copular verb and "stand" as a noun. The human analysts felt that "are" should be tagged as an auxiliary verb, and that "stand" should be a part of the verb, probably with the suffix "-ing" omitted. (The child actually repeated the sentence correctly in the next utterance.) This sentence thus resulted in two word-class disagreements from a single omission. A similar case occurred for "The rain stop" in which "stop" was automatically tagged as a noun, presumably because a suffix of "-s" or "-ed" had been omitted by the child. Omission of morphological endings is a well-known characteristic of the language of children with impaired hearing (Bamford & Bench, 1979). Another common characteristic of these language samples was repetition of words within sentences. For example, in "To keep keep dry from the rain" the first "keep" was correctly tagged as an infinitive verb, but the second "keep" was tagged as a noun and "dry" was tagged as a verb. The human analysts felt that "keep" should be tagged as a repeated copular verb and "dry" should be tagged as an adjective. These examples suggest that the automatic analysis of spoken language from children with hearing impairments might be improved by pre-processing the transcript to look for common misconstructions, omissions, or repetitions, and attempting to reconstruct the intended utterance before tagging. Other word-class disagreements arose from formulaic or idiomatic utterances such as "Pardon," "Sorry," and "Excuse me" which are very common in the spoken language of people with impaired hearing for obvious reasons. These are straight forward cases that are listed as examples in the ICE tagset manual, but are not handled correctly by AUTASYS. AUTASYS might also do a better analysis of children's spoken language if some of the finer distinctions between word-classes (such as pronoun/anticipatory it/prop it/cleft it, noun/nominal adjective, and conjunction/connective) were removed on the grounds that they are less likely to be used than more common constructions by children with limited language. Many of the instances where these tags were used by AUTASYS resulted in disagreements with the hand analysis.

Subclass results

In all but two cases, the percentage of subclass disagreements was well below 10%, and was not significantly correlated with either MLU ($r = -0.22$, $p = 0.28$) or percentage of word-class disagreements ($r = 0.03$, $p = 0.88$). AUTASYS seemed to be fairly accurate in subclass distinctions if the word-class

was accurately determined. Most of the subclass disagreements occurred for verb transitivity (108 instances), verb case (51 instances), pronoun type (45 instances), or noun type (17 instances). In many cases, AUTASYS overstated a verb's transitivity e.g. intransitive verbs were tagged as monotransitive, and monotransitive verbs were tagged as ditransitive. The largest group of verb case disagreements was for present tense verbs that were tagged automatically as infinitives. Most of the pronoun disagreements were for demonstrative pronouns that were tagged as relative, and interrogative pronouns automatically tagged as nominal. The noun errors arose when common nouns were automatically tagged as proper nouns, or vice versa. It was not clear to the authors whether the subclass disagreements were related to any particular characteristics of the spoken language of children with impaired hearing. A chi-squared analysis indicated significantly more subclass disagreements for the children's language samples than the control adult samples ($\chi^2 = 8.14$, $p < 0.01$).

Patterns of word-class usage

One of the potential uses of automatic word-tagging would be to produce patterns of word-class usage from language samples. This can be done relatively easily from the output files produced by AUTASYS. Figure 1 shows the patterns derived from groups of samples using the hand analysis, classified by their MLU into the four inter-quartile ranges. As one might expect, these patterns show clear changes as the MLU increases, and each group pattern is significantly different from every other pattern, as determined by chi-squared analyses ($p < 0.001$). We are also interested in whether the word-class patterns derived from automatic analysis would differ significantly from those derived from hand analyses. To answer this question, chi-squared analyses were carried out for each of the five group samples shown in Figure 1, and only one sample yielded a significant difference ($\chi^2 = 23.9$, $p < 0.05$ for the group with MLU less than 2.5) between the patterns derived from the automatic and hand analyses. Thus it seems that the present level of performance of AUTASYS is adequate for a study of patterns of word-class usage based on samples similar to those in this paper.

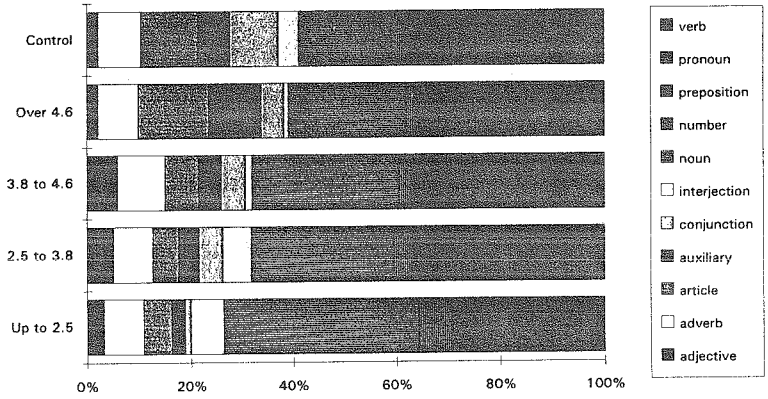

Figure 1. Patterns of word-class usage for samples grouped by mean length of utterance into the four interquartile ranges.

Clinical significance

Although the results of automatic word-tagging may provide a statistically reliable and objective description of some syntactic aspects of a child's language, they are of limited utility to a clinician whose primary goal is to improve the child's language. The clinician would like to know what the

processes are that are influencing the word patterns, and how these patterns can be changed through habilitation or learning programs. Often this will require an analysis of specific errors, deficiencies or abnormalities in the child's syntactic development, such as the omission of morphemes marking plurality, the use of inappropriate verb tenses, the use of a very restricted set of sentence structures, etc. The clinician would also like to know whether these characteristics are changing in response to therapy. As mentioned in the introduction, it is one of the aims of CASALA to produce this sort of information automatically for the clinician. It is possible that a program like AUTASYS may provide the first step towards this goal, but the translation of word-tagging information into clinically useful information is non-trivial. It is also possible that the discrepancies between the automatic and hand analyses observed in this study may make it very difficult to produce reliable information in some of the clinically important areas. For example, 30 to 60% of children's utterances contained a word-class disagreement. In the presence of these disagreements, it is unlikely that the detailed structure of the utterance could be parsed without error. If the proportion of utterances containing disagreements is to be reduced to an acceptable level (say 10 to 15%) the performance of the word-tagging program will need to be increased beyond that of the present version of AUTASYS 3.0.

CONCLUSIONS

The main conclusions of the study are as follows:

a) An automatic word-tagger (AUTASYS 3.0) resulted in an average of 90% word-class agreement with human word-taggers for spoken language samples of children with impaired hearing.

b) The agreement between automatic and human word-taggers might be improved if the automatic word-tagger was optimised to take into account common characteristics of the spoken language of children with impaired hearing.

c) The present level of agreement between automatic and human word-taggers is sufficient for the production of word-class usage patterns that differ by a statistically insignificant amount.

d) The present level of agreement between AUTASYS 3.0 and human word-taggers is likely to be inadequate for the automatic parsing of an acceptable proportion of utterances produced by children with impaired hearing.

ACKNOWLEDGMENTS

This work was carried out with financial support from the National Health and Medical Research Council and the Cooperative Research Centre for Cochlear Implant, Speech and Hearing Research. The authors would like to acknowledge the help of staff of the University of Melbourne/Royal Victorian Eye and Ear Hospital Cochlear Implant Clinic in providing the language samples for analysis.

REFERENCES

Bamford, J.M. & Bench, J. (1979) *A grammatical analysis of the speech of partially-hearing children,* in Crystal, D. (Ed) *Working with LARSP,* (Edward Arnold, London).

Crystal, D., Fletcher, P., and Garman, M. (1989) *The grammatical analysis of language disability (second edition),* (Whurr, London).

Greenbaum, S. (1992) *The ICE tagset manual,* (Survey of English Usage, University College London).

Lee, L. (1974) *Developmental sentence analysis,* (Northwestern University Press, Evanston, Illinois).

Lund, N.J. and Duchan, J.F. (1993) *Assessing children's language in naturalistic contexts (third edition),* (Prentice Hall, Englewood Cliffs, New Jersey).

Yibin, N. (1993) *Appendix to the ICE tagset manual: A list of closed-class items and a quick reference to the manual,* (Survey of English Usage, University College London).