School of Communication

CATA Options for the Coding of Open-Ended Survey Data

Paul Skalski

Cleveland State University

Background

- 1960s: Phillip Stone and colleagues create the General Inquirer, which pioneered using computers for the content analysis of text.
- Initially a mainframe program but…
- …now appears as a Java program on the Internet, for instant use, with 182 coding categories (dictionaries).

Background

- Skalski (2002) compiled a list of and information about all available computer assisted text analysis (CATA) programs.
- Initially: 20 programs identified, most for Windows (13), and most commercial (17).
- Updated list: Content Analysis Guidebook Online (30 programs, 24 currently active).

Which for open-ended survey data?

- Several functions to consider:
- 1. Frequency output
- 2. Alphabetical output
- 3. Multi-unit data file output
- 4. KWIC or concordance
- 5. Dictionary coding
- 6. Special analyses

Which for open-ended survey data?

- Several functions to consider:
- **1. Frequency output**
- 2. Alphabetical output
- **3. <u>Multi-unit data file output</u>**
- 4. KWIC or concordance
- **5. <u>Dictionary coding</u>**
- **6. Special analyses**

Important: Multi-unit data file output

- Displaying output in case-by-variable form, e.g.,

	Var 1	Var 2	Var 3	Var 4	Etc…
Case 1	3	4	0	0	
Case 2	2	0	2	1	
Etc….					

- Important for matching up computer coded open-ended data with closed-ended data.

Considerations: Dictionary Coding

- Issue of Measurement: Conceptualization → Operationalization
- Two options:
  - 1. Standard dictionaries
  - 2. Custom dictionaries
- Third option: Emergent coding (does not require a dictionary)

Option 1: Standard Dictionaries

- These are measures built into a program, or, trusting someone else's operationalization of your concepts of interest.
- Some seem pretty impressive, but…
- PROBLEM: Most CATA programs do not reveal the full dictionary information, and those that go beyond simple counts do not reveal the full algorithm.
- Raises validity questions.

Example Standard Dictionary: "Sports" from WordStat includes…

- Aerobics
- Baseball
- Boxing
- Bowling
- Skating
- Skiing
- Soccer
- Sport
- Swimming.

Option 2: Custom Dictionaries

- This complicated process involves coming up with your own operationalization of concepts of interest.
- More control and confidence perhaps, but…
- PROBLEMS:
  - Disambiguation (e.g., “fine”)
  - Negation (e.g., “I am not patriotic”)
  - Colloquial speech
- These also apply to standard dictionaries.

Option 2: Custom Dictionaries

- Steps/advice for valid dictionary creation:
- STEP ONE—identify all words consistent with conceptual definitions of each concept.
- STEP TWO—Use dictionary building tool:
  - E.g., Wordstat is programmed to allow the addition to a base dictionary of antonyms, synonyms, similar terms, hypernyms, hyponyms, holonyms, and other word classes.
  - Also allows "wild card" specifications of root words (e.g., including "pleasur*")
  - Also need to use exclude list.

Option 3: Emergent Coding

- This is when dimensions or patterns of text are derived from the data at hand (i.e., the texts under investigation), without any pre-set dictionaries.
- Easy, but…
- PROBLEM: Less accepted due to deviation from positivist, a priori assumptions of C.A.
- May still be useful in early stages of study.
- Also can get better technologically (AI).

Which for open-ended survey data?

- Here are example analyses by three of my favorite CATA programs, one for each type of coding previously discussed:
  - 1. CATPAC (emergent coding)
  - 2. Diction (standard dictionary)
  - 3. WordStat (custom dictionary)
- Note: Diction and Wordstat can do both types of dictionary coding.

School of Communication

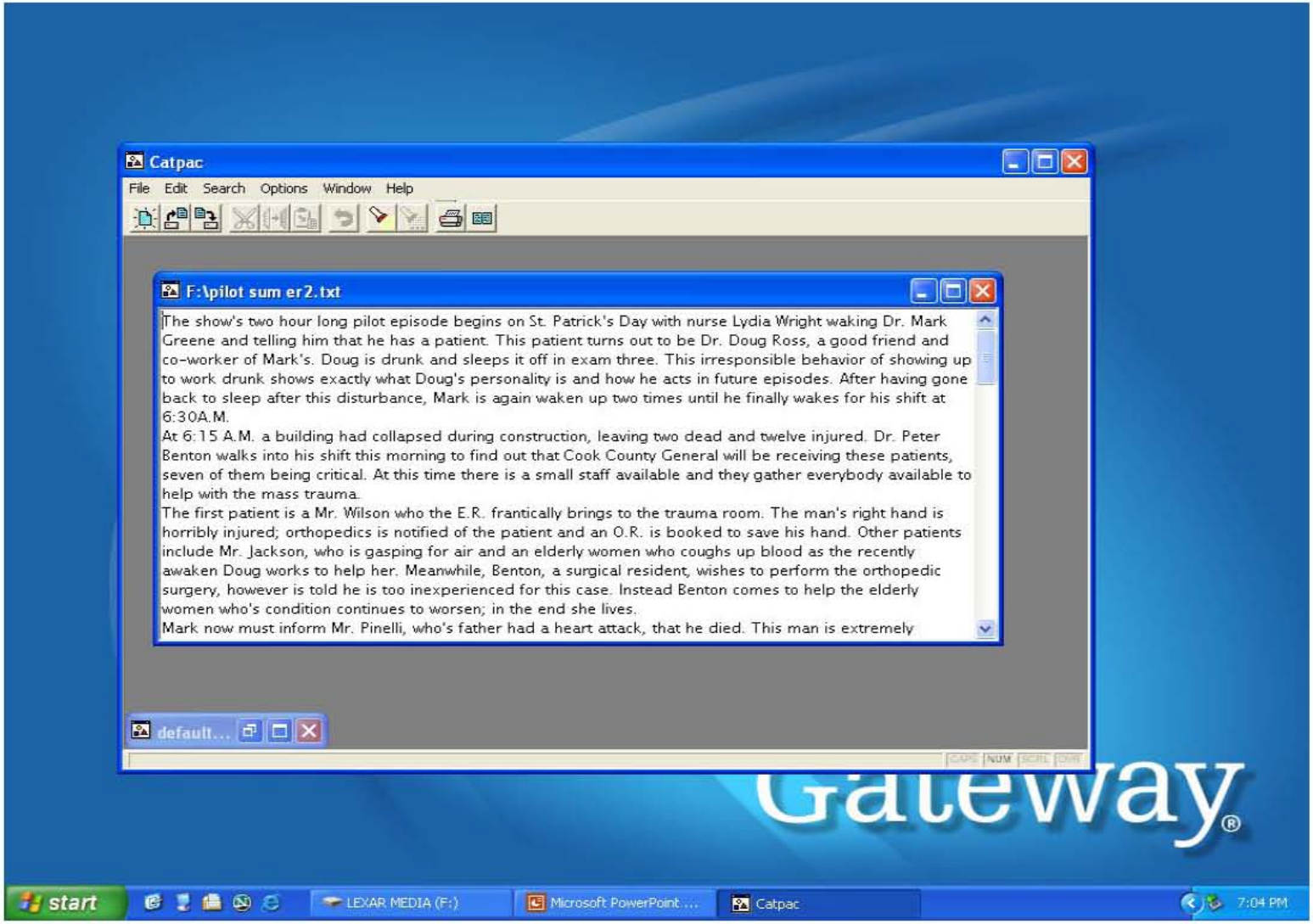
CATPAC

Cleveland State University College of Liberal Arts and Social Sciences

About CATPAC

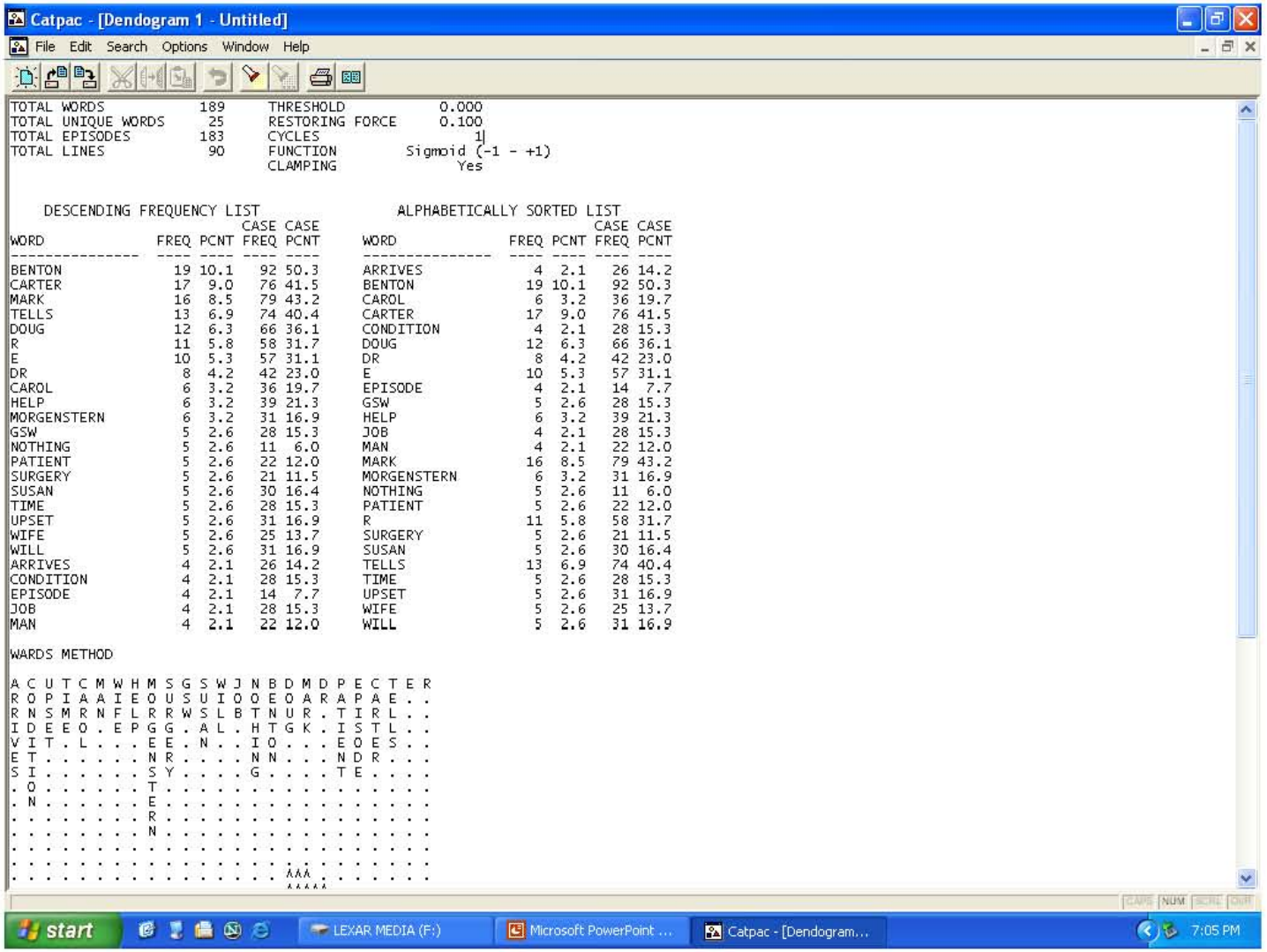
- Created by Joseph Woelfel (University of Buffalo)
- Part of the GALILEO suite of software that analyze and display various types of networks
- CATPAC uses a neural network approach, identifying the most frequent words and determining patterns of connection based on co-occurrence
- A scanning window is used to measure the association/co-occurrence
- Uses cluster analysis to present results of this co-occurrence procedure

The CATPAC Interface



- Text input will appear in CATPAC main screen

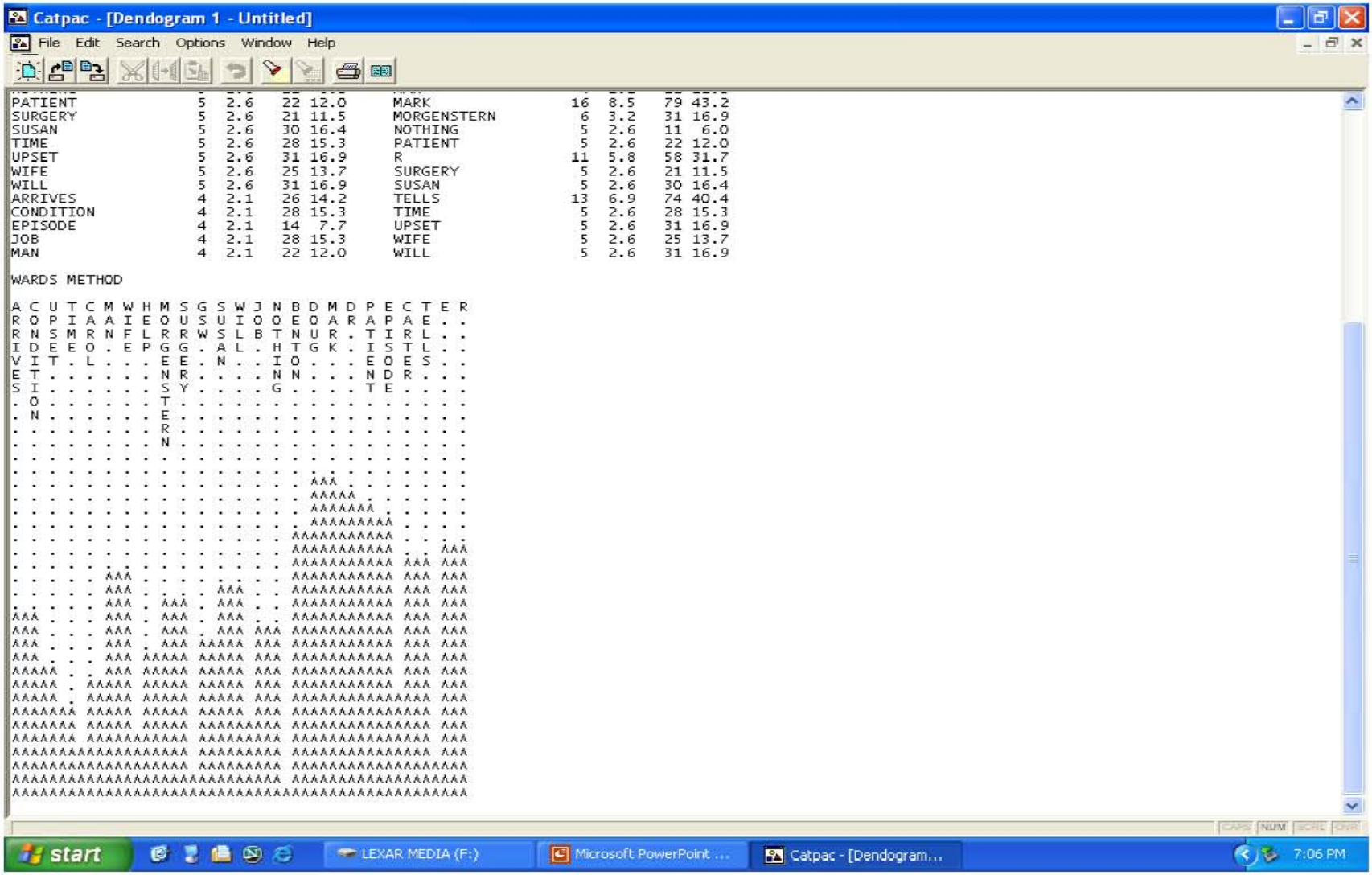
CATPAC Output: Descending Frequency List, Alphabetically Sorted List



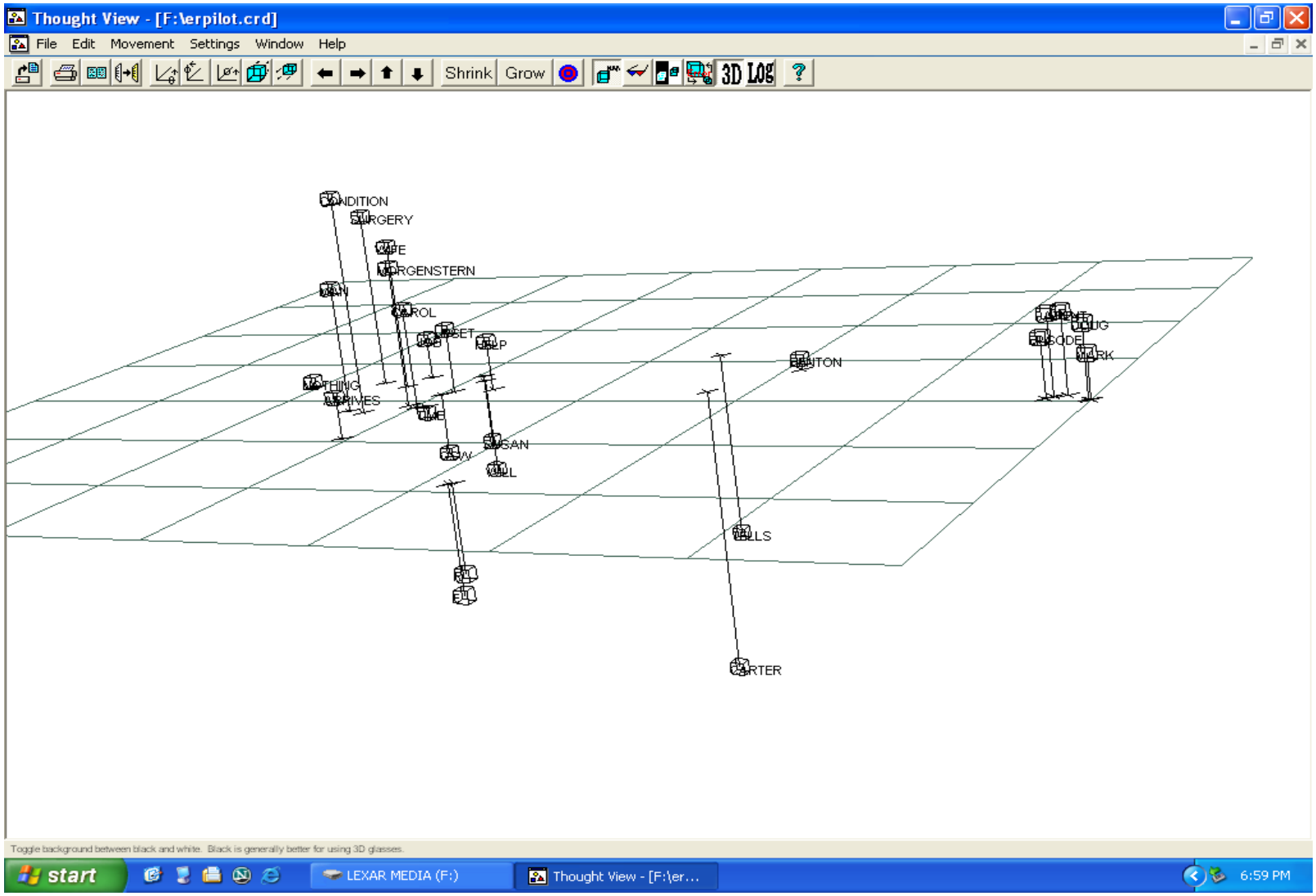
WORD	FREQ	PCNT	CASE FREQ	CASE PCNT
BENTON	19	10.1	92	50.3
CARTER	17	9.0	76	41.5
MARK	16	8.5	79	43.2
TELLS	13	6.9	74	40.4
DOUG	12	6.3	66	36.1
R	11	5.8	58	31.7
E	10	5.3	57	31.1
DR	8	4.2	42	23.0
CAROL	6	3.2	36	19.7
HELP	6	3.2	39	21.3
MORGENSTERN	6	3.2	31	16.9
GSW	5	2.6	28	15.3
NOTHING	5	2.6	11	6.0
PATIENT	5	2.6	22	12.0
SURGERY	5	2.6	21	11.5
SUSAN	5	2.6	30	16.4
TIME	5	2.6	28	15.3
UPSET	5	2.6	31	16.9
WIFE	5	2.6	25	13.7
WILL	5	2.6	31	16.9
ARRIVES	4	2.1	26	14.2
CONDITION	4	2.1	28	15.3
EPISODE	4	2.1	14	7.7
JOB	4	2.1	28	15.3
MAN	4	2.1	22	12.0

WORD	FREQ	PCNT	CASE FREQ	CASE PCNT
ARRIVES	4	2.1	26	14.2
BENTON	19	10.1	92	50.3
CAROL	6	3.2	36	19.7
CARTER	17	9.0	76	41.5
CONDITION	4	2.1	28	15.3
DOUG	12	6.3	66	36.1
DR	8	4.2	42	23.0
E	10	5.3	57	31.1
EPISODE	4	2.1	14	7.7
GSW	5	2.6	28	15.3
HELP	6	3.2	39	21.3
JOB	4	2.1	28	15.3
MAN	4	2.1	22	12.0
MARK	16	8.5	79	43.2
MORGENSTERN	6	3.2	31	16.9
NOTHING	5	2.6	11	6.0
PATIENT	5	2.6	22	12.0
R	11	5.8	58	31.7
SURGERY	5	2.6	21	11.5
SUSAN	5	2.6	30	16.4
TELLS	13	6.9	74	40.4
TIME	5	2.6	28	15.3
UPSET	5	2.6	31	16.9
WIFE	5	2.6	25	13.7
WILL	5	2.6	31	16.9

CATPAC Output: Dendogram



CATPAC Output: 3D Plot (using ThoughtView, another part of Galileo Suite)



Sample Findings

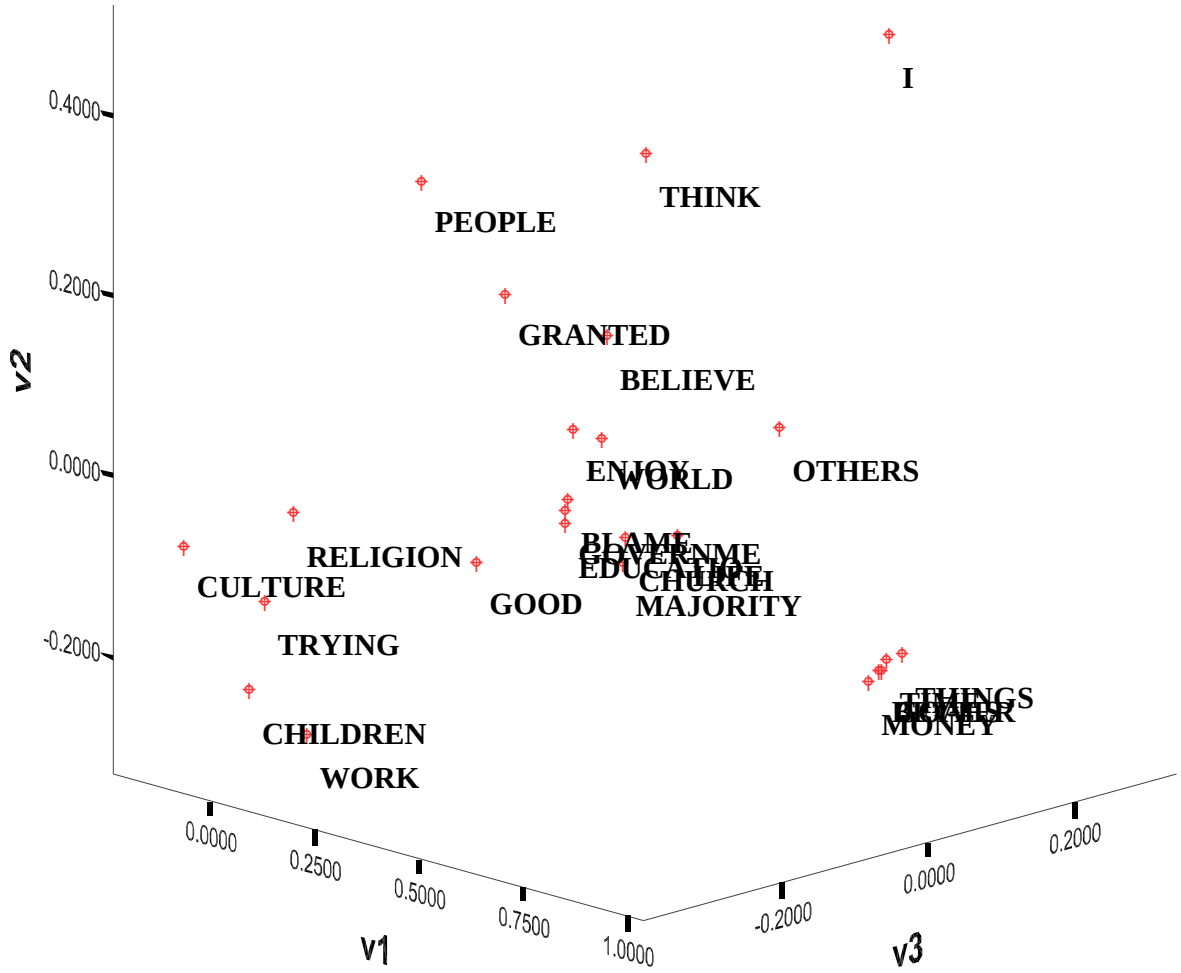
• Neuendorf and Skalski (2008) analyzed open ended survey responses to question asking:

•Describe yourself as a typical American.

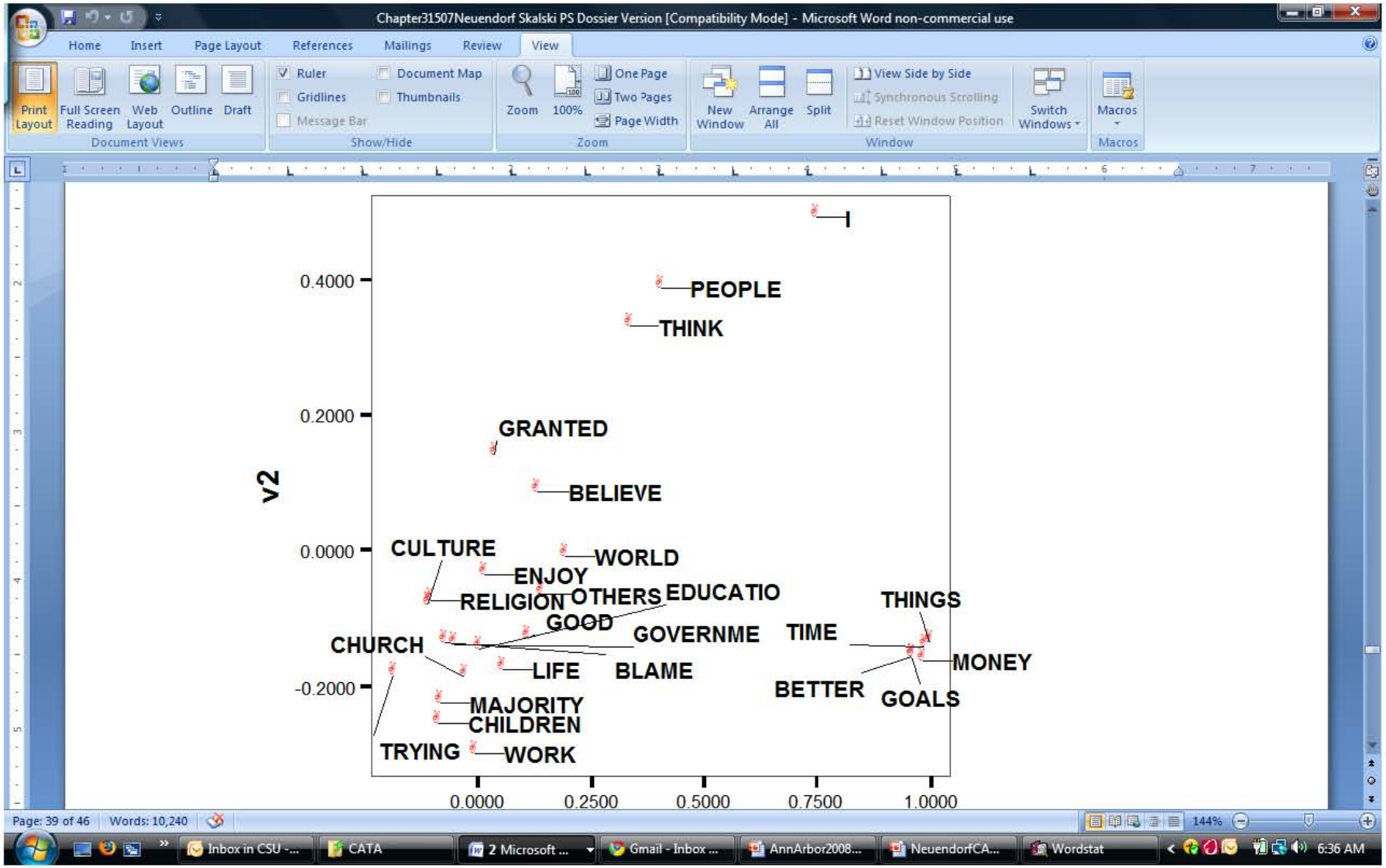
•All responses entered as one case delimited file.

•Results showed clustering of practical vs. other considerations, as in following:

CATPAC Output: 3D Plot (using SPSS)



CATPAC Output: 2D Plot (using SPSS)



School of Communication

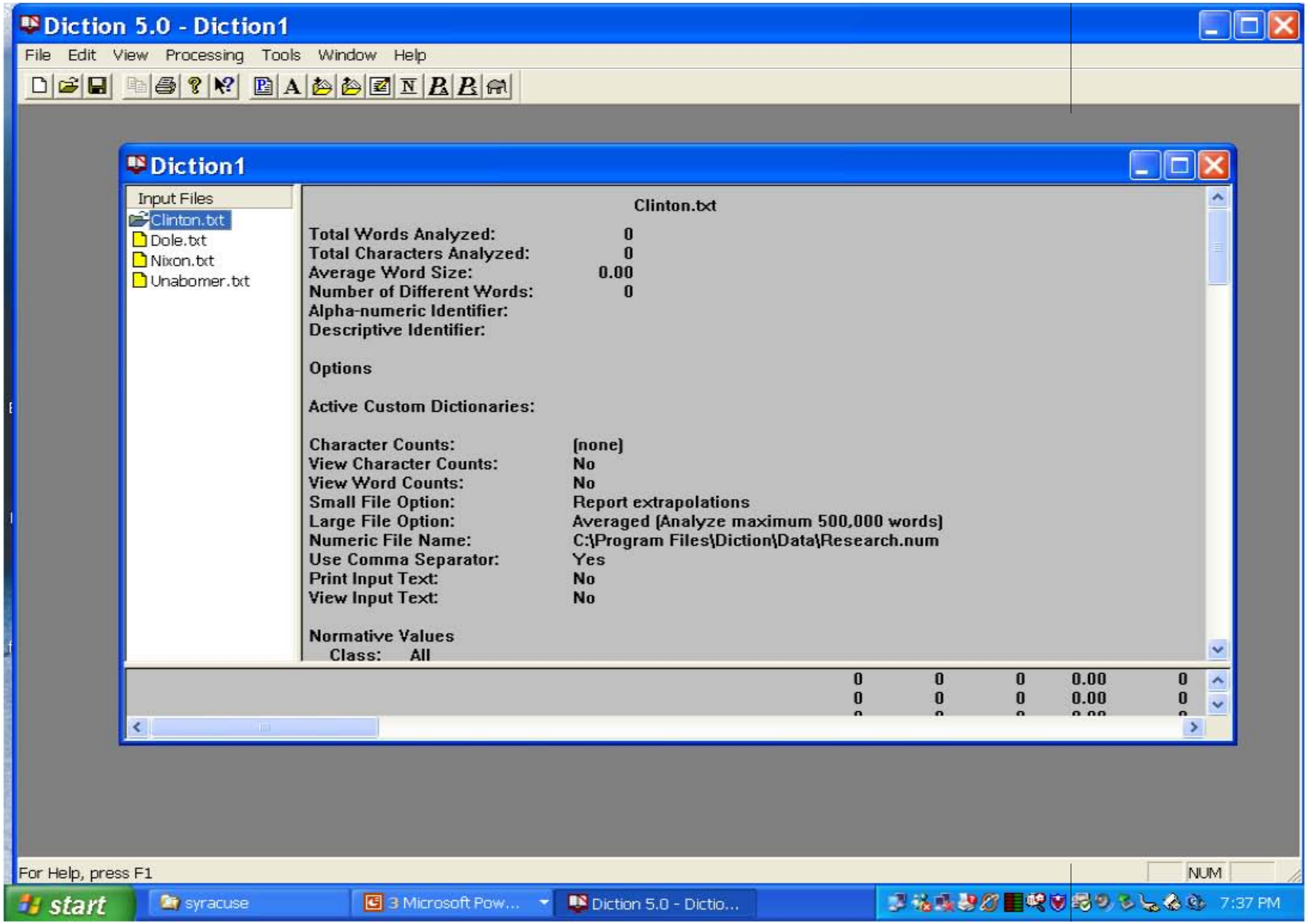
DICTION

Cleveland State University College of Liberal Arts and Social Sciences

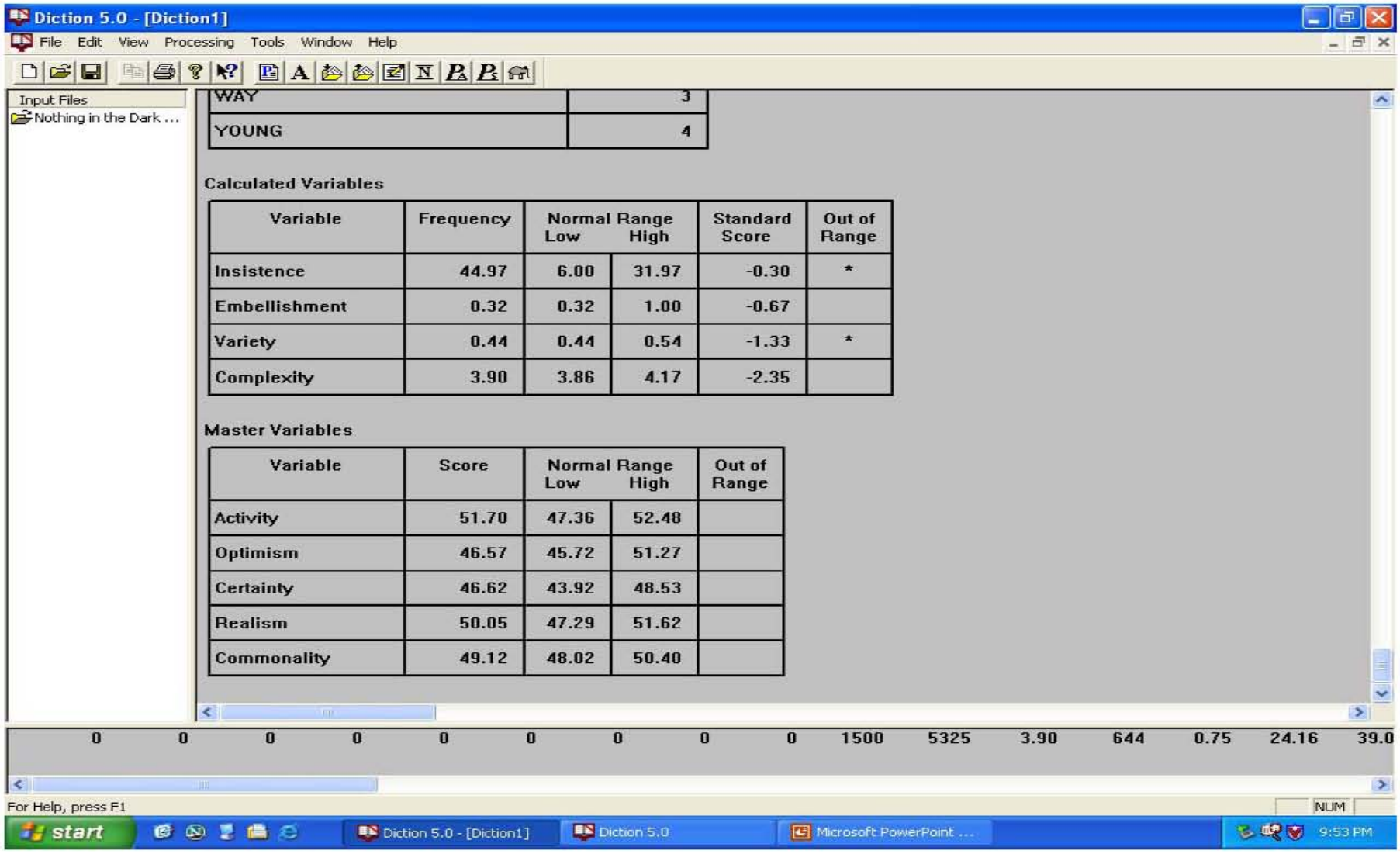
About Diction

- Created by Roderick P. Hart (University of Texas) originally for the purpose of analyzing political discourse
- To measure “semantic features”, uses a series of 31 standard dictionaries and five “Master Variables” (scales constituted of combinations of the 31):
  - Activity
  - Optimism
  - Certainty
  - Realism
  - Commonality
- Users can create custom dictionaries in addition to standard dictionaries.
- The program can accept individual or multiple passages.

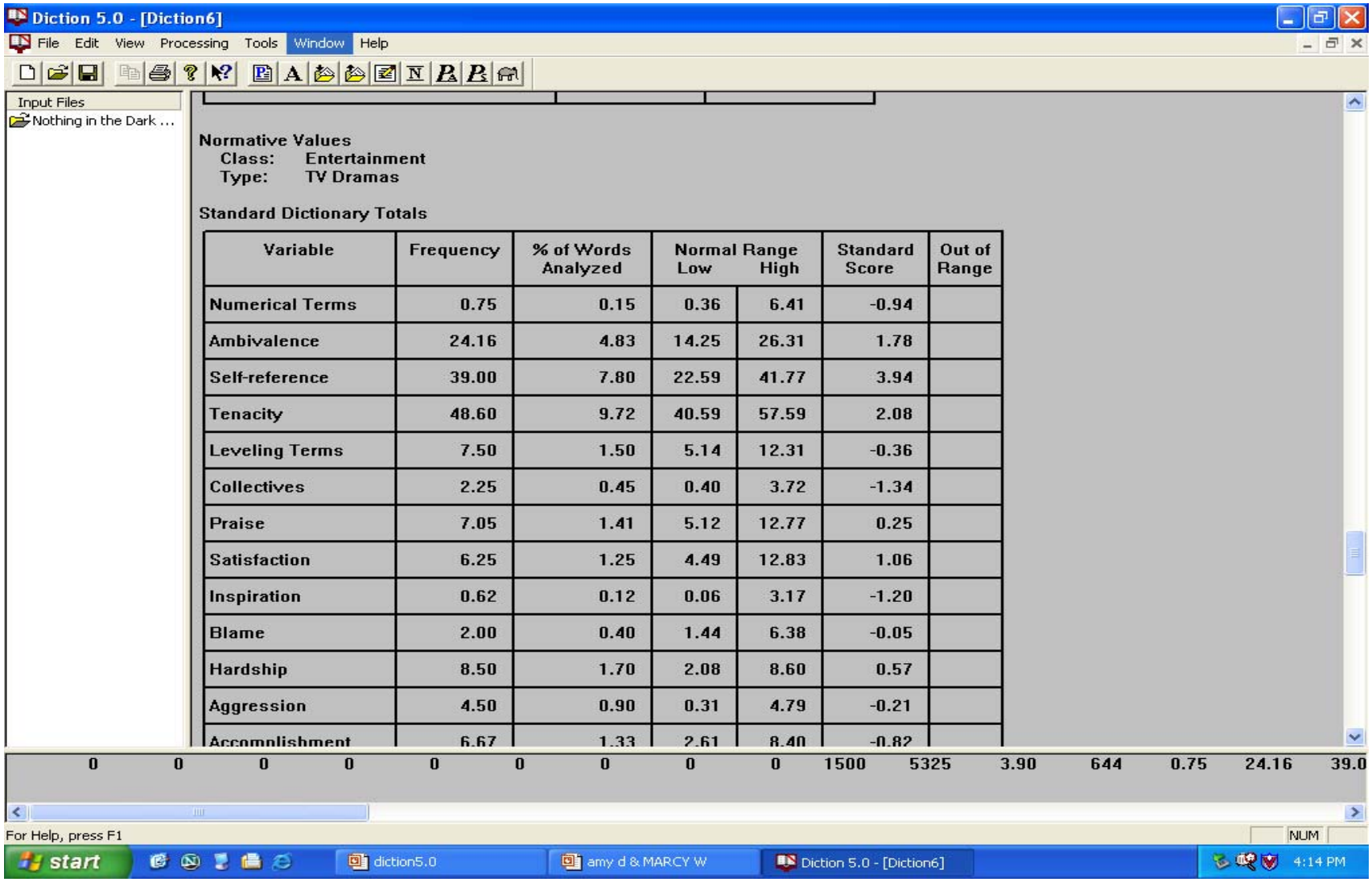
The Diction Interface



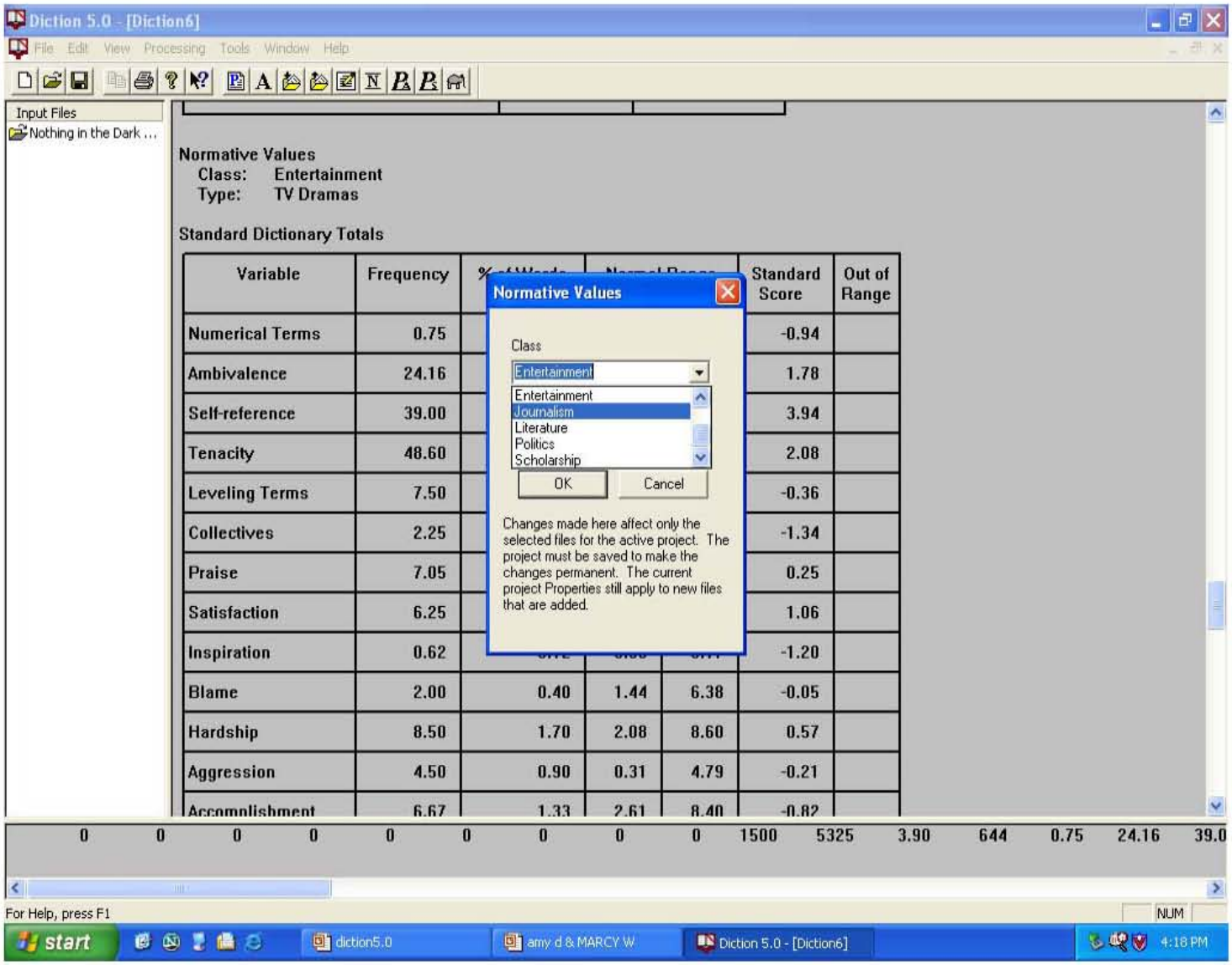
Diction Output: Calculated & Master Variables



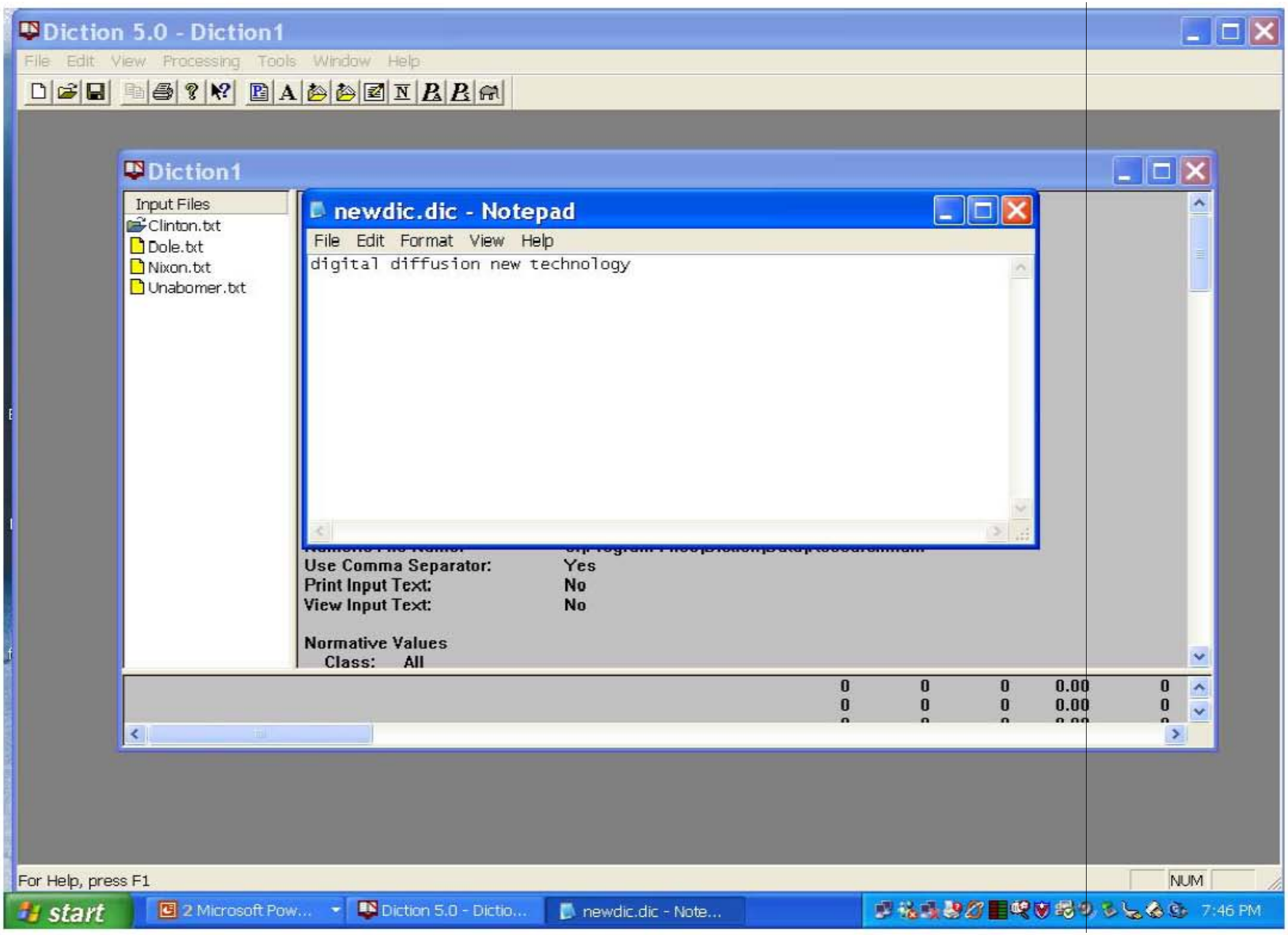
Diction Output: Dictionary Totals with Normative Values



Diction Output: Interactively Changing Normative Values

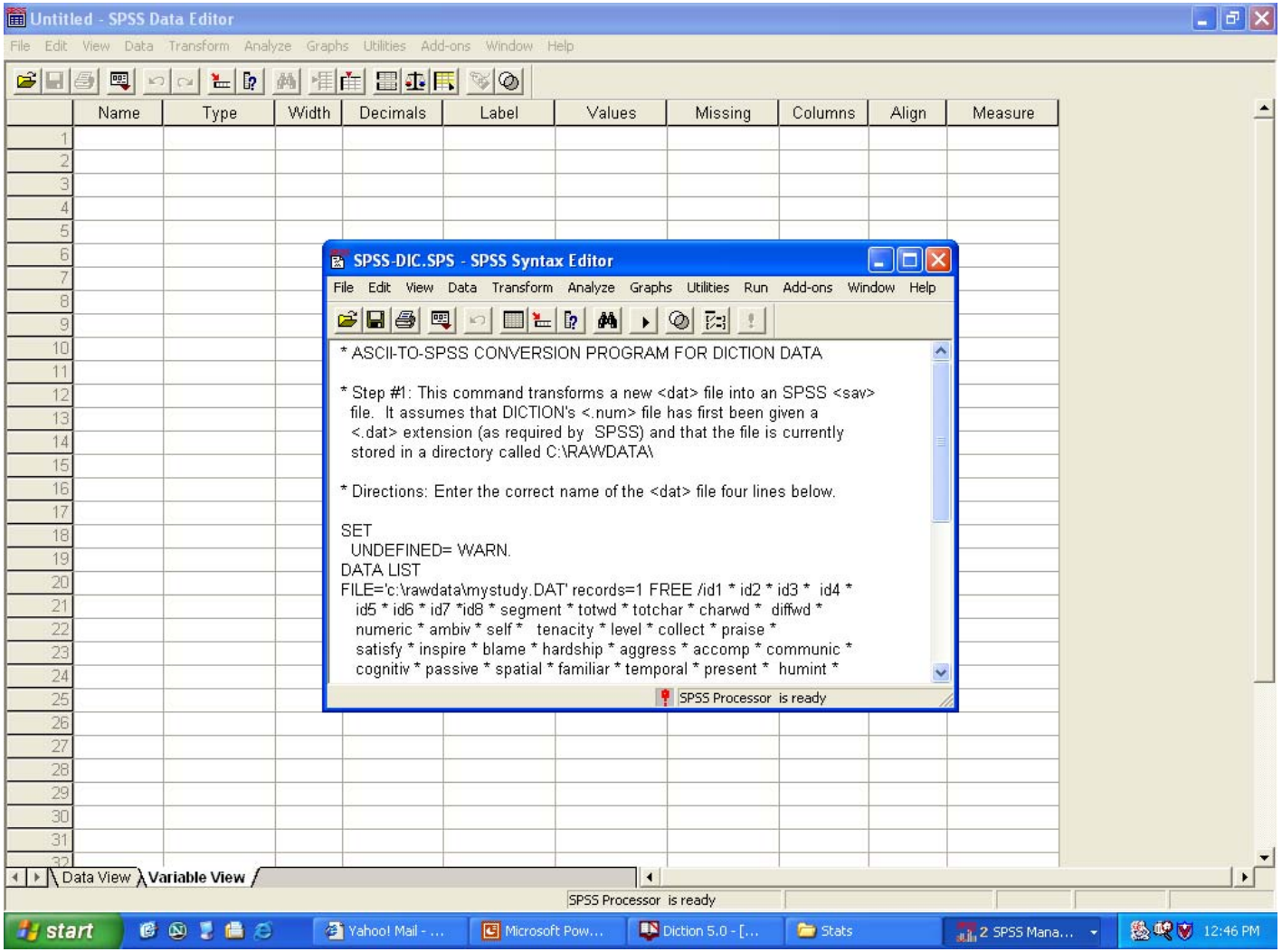


Diction: Custom Dictionaries as Simple .txt Files



Diction Output: Data file may be exported to SPSS

SPSS Syntax Editor



Sample Findings

• Neuendorf and Skalski (2008) analyzed the first State of the Union address by each of the nine last Presidents (Eisenhower to Bush).

•Texts from online archive; had to be imported as separate files.

•Results of Bush's first State of the Union address vs. other political speeches follows:

Diction Output: Master Dictionary Scores and Comparisons

Variable	Frequency	% of Words Analyzed	Normal Range Low	Normal Range High	Standard Score
Numerical Terms	6.42	1.28	0.3	15.04	-0.17
Ambivalence	7.57	1.51	6.49	19.21	-0.83
Self-reference	7.8	1.56	0	15.1	0.1
Tenacity	23.69	4.74	23.32	39.76	-0.95
Leveling Terms	7.69	1.54	5.02	12.76	-0.31
Collectives	13.57	2.71	4.04	14.46	0.83
Praise	8.99	1.8	2.77	9.59	0.83
Satisfaction	16.67	3.33	0.47	6.09	4.77
Inspiration	12.01	2.4	1.56	11.1	1.19
Blame	1.59	0.32	0.06	4.16	-0.25
Hardship	9	1.8	1.26	10.48	0.68
Aggression	8.59	1.72	1.07	9.79	0.73
Accomplishment	18.04	3.61	4.96	23.78	0.39
Communication	5.79	1.16	2.21	11.79	-0.25
Cognition	7.01	1.4	4.43	14.27	-0.48
Passivity	6.78	1.36	2.1	8.08	0.56
Spatial Terms	14.49	2.9	4.17	19.85	0.32
Familiarity	111.34	22.27	117.87	147.19	-1.45
Temporal Terms	18.45	3.69	8.36	21.82	0.5
Present Concern	9.33	1.87	7.02	16.6	-0.52
Human Interest	31.3	6.26	18.13	45.49	-0.04
Concreteness	27.52	5.5	10.7	28.5	0.89
Past Concern	1.8	0.36	0.97	6.19	-0.68
Centrality	3.17	0.63	1.19	7.54	-0.37
Rapport	2.84	0.57	0.42	4.26	0.26
Cooperation	8.73	1.75	0.36	8.44	1.07
Diversity	1.69	0.34	0.07	3.81	-0.14
Exclusion	0.72	0.14	0	4.31	-0.65
Liberation	3.62	0.72	0	4.72	0.57
Denial	3.09	0.62	2.57	10.35	-0.87
Motion	2.27	0.45	0.17	4.35	0

Diction Output: Master Dictionary Scores and Comparisons

Variable	Score	Normal Range		Out of Range
		Low	High	
Activity	50.1	46.74	55.48	
Optimism	57.22	46.37	52.25	*
Certainty	49.96	46.9	51.96	
Realism	49.75	46.1	52.62	
Commonality	51.18	46.86	52.28	

School of Communication

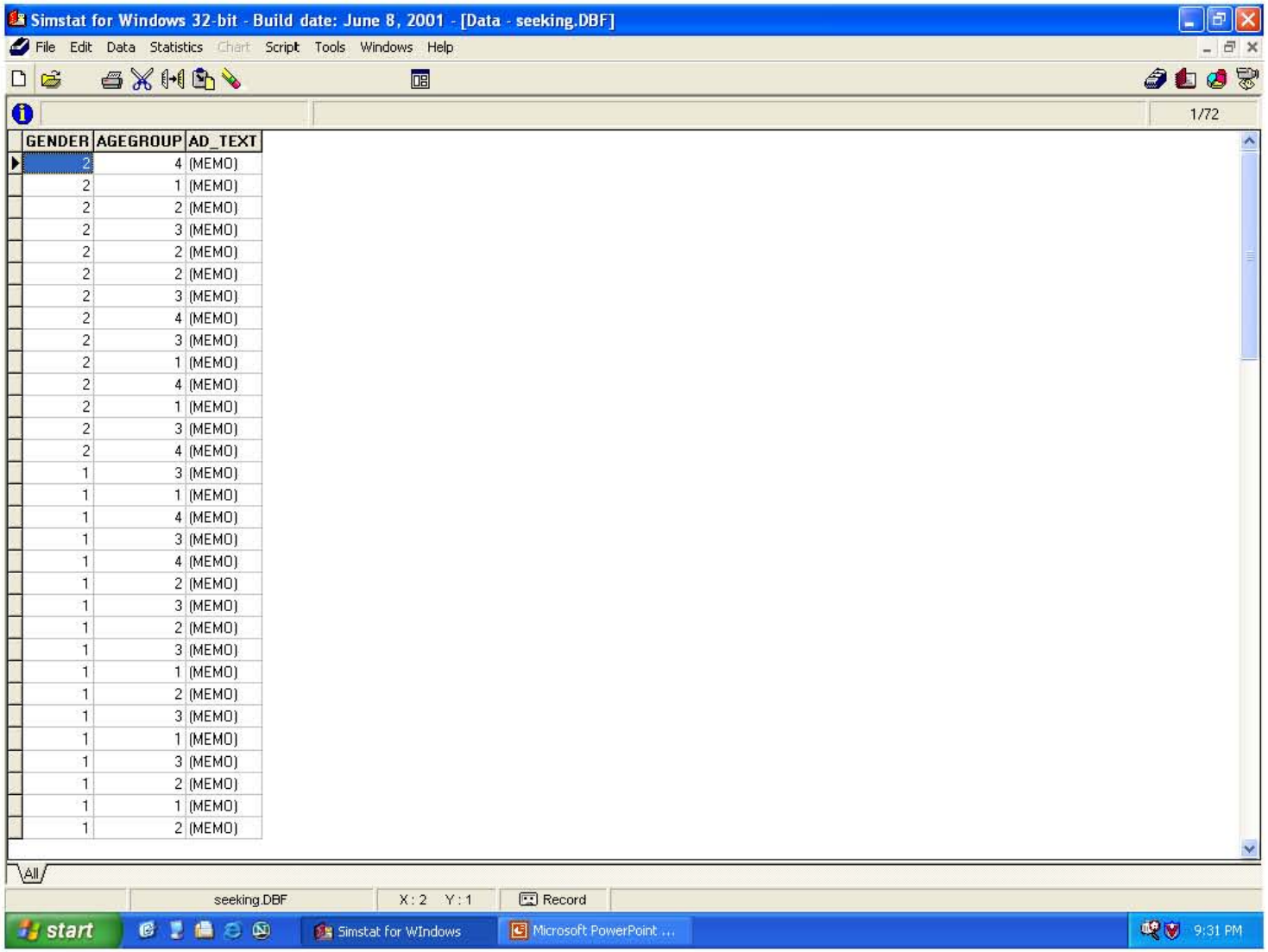
WORDSTAT

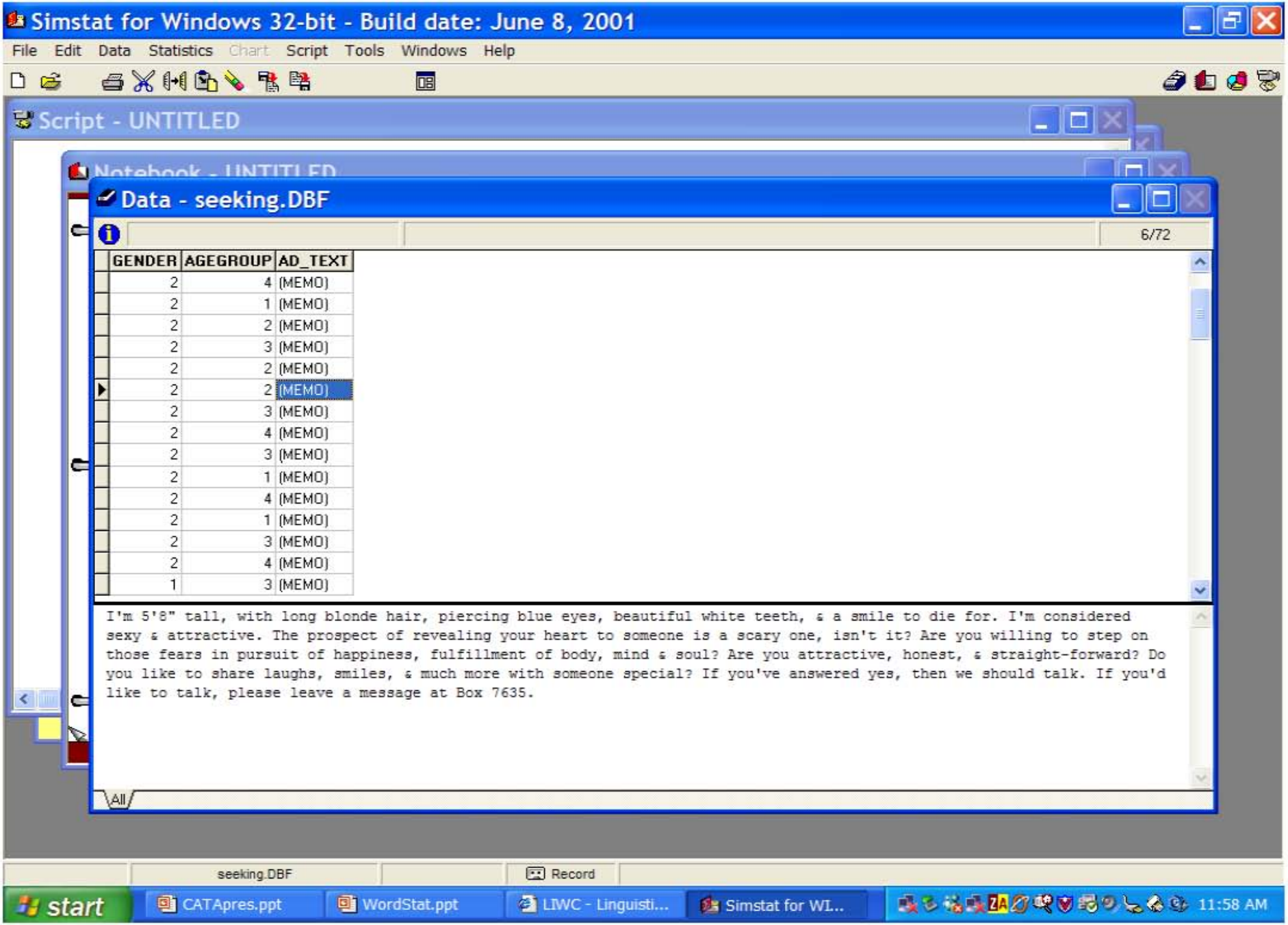
Cleveland State University College of Liberal Arts and Social Sciences

About WordStat

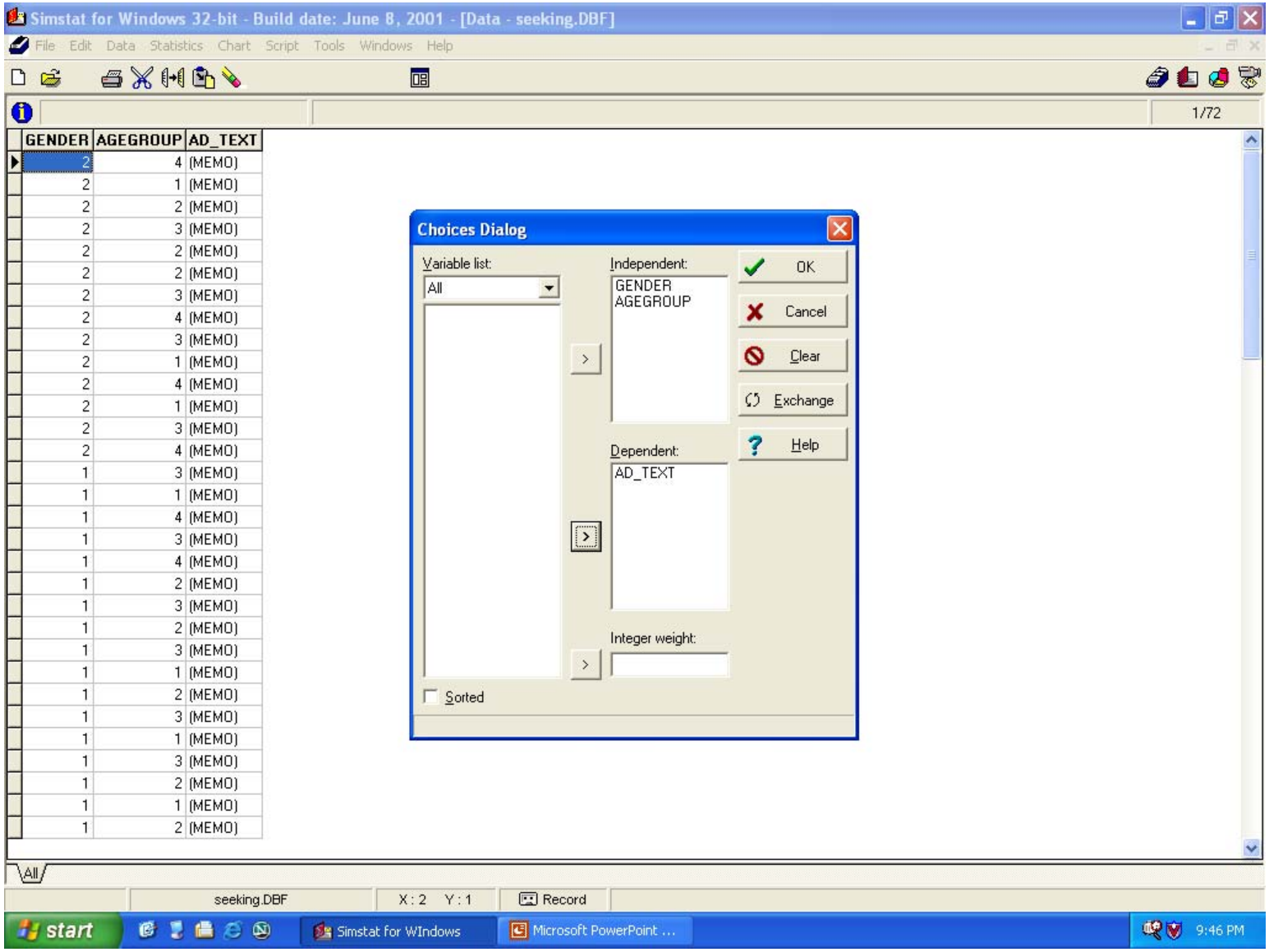
- Created by Normand Peladeau, as part of the SimStat suite for quantitative data analysis (a counterpart to SPSS)
- Must be run as part of SimStat
- Particularly suited to analyzing open-ended responses, in that the data set may include both numeric and textual variables—which can immediately be crosstabulated
- The "standard" dictionaries that are included are incomplete and should be avoided
- Also includes KWIC

The WordStat Interface (within SimStat)

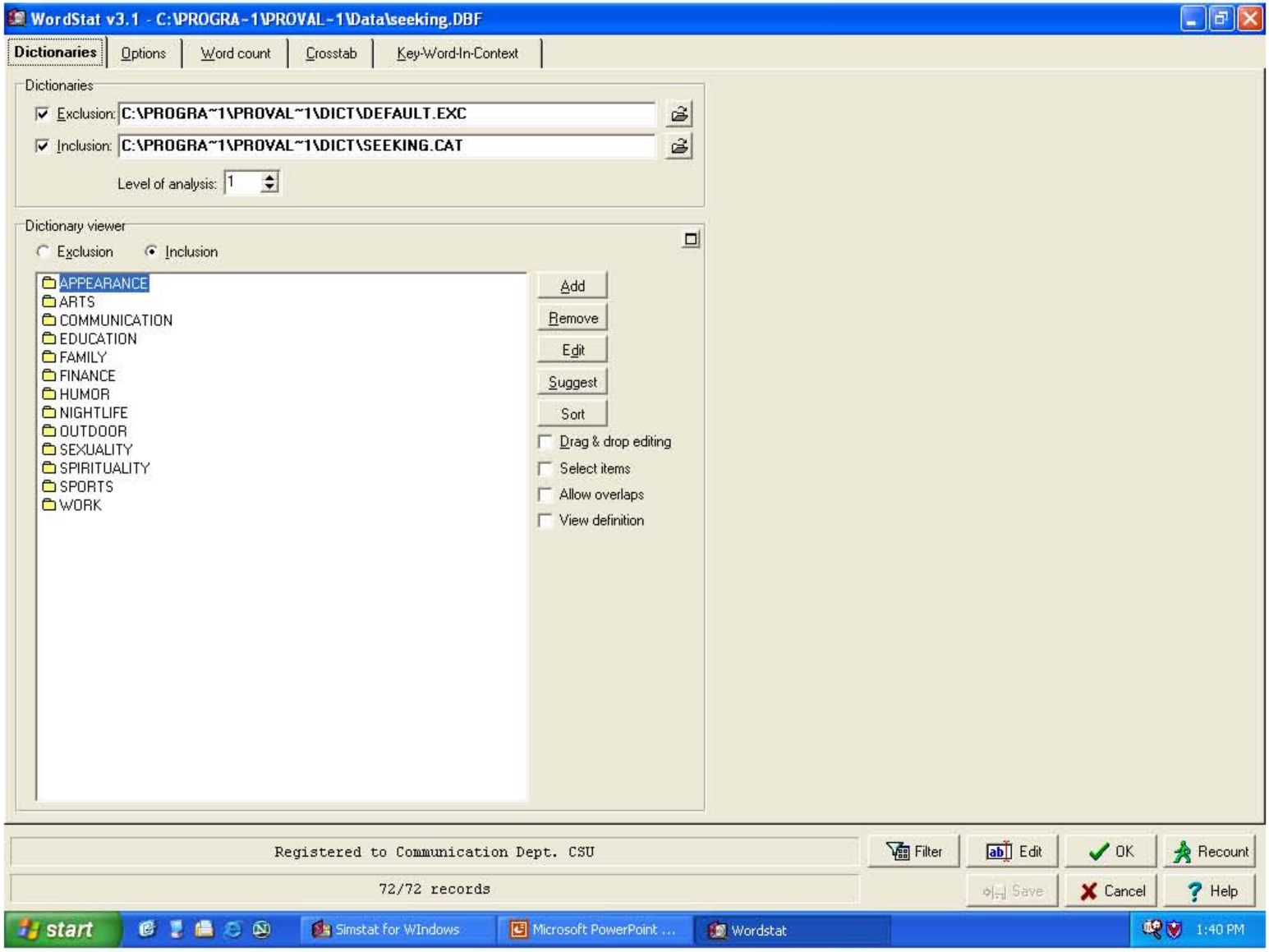



Simstat for Windows 32-bit - Build date: June 8, 2001
File Edit Data Statistics Chart Script Tools Windows Help
Script - UNTITLED
Notebook - UNTITLED
Data - seeking.DBF
6/72
GENDER AGEGROUP AD_TEXT
2 4 (MEMO)
2 1 (MEMO)
2 2 (MEMO)
2 3 (MEMO)
2 2 (MEMO)
2 2 (MEMO)
2 3 (MEMO)
2 4 (MEMO)
2 3 (MEMO)
2 1 (MEMO)
2 4 (MEMO)
2 1 (MEMO)
2 3 (MEMO)
2 4 (MEMO)
1 3 (MEMO)
I'm 5'8" tall, with long blonde hair, piercing blue eyes, beautiful white teeth, & a smile to die for. I'm considered sexy & attractive. The prospect of revealing your heart to someone is a scary one, isn't it? Are you willing to step on those fears in pursuit of happiness, fulfillment of body, mind & soul? Are you attractive, honest, & straight-forward? Do you like to share laughs, smiles, & much more with someone special? If you've answered yes, then we should talk. If you'd like to talk, please leave a message at Box 7635.
All
seeking.DBF
Record
start
CATApres.ppt
WordStat.ppt
LIWC - Linguisti...
Simstat for WI...
11:58 AM

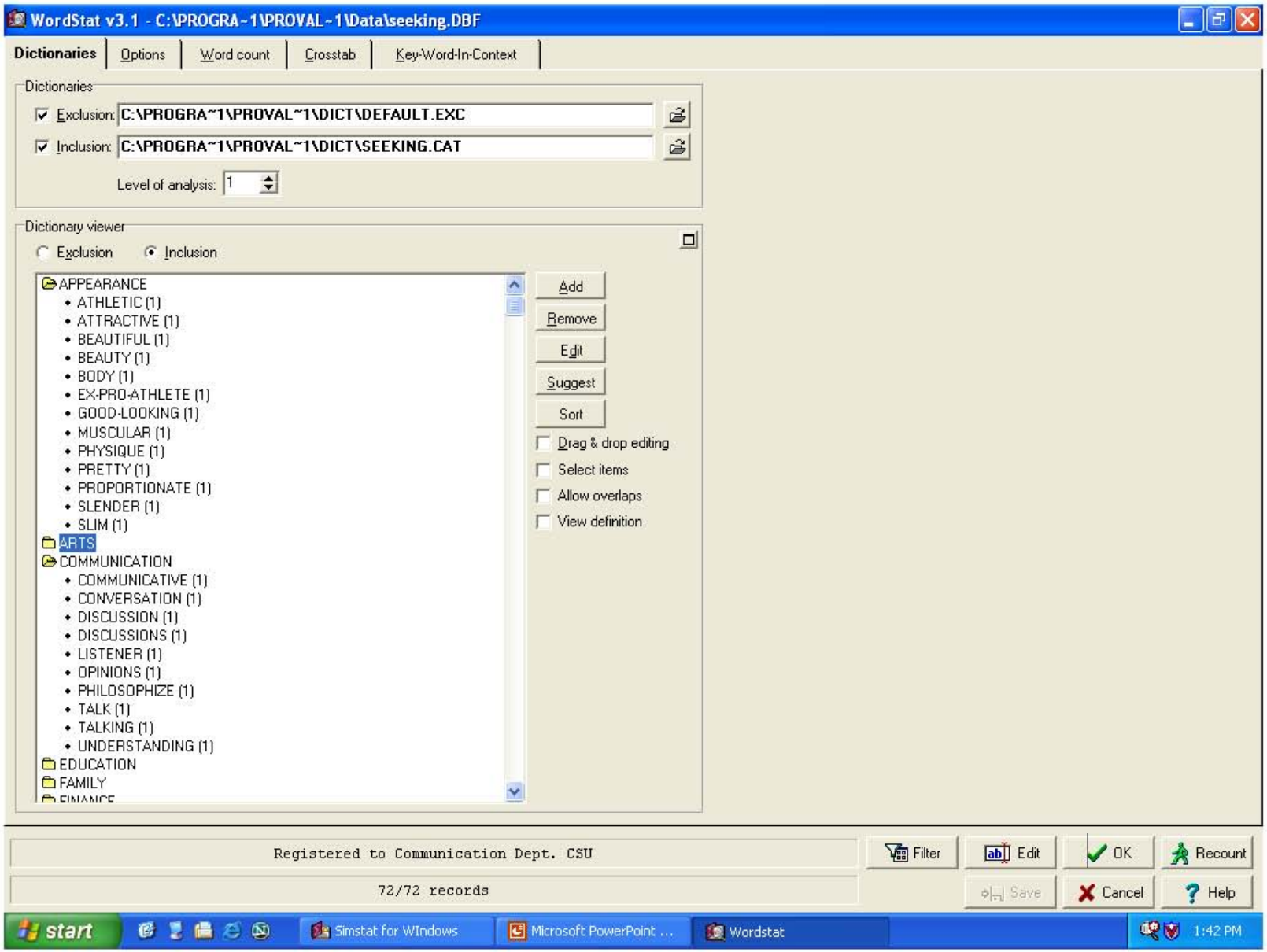
Selection of Independent & Dependent Variables—Including Textual Variable



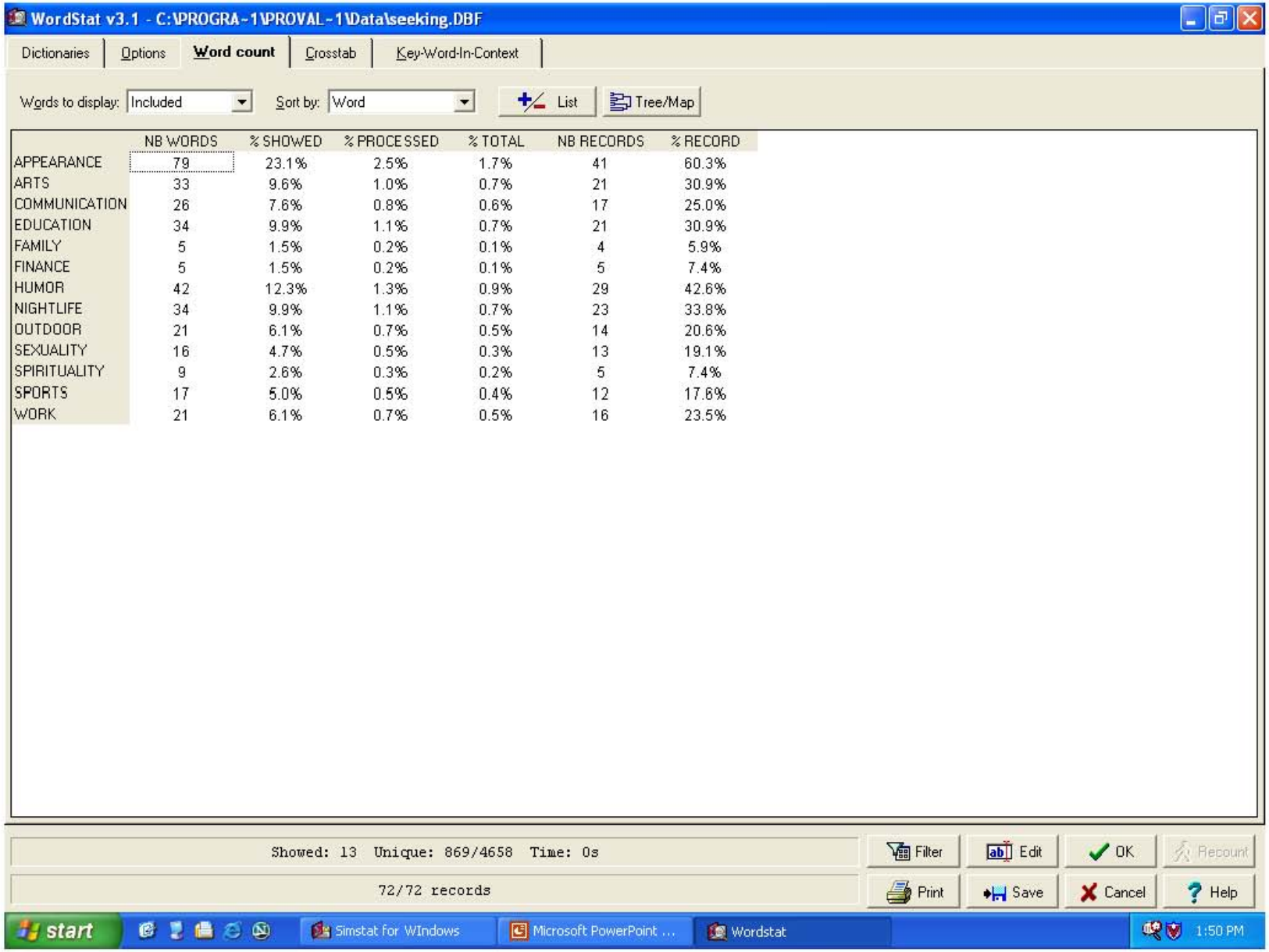
Standard WordStat “Dictionaries”



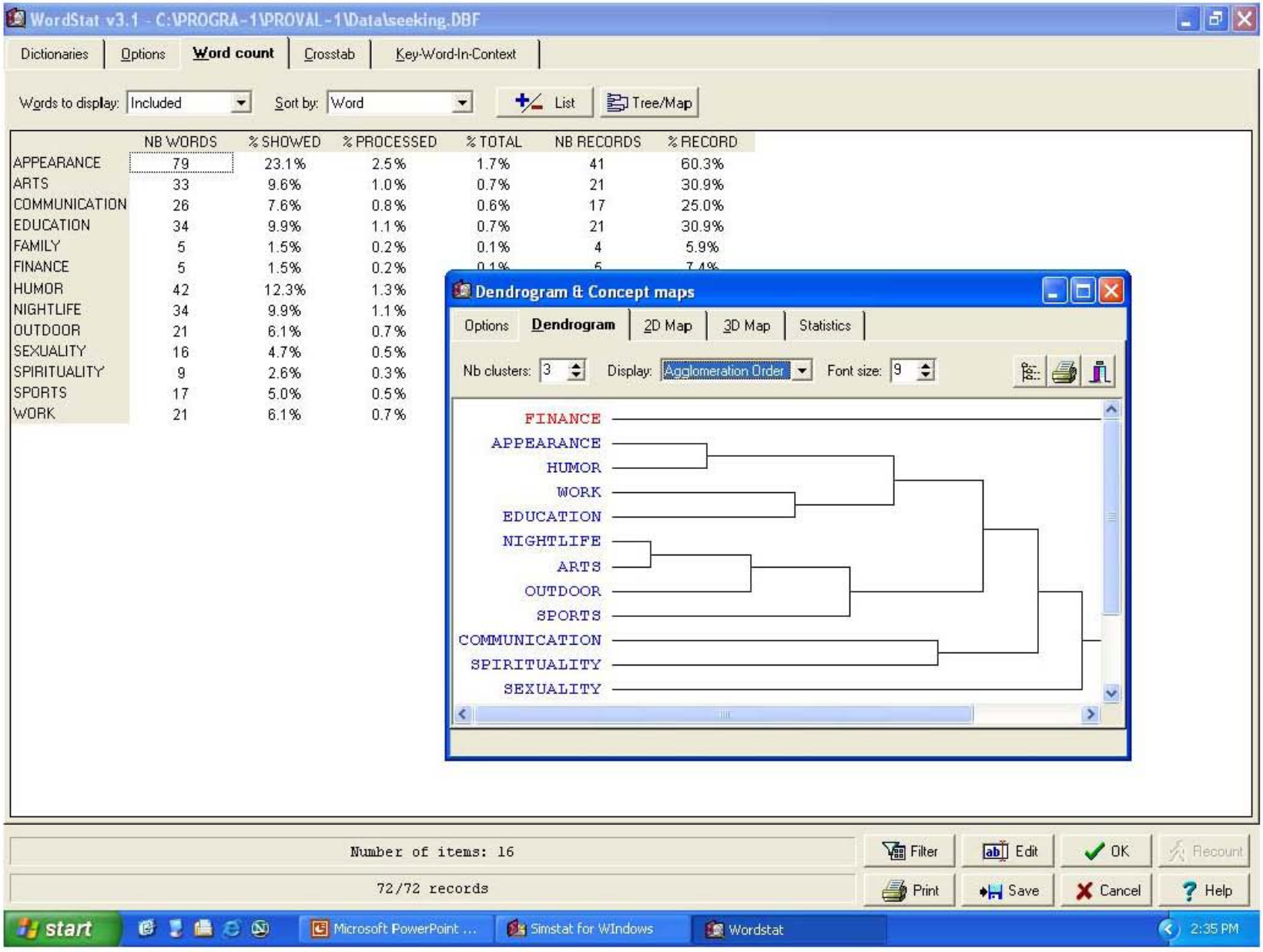
Breakdown of very limited WordStat “Dictionary”



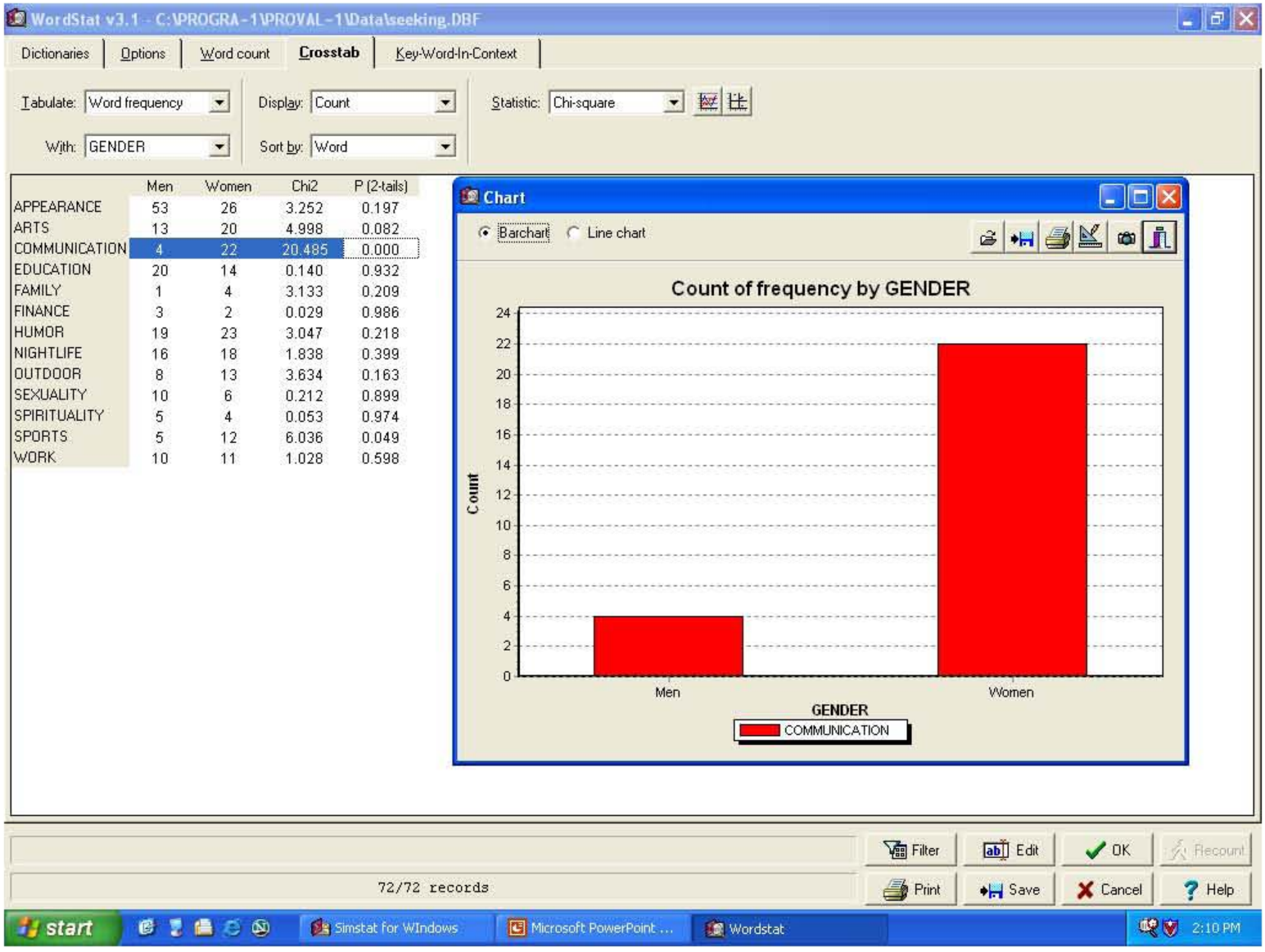
WordStat Output: Word counts



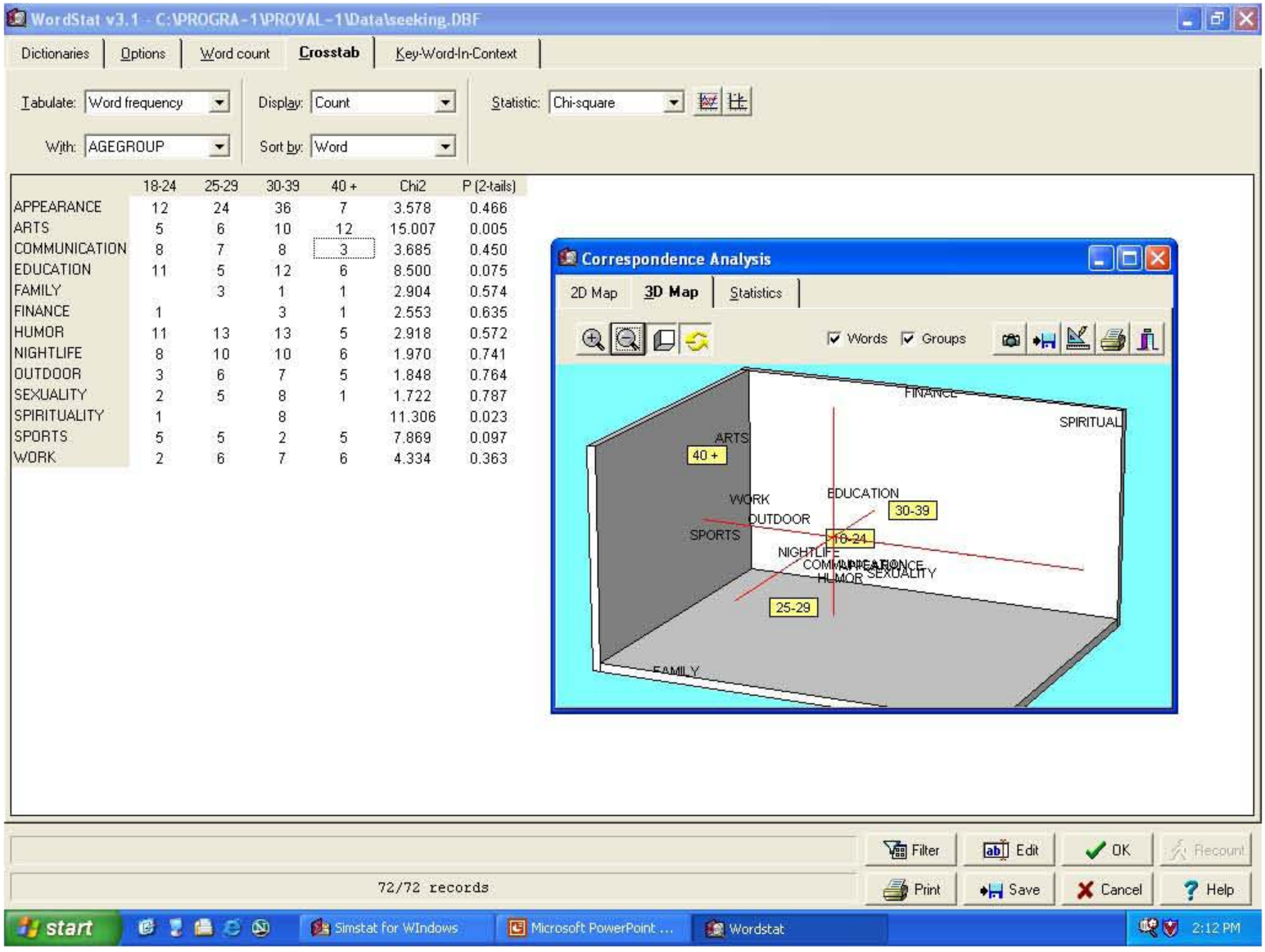
WordStat Output: Dendogram



WordStat Output: Crosstab with bar graph

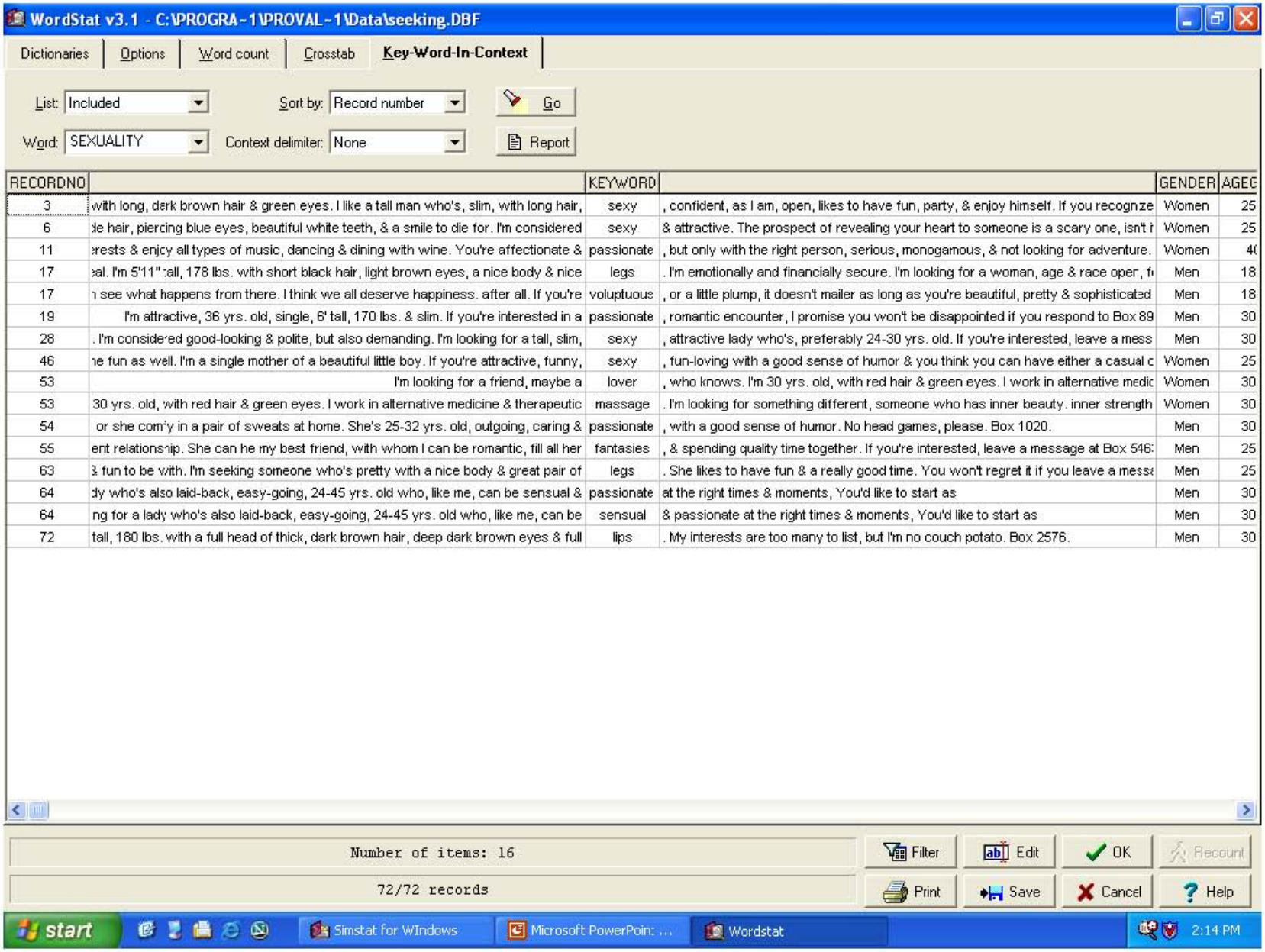


WordStat Output: Crosstab and 3D representation



	18-24	25-29	30-39	40 +	Chi2	P (2-tails)
APPEARANCE	12	24	36	7	3.578	0.466
ARTS	5	6	10	12	15.007	0.005
COMMUNICATION	8	7	8	3	3.685	0.450
EDUCATION	11	5	12	6	8.500	0.075
FAMILY		3	1	1	2.904	0.574
FINANCE	1		3	1	2.553	0.635
HUMOR	11	13	13	5	2.918	0.572
NIGHTLIFE	8	10	10	6	1.970	0.741
OUTDOOR	3	6	7	5	1.848	0.764
SEXUALITY	2	5	8	1	1.722	0.787
SPIRITUALITY	1		8		11.306	0.023
SPORTS	5	5	2	5	7.869	0.097
WORK	2	6	7	6	4.334	0.363

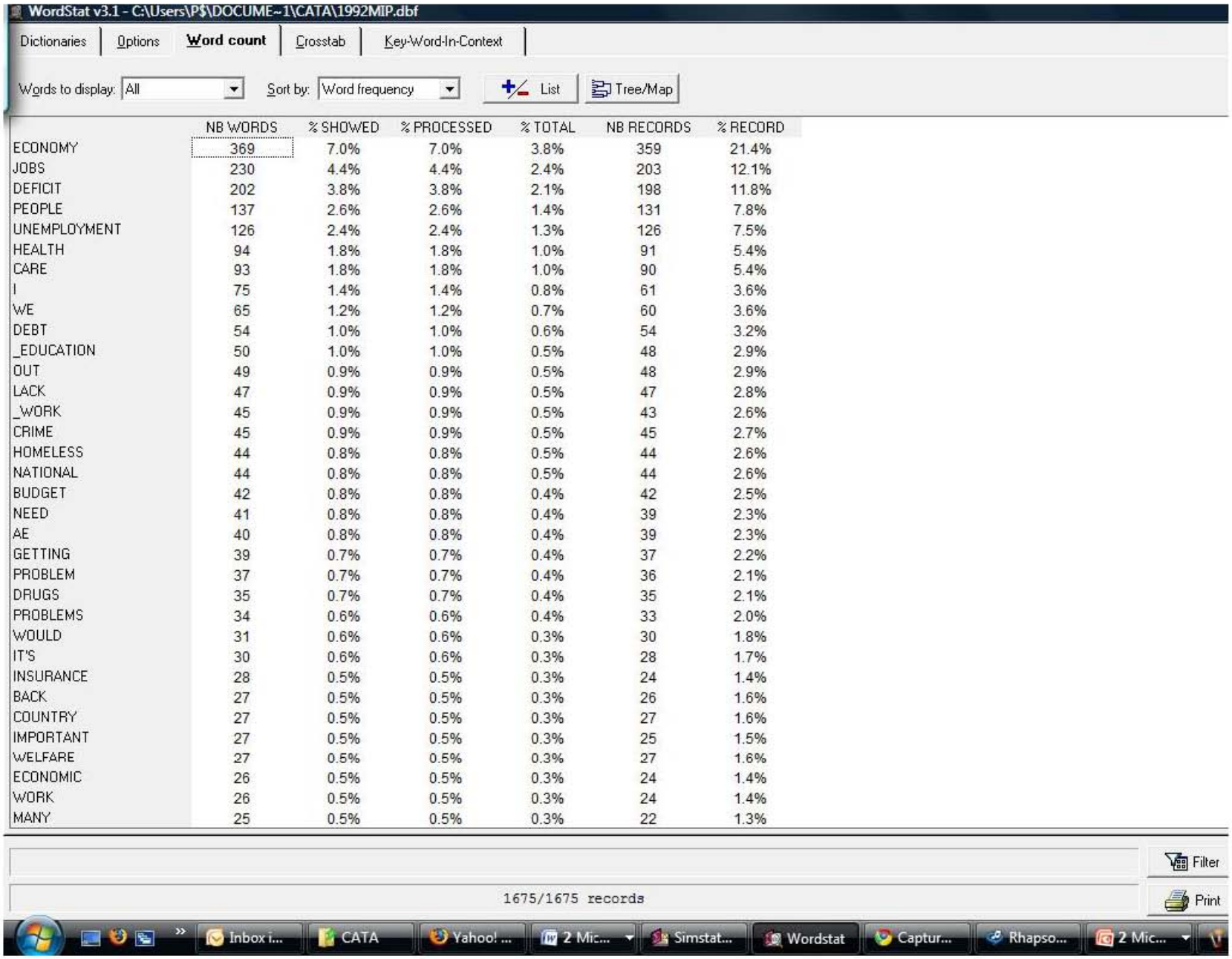
WordStat Output: KWIC



Sample Findings

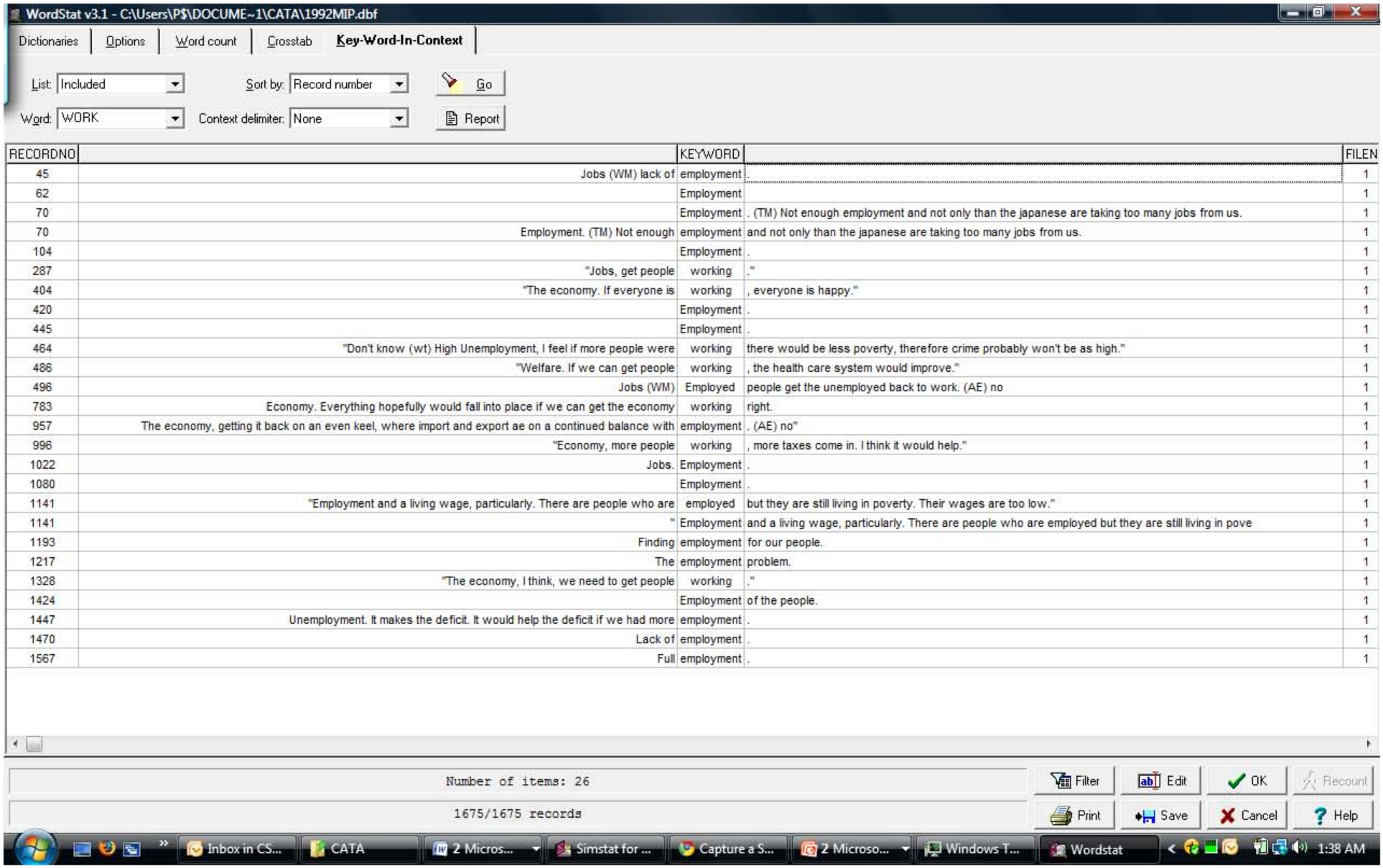
- Skalski (last night) did some basic analyses using the 1992 ANES open-ended data.
- Q: What do you think is the most important problem facing this country?
- Fairly easy to import responses into the program.
- Results as follows (with sample dictionary):

WordStat Output: Word Count



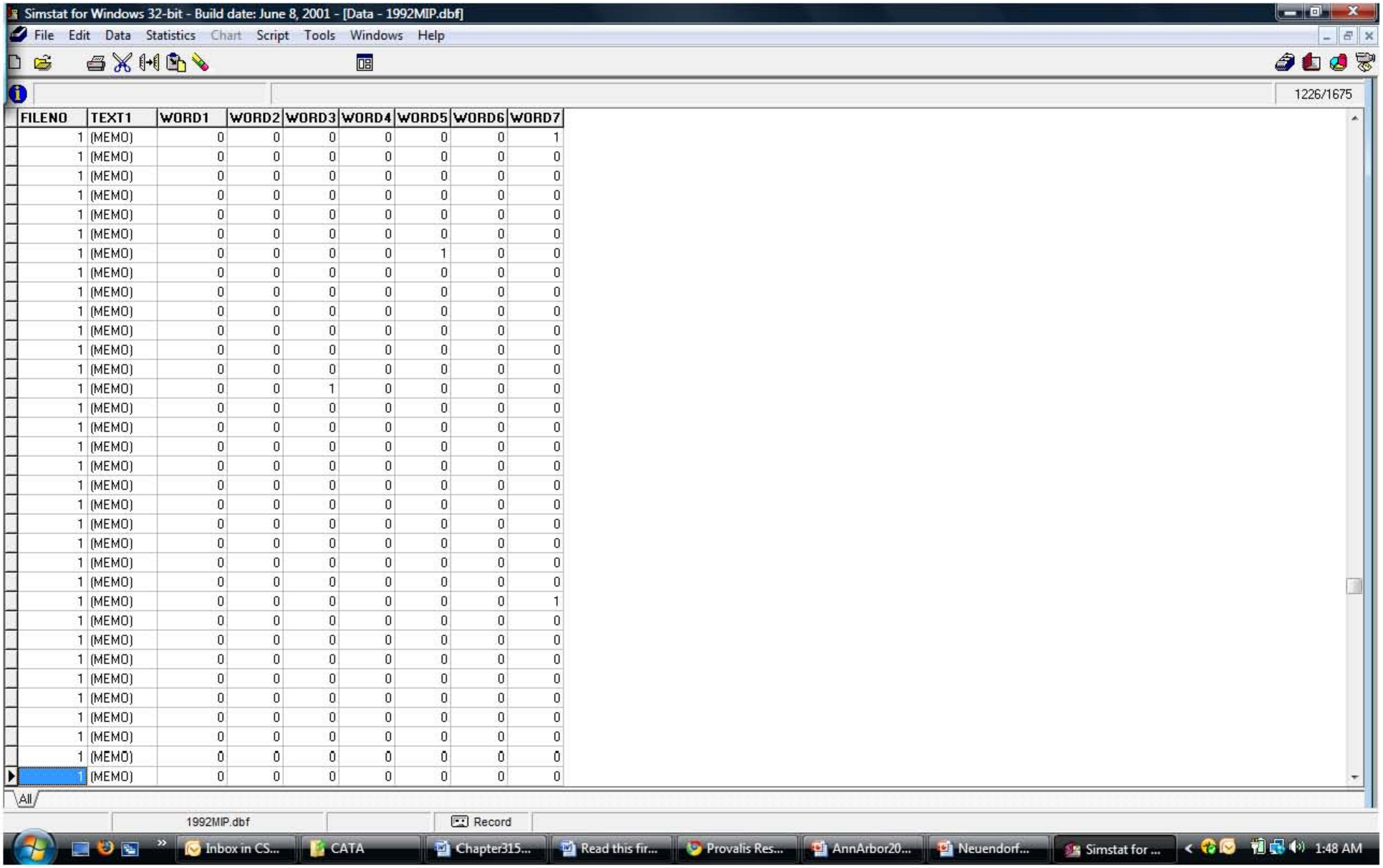
	NB WORDS	% SHOWED	% PROCESSED	% TOTAL	NB RECORDS	% RECORD
ECONOMY	369	7.0%	7.0%	3.8%	359	21.4%
JOBS	230	4.4%	4.4%	2.4%	203	12.1%
DEFICIT	202	3.8%	3.8%	2.1%	198	11.8%
PEOPLE	137	2.6%	2.6%	1.4%	131	7.8%
UNEMPLOYMENT	126	2.4%	2.4%	1.3%	126	7.5%
HEALTH	94	1.8%	1.8%	1.0%	91	5.4%
CARE	93	1.8%	1.8%	1.0%	90	5.4%
I	75	1.4%	1.4%	0.8%	61	3.6%
WE	65	1.2%	1.2%	0.7%	60	3.6%
DEBT	54	1.0%	1.0%	0.6%	54	3.2%
_EDUCATION	50	1.0%	1.0%	0.5%	48	2.9%
OUT	49	0.9%	0.9%	0.5%	48	2.9%
LACK	47	0.9%	0.9%	0.5%	47	2.8%
_WORK	45	0.9%	0.9%	0.5%	43	2.6%
CRIME	45	0.9%	0.9%	0.5%	45	2.7%
HOMELESS	44	0.8%	0.8%	0.5%	44	2.6%
NATIONAL	44	0.8%	0.8%	0.5%	44	2.6%
BUDGET	42	0.8%	0.8%	0.4%	42	2.5%
NEED	41	0.8%	0.8%	0.4%	39	2.3%
AE	40	0.8%	0.8%	0.4%	39	2.3%
GETTING	39	0.7%	0.7%	0.4%	37	2.2%
PROBLEM	37	0.7%	0.7%	0.4%	36	2.1%
DRUGS	35	0.7%	0.7%	0.4%	35	2.1%
PROBLEMS	34	0.6%	0.6%	0.4%	33	2.0%
WOULD	31	0.6%	0.6%	0.3%	30	1.8%
IT'S	30	0.6%	0.6%	0.3%	28	1.7%
INSURANCE	28	0.5%	0.5%	0.3%	24	1.4%
BACK	27	0.5%	0.5%	0.3%	26	1.6%
COUNTRY	27	0.5%	0.5%	0.3%	27	1.6%
IMPORTANT	27	0.5%	0.5%	0.3%	25	1.5%
WELFARE	27	0.5%	0.5%	0.3%	27	1.6%
ECONOMIC	26	0.5%	0.5%	0.3%	24	1.4%
WORK	26	0.5%	0.5%	0.3%	24	1.4%
MANY	25	0.5%	0.5%	0.3%	22	1.3%

WordStat Output: KWIC



RECORDNO		KEYWORD		FILEN
45	Jobs (WM) lack of	employment	.	1
62		Employment		1
70		Employment	. (TM) Not enough employment and not only than the japanese are taking too many jobs from us.	1
70	Employment. (TM) Not enough	employment	and not only than the japanese are taking too many jobs from us.	1
104		Employment	.	1
287	"Jobs, get people	working	."	1
404	"The economy. If everyone is	working	, everyone is happy."	1
420		Employment	.	1
445		Employment	.	1
464	"Don't know (wt) High Unemployment, I feel if more people were	working	there would be less poverty, therefore crime probably won't be as high."	1
486	"Welfare. If we can get people	working	, the health care system would improve."	1
496	Jobs (WM)	Employed	people get the unemployed back to work. (AE) no	1
783	Economy. Everything hopefully would fall into place if we can get the economy	working	right.	1
957	The economy, getting it back on an even keel, where import and export ae on a continued balance with	employment	. (AE) no"	1
996	"Economy, more people	working	, more taxes come in. I think it would help."	1
1022	Jobs.	Employment	.	1
1080		Employment	.	1
1141	"Employment and a living wage, particularly. There are people who are	employed	but they are still living in poverty. Their wages are too low."	1
1141	"	Employment	and a living wage, particularly. There are people who are employed but they are still living in pove	1
1193	Finding	employment	for our people.	1
1217	The	employment	problem.	1
1328	"The economy, I think, we need to get people	working	."	1
1424		Employment	of the people.	1
1447	Unemployment. It makes the deficit. It would help the deficit if we had more	employment	.	1
1470	Lack of	employment	.	1
1567	Full	employment	.	1

WordStat Output: Back to Data File



Final Points and Advice

- Strengths of CATA:
  - Quick (data preparation takes longest usually).
  - Can process large amounts of text with ease.
- Weaknesses of CATA:
  - Still not as good as human coding.
  - Computer cannot recognize certain message features, introducing measurement error.
  - Other validity concerns (e.g., dictionaries).

Final Points and Advice

- RECOMMENDATION: Use CATA at early analysis stages for exploratory purposes and data reduction.
- May also be used for automatic coding with good dictionaries and algorithms (?)
- Cross validation needed for confidence.
- Certain research Qs/data better suited.

Final Points and Advice

- Be wary of standard dictionaries.
  - Only trust those with full disclosure of details or other evidence of validity/effectiveness.
- Price matters:
  - Freeware: VBPro, General Inquirer, MCCALite
  - Relatively inexpensive ($100-$200): Student CATPAC, Diction, LIWC.
  - Expensive ($1000+): Full CATPAC, Wordstat.

In the Future….

- Datasets made available to encourage use and development of CATA programs.
- Programs with the ability to go beyond just words and examine larger units of analysis (e.g., sentences, paragraphs).
- Also machine learning from human coding examples (e.g., VCS) or combo programs.
- Coding of audio/visual information.

THE END

- Special thanks to Kim Neuendorf.
- Website: http://academic.csuohio.edu/kneuendorf/content/
- Questions?
  Email: p.skalski@csuohio.edu
- Thanks!