# Minnesota Contextual Content Analysis (MCCA)

Ken Litkowski
CL Research
ken@clres.com
http://www.clres.com
http://www.clres.com/cata/index.html

Don McTavish
University of Minnesota
Dmctavish@earthlink.net

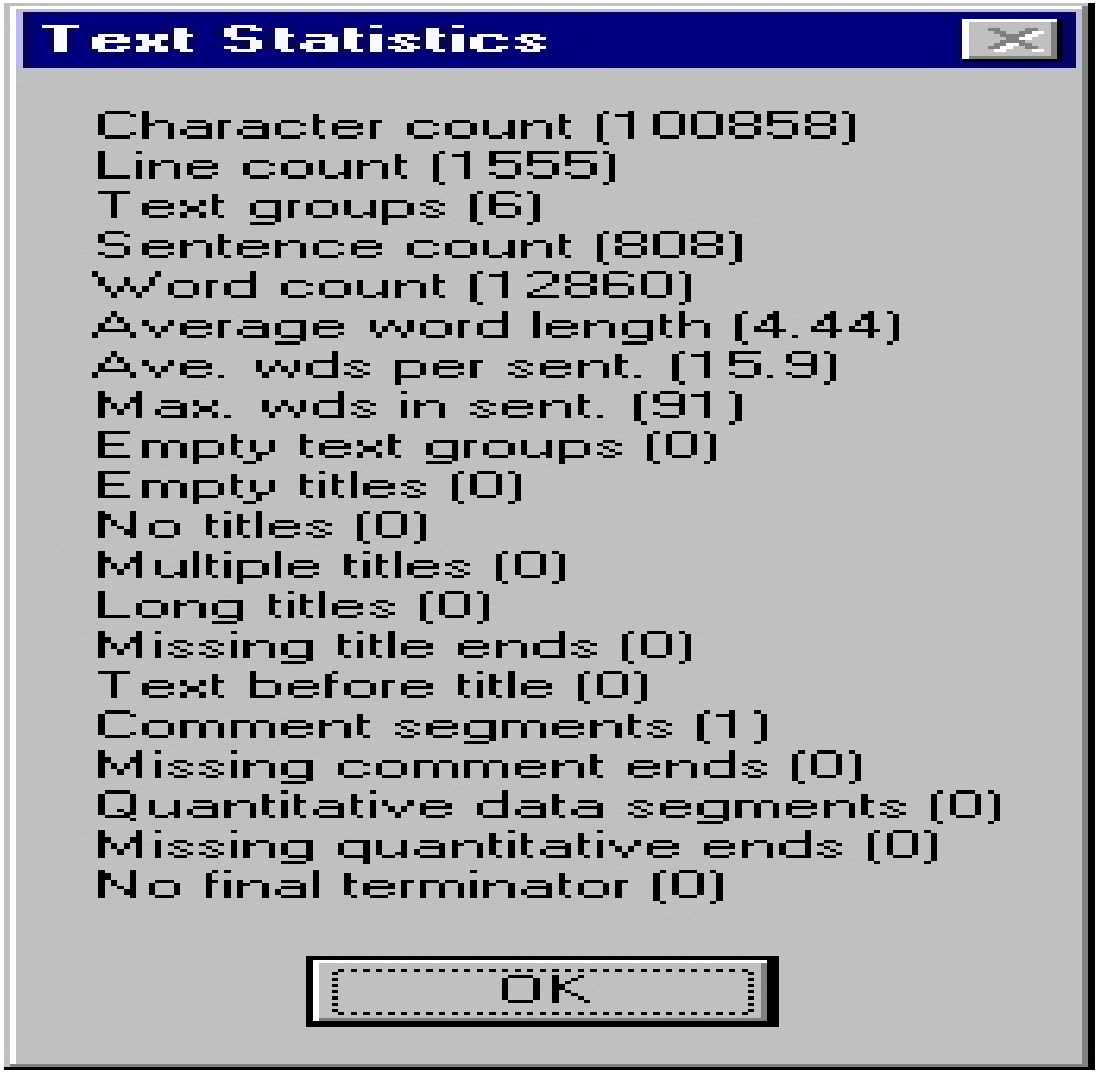
# Minnesota Contextual Content Analysis

- **Characterizing a text based on the relative frequency with which words in each category are used, compared to norms determined from general usage statistics for the English language**
- **Several statistics are generated from this analysis, for direct use or for further statistical analysis**
- **All words in one or more texts are divided into 116 idea categories (plus a "not classified" category)**
- **The MCCA dictionary groups word meanings into categories thought to express (singly or in combinations of categories) ideas important to an investigator**
- **Two kinds of normed scores (emphasis or E-scores and context or C-scores) are generated for each analyzed text**
- **Suitable for texts of any length (short open-ended questionnaire items, sentences such as Likert scale items, multi-person transcripts)**
- **Sample: Presidential announcement speeches by Bradley, Buchanan, Bush, Forbes, Gore, and McCain**

# Processing a Text

- **Text Preparation**
  - Ordinary text file, with a title field and a marker to separate texts
  - Quick overall prescreen of text
  - Specify options (e.g., use of a stop list for display and/or analysis)
- **MCCA Dictionary**
  - Over 11,000 words classified into one or more categories (with inflected forms as distinct entries
  - Dictionary stored in CL Research's DIMAP dictionary maintenance software
- **Processing**
  - MCCA dictionary is loaded
  - Tokenizer identifies each word and creates underlying statistics (about 10 minutes to process 2 MB file), disambiguating by context
  - Statistics available at completion of processing (user can click buttons or tab to examine results for individual documents or across documents)

# Text Statistics



```
Character count [100858]
Line count [1555]
Text groups [6]
Sentence count [808]
Word count [12860]
Average word length [4.44]
Ave. wds per sent. [15.9]
Max. wds in sent. [91]
Empty text groups [0]
Empty titles [0]
No titles [0]
Multiple titles [0]
Long titles [0]
Missing title ends [0]
Text before title [0]
Comment segments [1]
Missing comment ends [0]
Quantitative data segments [0]
Missing quantitative ends [0]
No final terminator [0]
```

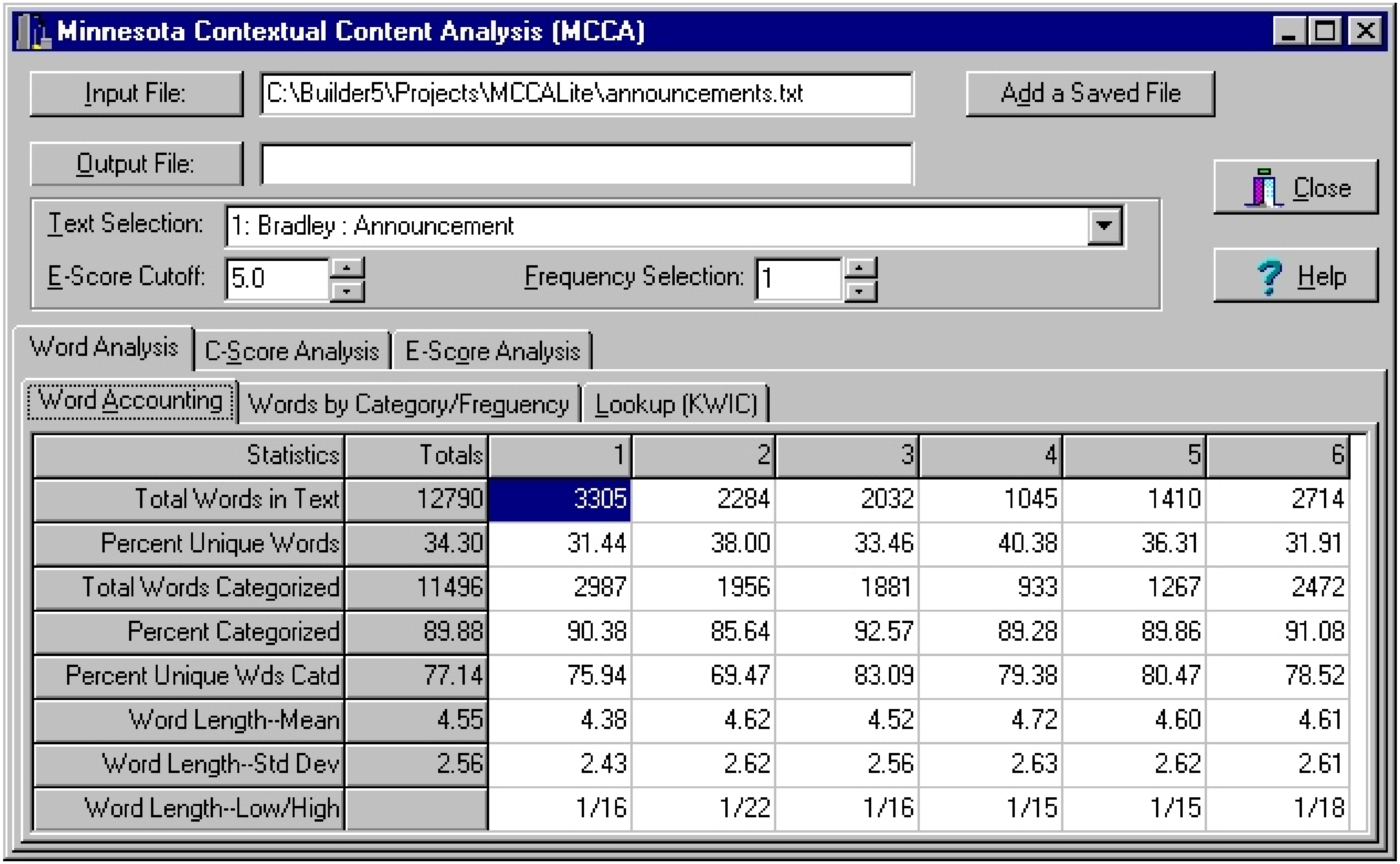
# Output Available

- Word Accounting, Lookup (KWIC), Words in Category, Word List
- C(ontext)-Score: Weighted scores and plots, Distance Matrix
- E(mphasis)-Score: High Categories, Selected Plots, Difference Analysis, Diagnostic Groups, Distance Matrix
- Other: Co-Occurrence of categories, SPSS Output (for further analyses), KYST Output (for multidimensional scaling)

# Word Analysis

- Word Accounting (for all text groups in the file and for each individual text group)
  - the total number of words; the percentage of unique words in the texts; the total number of words for which a category was available; percentage of tokens categorized, the percentage of unique words that were categorized, and average length, the standard deviation of the lengths, and the shortest and longest lengths.
- Words alphabetically, by category, by frequency
- Alphabetical list of words with counts for selection of keywords in context

# Word Accounting



Input File: C:\Builder5\Projects\MCCALite\announcements.txt

Output File:

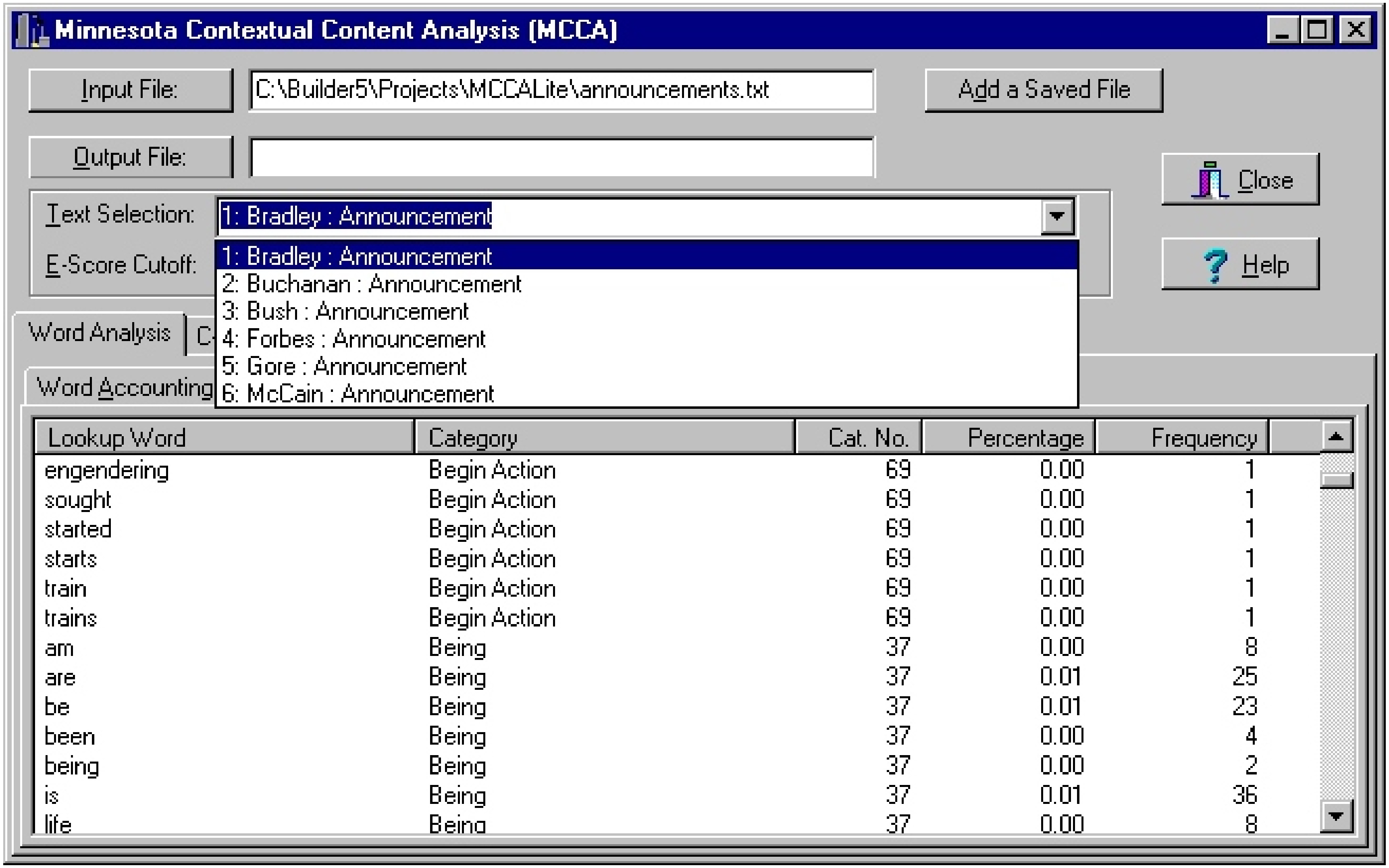
Text Selection: 1: Bradley : Announcement

E-Score Cutoff: 5.0

Frequency Selection: 1

| Statistics | Totals | 1 | 2 | 3 | 4 | 5 | 6 |
|---:|---:|---:|---:|---:|---:|---:|---:|
| Total Words in Text | 12790 | 3305 | 2284 | 2032 | 1045 | 1410 | 2714 |
| Percent Unique Words | 34.30 | 31.44 | 38.00 | 33.46 | 40.38 | 36.31 | 31.91 |
| Total Words Categorized | 11496 | 2987 | 1956 | 1881 | 933 | 1267 | 2472 |
| Percent Categorized | 89.88 | 90.38 | 85.64 | 92.57 | 89.28 | 89.86 | 91.08 |
| Percent Unique Wds Catd | 77.14 | 75.94 | 69.47 | 83.09 | 79.38 | 80.47 | 78.52 |
| Word Length--Mean | 4.55 | 4.38 | 4.62 | 4.52 | 4.72 | 4.60 | 4.61 |
| Word Length--Std Dev | 2.56 | 2.43 | 2.62 | 2.56 | 2.63 | 2.62 | 2.61 |
| Word Length--Low/High | | 1/16 | 1/22 | 1/16 | 1/15 | 1/15 | 1/18 |

# Frequency Analysis



Minnesota Contextual Content Analysis (MCCA)

Input File: C:\Builder5\Projects\MCCALite\announcements.txt

Output File:

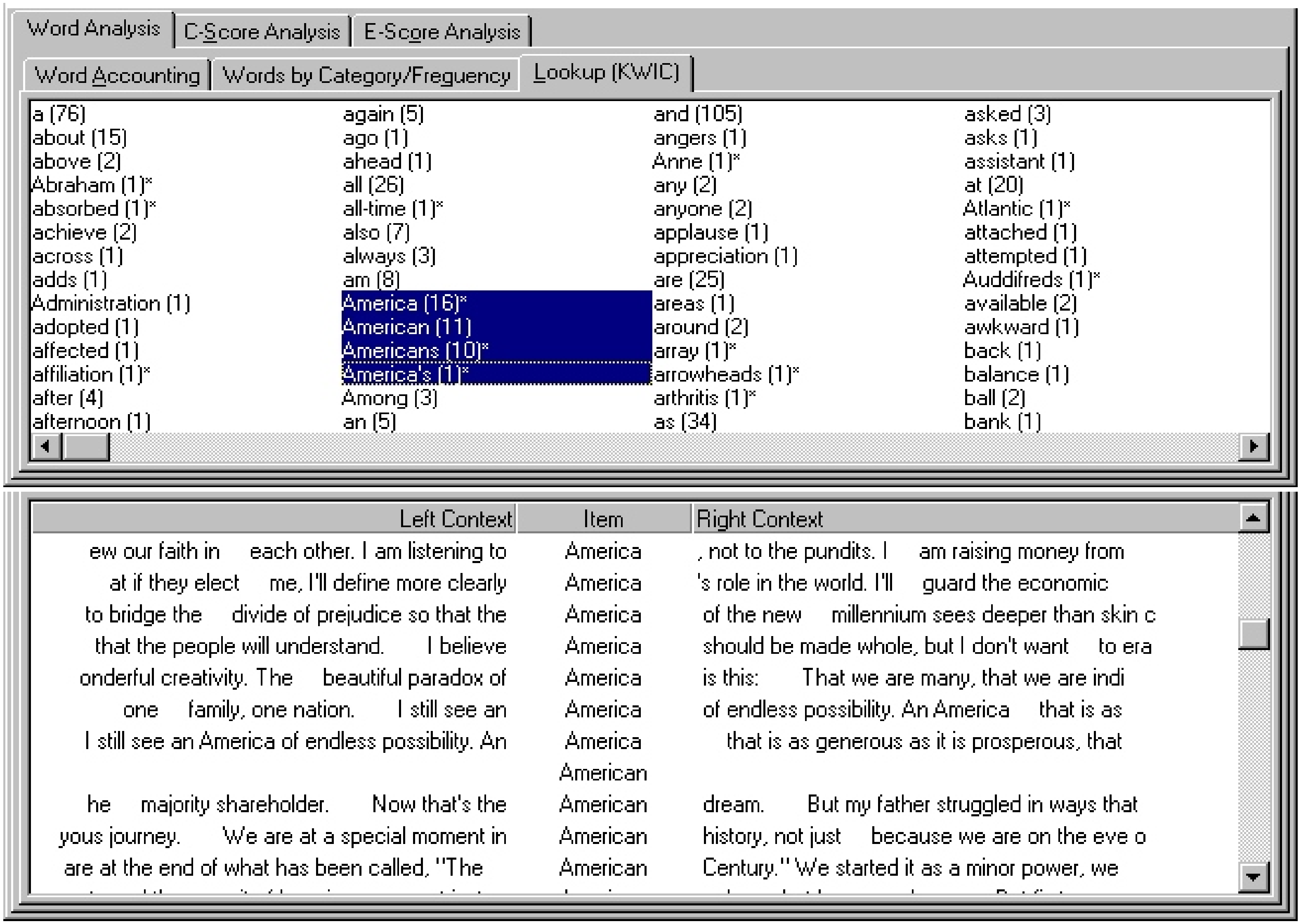
Text Selection: 1: Bradley : Announcement

E-Score Cutoff:

1: Bradley : Announcement
2: Buchanan : Announcement
3: Bush : Announcement
4: Forbes : Announcement
5: Gore : Announcement
6: McCain : Announcement

| Lookup Word | Category | Cat. No. | Percentage | Frequency |
|---|---|---|---|---|
| engendering | Begin Action | 69 | 0.00 | 1 |
| sought | Begin Action | 69 | 0.00 | 1 |
| started | Begin Action | 69 | 0.00 | 1 |
| starts | Begin Action | 69 | 0.00 | 1 |
| train | Begin Action | 69 | 0.00 | 1 |
| trains | Begin Action | 69 | 0.00 | 1 |
| am | Being | 37 | 0.00 | 8 |
| are | Being | 37 | 0.01 | 25 |
| be | Being | 37 | 0.01 | 23 |
| been | Being | 37 | 0.00 | 4 |
| being | Being | 37 | 0.00 | 2 |
| is | Being | 37 | 0.01 | 36 |
| life | Being | 37 | 0.00 | 8 |

# Keywords in Context



a (76)
about (15)
above (2)
Abraham (1)*
absorbed (1)*
achieve (2)
across (1)
adds (1)
Administration (1)
adopted (1)
affected (1)
affiliation (1)*
after (4)
afternoon (1)

again (5)
ago (1)
ahead (1)
all (26)
all-time (1)*
also (7)
always (3)
am (8)
America (16)*
American (11)
Americans (10)*
America's (1)*
Among (3)
an (5)

and (105)
angers (1)
Anne (1)*
any (2)
anyone (2)
applause (1)
appreciation (1)
are (25)
areas (1)
around (2)
array (1)*
arrowheads (1)*
arthritis (1)*
as (34)

asked (3)
asks (1)
assistant (1)
at (20)
Atlantic (1)*
attached (1)
attempted (1)
Auddifreds (1)*
available (2)
awkward (1)
back (1)
balance (1)
ball (2)
bank (1)

| Left Context | Item | Right Context |
|---|---|---|
| ew our faith in each other. I am listening to | America | , not to the pundits. I am raising money from |
| at if they elect me, I'll define more clearly | America | 's role in the world. I'll guard the economic |
| to bridge the divide of prejudice so that the | America | of the new millennium sees deeper than skin c |
| that the people will understand. I believe | America | should be made whole, but I don't want to era |
| onderful creativity. The beautiful paradox of | America | is this: That we are many, that we are indi |
| one family, one nation. I still see an | America | of endless possibility. An America that is as |
| I still see an America of endless possibility. An | America | that is as generous as it is prosperous, that |
|  | American |  |
| he majority shareholder. Now that's the | American | dream. But my father struggled in ways that |
| yous journey. We are at a special moment in | American | history, not just because we are on the eve o |
| are at the end of what has been called, "The | American | Century." We started it as a minor power, we |

CL Research

ICA/CATA-2001

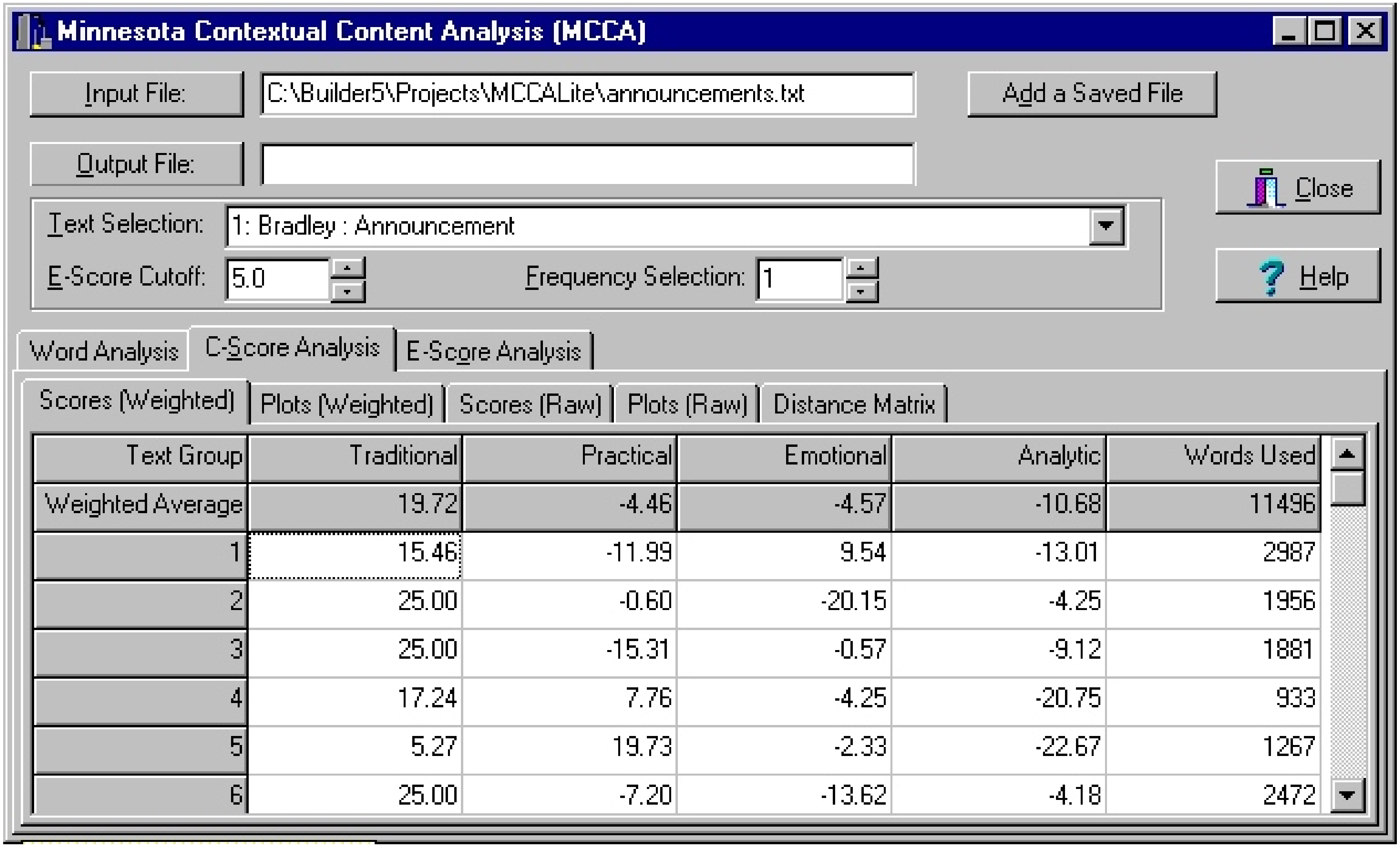
# Word Accounting Observations

- Texts range from 1045 to 3305 words
- High percent categorized indicates broad conversational style (no technical words, no proper nouns), least so for Buchanan
- Percent unique words at low end of expected 30 to 50 percent range, indicating some degree of repetitiveness, Bradley and McCain the most, and Forbes the least
- General impression indicates these announcements are trying to communicate to broad groups

# Context Score Analysis

- Analysis of words across four social contexts (practical, traditional, emotional, analytic)
  - Each context dimension is a function of the emphasis in the text across a large number of idea categories and is represented by a vector of weights
  - Contexts are experimental, empirically-derived profiles of relative emphasis on each idea category
- Normed and raw plots show emphasis on different dimensions (facilitating genre analysis)
- Distance matrix across texts allows broad interpretation of style differences
  - Standard euclidean distance computation
  - The larger the distance measure, the greater the social context distance between the two positions
  - The larger the distance measure, it is hypothesized that communications difficulties are likely to be encountered because there is little shared perspective

# Context Scores



Input File: C:\Builder5\Projects\MCCALite\announcements.txt

Output File:

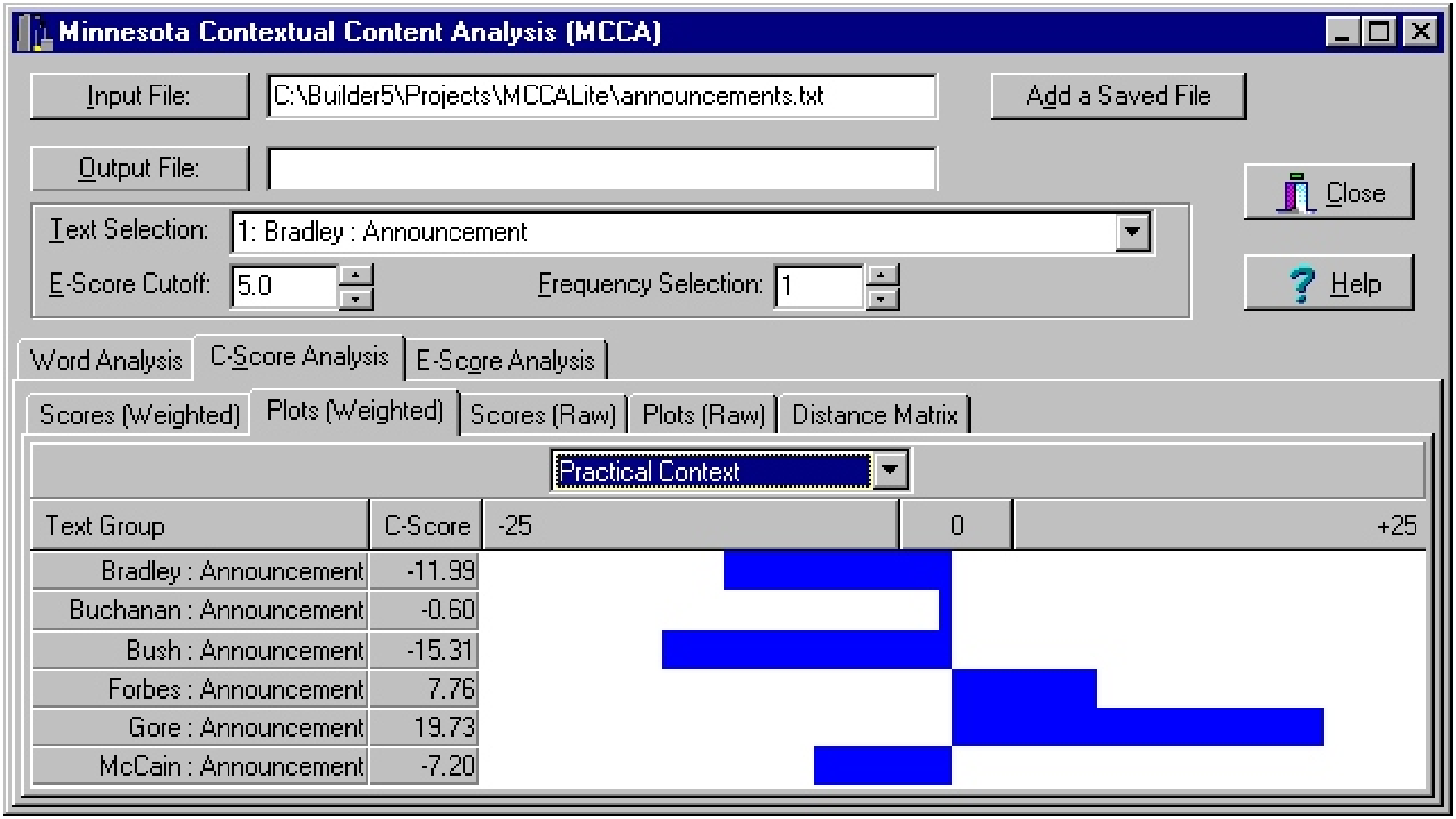
Text Selection: 1: Bradley : Announcement

E-Score Cutoff: 5.0

Frequency Selection: 1

| Text Group | Traditional | Practical | Emotional | Analytic | Words Used |
|---|---|---|---|---|---|
| Weighted Average | 19.72 | -4.46 | -4.57 | -10.68 | 11496 |
| 1 | 15.46 | -11.99 | 9.54 | -13.01 | 2987 |
| 2 | 25.00 | -0.60 | -20.15 | -4.25 | 1956 |
| 3 | 25.00 | -15.31 | -0.57 | -9.12 | 1881 |
| 4 | 17.24 | 7.76 | -4.25 | -20.75 | 933 |
| 5 | 5.27 | 19.73 | -2.33 | -22.67 | 1267 |
| 6 | 25.00 | -7.20 | -13.62 | -4.18 | 2472 |

# Context Score Plots



Input File: C:\Builder5\Projects\MCCALite\announcements.txt

Text Selection: 1: Bradley : Announcement

E-Score Cutoff: 5.0

Frequency Selection: 1

Practical Context

| Text Group | C-Score | -25 | 0 | +25 |
|---|---|---|---|---|
| Bradley : Announcement | -11.99 | | | |
| Buchanan : Announcement | -0.60 | | | |
| Bush : Announcement | -15.31 | | | |
| Forbes : Announcement | 7.76 | | | |
| Gore : Announcement | 19.73 | | | |
| McCain : Announcement | -7.20 | | | |

# Context Score Observations

- Announcement speeches are decidedly non-analytic and almost completely non-emotional (except for Bradley)
- Republicans (Buchanan, Bush, Forbes, and McCain) focus on traditional dimension to the maximum (norms and expectations for proper behavior)
- Gore with emphasis on practical dimension (focus on goal accomplishment)
  - Of all dimensions, contrast between Bush and Gore is greatest on the practical
- Gore and Bush most different among all pairs, while Buchanan and McCain are most similar, with Bradley and Bush next most similar

# Emphasis Scores

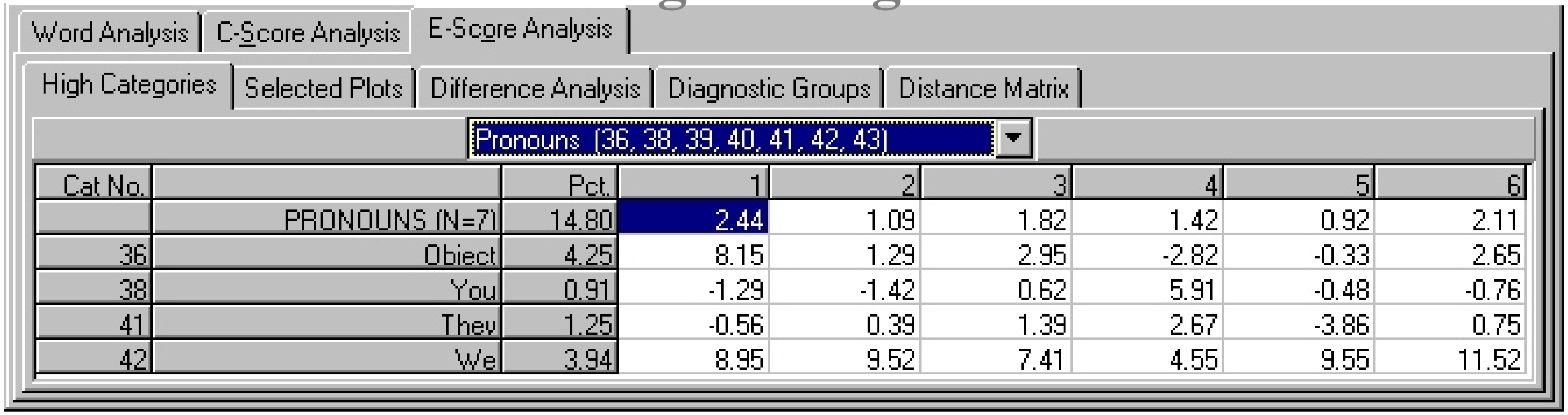
- **Shows the emphasis (called E-scores) placed on each of many (116) idea categories**
- **An idea category consists of a group of words which reflect a given idea or meaning**
- **Scores are "normed" against expected usage of the words in an idea category so that positive E-scores indicate an over-emphasis of the idea category and negative E-scores indicate a relative omission of a given idea in the text**
  - Normed scores are computed in a z-score-like fashion, contrasting category proportions with the expected probability of use of a given idea category, divided by a standard deviation of expected category usage across the four social contexts.

# Emphasis Score Analyses

- **Distance Matrix**
  - A "probability" distance between each pair of text groups in the input file over all 116 E-Score categories
  - Texts that are "more" similar to one another have lower "distances" between them.
- **High Categories**
  - Grouped into 23 super-categories, but showing only categories meeting a cutoff score
- **Difference Analysis**
  - Differences in E-scores between a reference text and all the others
- **Diagnostic Groups**
  - 43 emphasis score (EScore) combinations
- **Selected Plots**
  - Categories meeting a cutoff score showing over- or underemphasis

# Emphasis Score Analyses

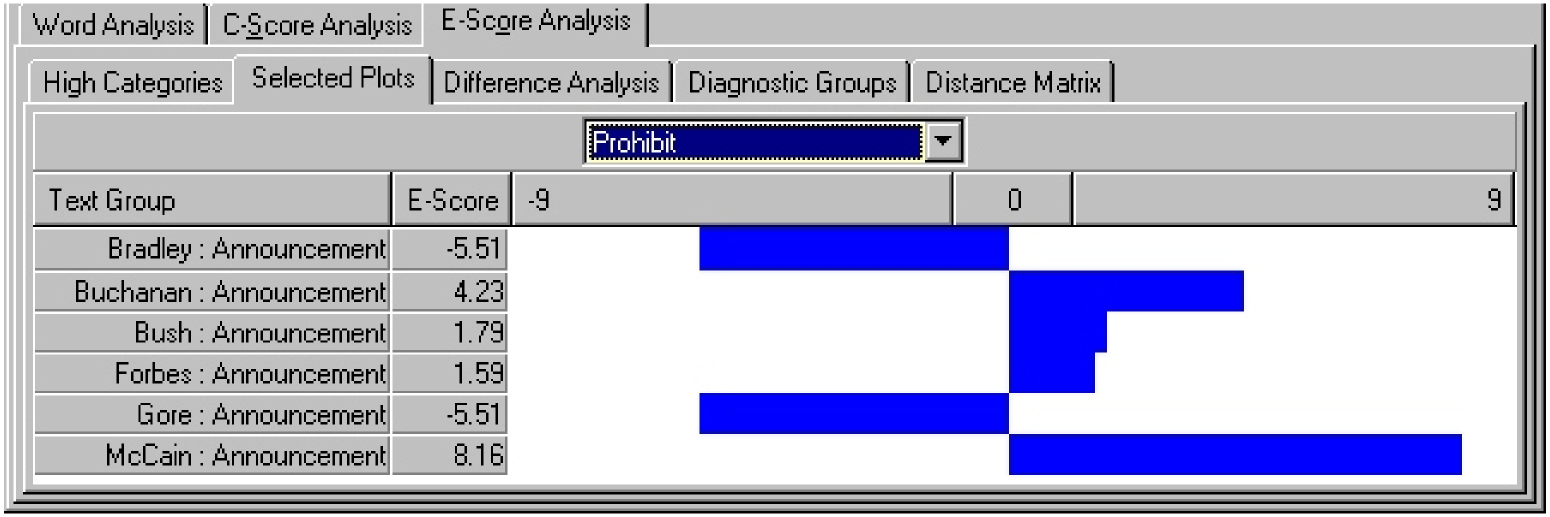
## High Categories



Pronouns (36, 38, 39, 40, 41, 42, 43)

| Cat No. | | Pct. | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|---|
| | PRONOUNS (N=7) | 14.80 | 2.44 | 1.09 | 1.82 | 1.42 | 0.92 | 2.11 |
| 36 | Object | 4.25 | 8.15 | 1.29 | 2.95 | -2.82 | -0.33 | 2.65 |
| 38 | You | 0.91 | -1.29 | -1.42 | 0.62 | 5.91 | -0.48 | -0.76 |
| 41 | They | 1.25 | -0.56 | 0.39 | 1.39 | 2.67 | -3.86 | 0.75 |
| 42 | We | 3.94 | 8.95 | 9.52 | 7.41 | 4.55 | 9.55 | 11.52 |

## Selected Plots



Prohibit

| Text Group | E-Score |
|---|---|
| Bradley : Announcement | -5.51 |
| Buchanan : Announcement | 4.23 |
| Bush : Announcement | 1.79 |
| Forbes : Announcement | 1.59 |
| Gore : Announcement | -5.51 |
| McCain : Announcement | 8.16 |

# Emphasis Score Observations (1)

- Pairwise distances among the six candidates are very similar, indicating equidistant emphases on different ideas
- Most prominent deviation from the norm exhibited by Bush in emphasis on failure (45.71), also strongly emphasized by McCain (38.60), Bradley (10.76), and Buchanan (11.34)
- Next most prominent was Gore's use of postive adjectives expression the idea of "good" (22.75), also emphasized by Bush (11.14). Bush also emphasized positive adjectives expressing "tenderness" (10.81). Forbes emphasized "happy" (8.05) and Bradley "good" (8.44)
- All candidates showed relative overemphasis on movement in space (up, down, back, close): Bush (11.50), Bradley (12.78), Gore (14.89), Buchanan (17.14), Forbes (8.30), McCain (8.32)

# Emphasis Score Observations (2)

- **McCain (11.42) and Bush (11.70) emphasis on traditional symbols “children” and “young”**
- **Gore (15.00) and Bush (14.46) emphasis on status words such as “weak”, “poor” and “underprivileged”**
- **Guidance verbs: “guide” (Gore - 9.99), “prohibit” (McCain - 8.16, Buchanan - 4.23), “submit” (McCain - 7.45)**
- **Deviance verbs: “deviant behavior” (Buchanan - 6.67, Gore - 6.04, McCain -5.32), “creative process” (Gore - 6.83, Bush - 7.76), “about changing” (McCain - 5.97)**

# Other Analyses

- **Cooccurrence Analysis: shows, for each category, which are the four most frequent categories that follow**
- **Multidimensional Scaling: context and emphasis scores are analyzed using KYST**
- **C-score and E-score output for further analysis, particularly with other data (see other papers available at web site for examples)**