Using corpora in the field of Augmentative and Alternative Communication (AAC) to provide visual representations of vocabulary use by non-speaking individuals

Russell Thomas Cross
Prentke Romich Company
rtc@prentrom.com

1 Introduction

The field of Augmentative and Alternative Communication (AAC) is a discipline focuses on the needs of individuals with severe speech or language problems, the severity of which is such that there is a need to supplement any existing speech or replace speech altogether.

Aided communication methods include low-tech solutions such as paper and pencil to communication books or boards populated by words and/or symbols, or devices that produce voice output (speech generating devices or SGDs) along with text output. Electronic communication aids allow the user to use picture symbols, letters, and/or words and phrases to create messages. Some devices can be programmed to produce different spoken languages.

The success individuals may have in using an SGD is heavily influenced by the amount of time spent by parents and spouses, educators, Speech and Language Therapists, in helping them to learn how to use the system (Arnott & Alm, 2013; Ball & Lasker, 2013; Travis & Geiger, 2010).

2 Improving performance using automated data logging (ADL)

Automatic data logging is a feature of some voice output communication aids. Such data can be useful in providing clinicians with information on how a client is using a device and, more importantly, how well that client is using it to communicate effectively. There are limitations to the data, which include;

- Absence of input from communication partners
- Absence of any multi-modal elements.
- Absence of social/geographical context.
- Need to mark explicitly if someone else is using the device for modeling/teaching.

Given that these limitations are recognized, it is still possible to use the information in a fruitful and constructive way. For example, one simple measure of AAC use is to count words used, which can give an idea of an individual's knowledge of the lexicon available to them in their AAC system. Another is to measure the time period between linguistic events so as to get an idea of communication rate. A third is to look at the type of words being used and determine the spread of different parts of speech.

3 Visualizing the data

One challenge with machine-logged data is that in its raw form it can be difficult to interpret. It is possible to use manual and semi-automated systems such as SALT (Miller & Chapman, 1983) AQUA (Lesher, Moulton, Rinkus, & Higginbotham, 2000), PERT (Romich, Hill, Seagull, Ahmad, Strecker, & Gotla, 2003) and QUAD (Cross, 2010) to convert such raw data into more user-friendly formats. Another method is to use specific data visualization software that is designed to convert numeric and textual data into graphic formats.

Cross (2013) developed a web-based automated data analysis software that allows for the uploading of a log file to a secure server, where it can be parsed in a number of ways to as to present summary data in the form of a visual dashboard. The current version allows for data to be analyzed in terms of;

- Word frequency
- Parts of Speech
- Performance against target vocabulary
- Daily/Weekly/Monthly device use

It's also possible to search for specific instances of words and see them in context.

4 Using the Corpus of Contemporary American English

To provide a large corpus against which client-generated utterance could be matched, the Corpus of Contemporary America English (Davies, 2008-) was used. This was chosen because not only did it provide a very large database – far larger than any currently available in the field of AAC – but it also includes frequency data and grammatical tagging based on the CLAWS system (Garside, 1987). Both word frequency and syntax (mainly in the area of morphology) are important pieces of information when monitoring the performance of an aided communicator (Binger, 2008; Binger & Light, 2008). Furthermore, such information can inform educational and clinical intervention programs (Cross, 2013).

Another feature of the database is that words are lemmatized, providing a level of analysis that has implications for the teaching vocabulary as *word sets* rather than individual lexical items. For example, if a client demonstrates the use of *jump*, *jumps*, *jumped*, *walks*, and *walking*, teaching *jumping* and *walked* to "complete the set" makes sense.

5 Outline of how the system works

The basic operation of the server is fairly simple. It consists of three elements:

(a) **Uploaded Data File**: The primary input to the system is a plain text (TXT) file that has been created by the automated data logging feature of an SGD.

All individual uploads are aggregated over time and become the basis of a "merged file" that provides a personal database of language use. It is this aggregated database that is used for all the different types of analyses the system has to offer.

(b) **Comparison Database**: Certain analyses – such as the "Parts-of-Speech" analysis, use the database in order to identify and present words. The system makes use of color coding in order to represent these in order to create, for example, a bar chart:

(c) **Analysis "widgets":** Specific analyses can be performed by selecting a "widget" - a single-hit button that triggers a particular action. For example, a "Cloud" widget looks at all the words used in the merged file within a specific time period and then displays these as a word cloud picture, where the size of a word is directly proportional to its frequency of use.

As another example, a "Weekly Use" widget counts the number of times within a 15-minute period that the SGD is used. It then displays this as a graph.

The graphical results of using any of these widgets can be saved as PNG graphics files and then used to create reports and summaries.

6 Next Steps

Using client-generated data to improve the performance of individuals who use SGDs is still relatively new. The use of large scale corpora to provide enable comparisons to be made and individual performance to be tracked is also in its infancy. This means that the metrics being used are rather broad and need to be made more granular and specific. For example, the analysis of parts-of-speech uses the global categories of noun, verb, adjective etc. but a more precise breakdown using specific CLAWS tags would yield much more information.

Another challenge is to be able to use more flexible filters in the system so as to be able to break down the data into more focused conditions. Being able to have the server handle questions such as "how many times was the *–ing* participle used one month ago compared with this week" is pedagogically valuable.

References

Arnott, J. L., & Alm, N. (2013). Towards the improvement of Augmentative and Alternative Communication through the modelling of conversation. *Computer Speech & Language*, 27(6), 1194-1211.

Ball, L. J., & Lasker, J. (2013). Teaching Partners to Support Communication for Adults with Acquired Communication Impairment. Perspectives on *Augmentative and Alternative Communication*, 22(1), 4-15.

Binger, C. (2008). Grammatical Morpheme Intervention Issues for Students Who Use AAC. Perspectives on *Augmentative and Alternative Communication*, 17(2), 62-68.

Binger, C., & Light, J. (2008). The morphology and syntax of individuals who use AAC: research review and implications for effective practice. *Augmentative and Alternative Communication*, 24(2), 123-138.

Cross, R. T. (2010). Developing Evidence-Based Clinical Resources Embedding. In Hazel Roddam and Jemma Skeat *Evidence-Based Practice in Speech and Language Therapy* (pp. 114-121): John Wiley & Sons, Ltd.

Cross, R. T. (2012). Using AAC device-generated data to develop therapy sessions. Paper presented at the *American Speech Hearing and Language Association Annual Convention*, Atlanta, GA.

Cross, R. T. (2013). The Value and Limits of Automated Data Logging and Analysis in AAC Devices. Paper presented at the *ASHA Convention*, Chicago, IL.

Davies, M. (2008-). *The Corpus of Contemporary American English: 425 million words, 1990-present*. Available online at http://corpus.byu.edu/coca

Garside, R. (1987). The CLAWS Word-tagging System. In R. Garside, G. Leech & G. Sampson (Eds.), *The Computational Analysis of English: A Corpus-based Approach* (pp. 30-41). London: Longman.

Lesher, G. W., Moulton, B. J., Rinkus, G., & Higginbotham, D. J. (2000). A Universal Logging Format for Augmentative Communication. Paper presented at the *2000 CSUN Conference*, Los Angeles. http://www.csun.edu/cod/conf/2000/proceedings/0088 Lesher.htm

Miller, J., & Chapman, R. (1983). *SALT: Systematic Analysis of Language Transcripts*. San Diego: College Hills Press.

Romich, B. A., Hill, K. J., Seagull, A., Ahmad, N., Strecker, J., & Gotla, K. (2003). AAC Performance Report Tool: PERT. Paper presented at the *Rehabilitation Engineering Society of North America (RESNA) 2003 Annual Conference*, Arlington, VA.

Travis, J., & Geiger, M. (2010). The effectiveness of the Picture Exchange System (PECS) for children with autism spectrum disorder (ASD): A South African pilot study. *Child Language Teaching and Therapy*, 26(1), 39-59.