Neologisms
• ADI-R: “non-words”.
• Volden & Lord (1991): “words that are not included in the standard lexicon of adult native language speakers”.
• Lord (1996): “use of made-up non-words as if they were words”.

Idiosyncratic Word Use
• Kanner (1946): “peculiar and out of place in ordinary conversation”.
• ADI-R: “obviously peculiar” words.
• Volden & Lord (1991): “standard, familiar words or phrases [used] in idiosyncratic, but meaningful ways”.

Use word frequencies tabulated from the Wall Street Journal (WSJ) training set of the Penn Treebank (40,000 sentences, 1 million words).

Neologisms: Words from a transcript that do not appear in the WSJ (known as out-of-vocabulary words, or OOVs) are potential neologisms.
1. Raw OOV rate: number of OOVs used divided by the number of total words used.
2. OOV type rate: number of unique OOVs used by a child divided by the number of unique words used.

Unusual Words: Words from a transcript that appear very rarely in the WSJ may be considered unusual.
1. Low frequency words: percentage of words used by a child with WSJ frequency <= 100.
2. Mid-frequency words: percentage of words used with WSJ frequency > 100, <= 10,000.
3. High frequency words: percentage of words used with WSJ frequency > 10,000.

Results: Neologisms and OOVs

Background

Motivation & Objectives
Definitions of neologism and idiosyncratic words or phrases, as presented in ADOS manual, are vague and over-inclusive: “repetitive”, “inappropriately formal”, “unusual use of words or formation of utterances”, “idiosyncratic”, “neologisms”. Broad definition may lead to errors in clinical judgment about autism-perticular language.

Objective
Determine whether significant differences between TD and ASD groups can be obtained by measuring neologism and unusual word use with 1) manual methods based on specific criteria, and 2) automated methods based on natural language processing (NLP) techniques.

Method
Automated Methods
Use word frequencies tabulated from the Wall Street Journal (WSJ) training set of the Penn Treebank (40,000 sentences, 1 million words).

Neologisms: Words from a transcript that do not appear in the WSJ (known as out-of-vocabulary words, or OOVs) are potential neologisms.

1. Raw OOV rate: number of OOVs used divided by the number of total words used.
2. OOV type rate: number of unique OOVs used by a child divided by the number of unique words used.

Unusual Words: Words from a transcript that appear very rarely in the WSJ may be considered unusual.

1. Low frequency words: percentage of words used by a child with WSJ frequency <= 100.
2. Mid-frequency words: percentage of words used with WSJ frequency > 100, <= 10,000.
3. High frequency words: percentage of words used with WSJ frequency > 10,000.

Manual Methods
Use extracted WSJ OOVs and Amazon Mechanical Turk, a web-based interface in which untrained “workers” read transcripts and identify any strange or unusual words.

Neologisms: Detailed examination of WSJ OOVs by trained linguist in order to exclude existing words, non-words subsequently explained by the child, and non-words whose meaning can be extrapolated from productive morphological processes (e.g., adding -er or -ish to an existing word).

Unusual Words: Raw Mechanical Turk data: Percentage of sentences containing words identified as strange or unusual by untrained Mechanical Turk workers.
2. Mechanical Turk data + detailed linguistic examination à la Volden & Lord (1991): Review sentences identified by workers, and determine whether the word has a:
   • non-developmental syntax or morphology error
   • developmental syntax or morphology error
   • semantic error: a non-word or inappropriate word

Results: Unusual Words and Word Use

References & Acknowledgements

Conclusions
• Given that OOVs include words that neither are neologisms nor exemplify idiosyncratic word use, the simple coarse measure of OOV rate performed remarkably well.
• NLP-based frequency statistics may capture unusual word usage patterns.
• Defining categories of unusual word use is critical, requires linguistic analysis and clinical expertise.
• Categorization by untrained annotators does not yield group differences revealed by expert annotation.

Future Work
• Explore whether measures of unusual word use can play a role in reducing ASD/DLD diagnostic substitution.
• Further refine definitions/criteria for neologistic and unusual word use to distinguish ASD from DLD.
• Use automated methods, including corpus-based NLP techniques and measures of syntactic complexity, to investigate more subtle differences in ASD vs. DLD word use.
• Combine measures of unusual word use with measures of repetitive language.
• Build algorithms to distinguish underlying causes of unusual word use (e.g., socialcommunication issues vs. word-retrieval problems), since uncertainty about source of unusual word use can contribute to diagnostic substitution.

Sponsors: National Institute on Deafness and Other Communicative Disorders, NIDCD 1R01DC007129-01; Innovative Technology for Autism grant from Autism Speaks; Autism Speaks Student Fellowship.