

# Automatic Text Analysis

# 8

This chapter provides an introduction to the application of automatic text analysis (ATA) in online behavioral research. We use the term *ATA* synonymously with the terms *computer content analysis*, *computer-assisted content analysis*, *computer-assisted text analysis*, and *computerized text analysis*. *ATA* has been defined as a set of methods that automatically extract statistically manipulable information about the presence, intensity, or frequency of thematic or stylistic characteristics of textual material (Shapiro & Markoff, 1997). In line with a quantitative notion of measurement, we focus exclusively in this chapter on *ATA* tools that extract quantitative information that can be subjected to statistical analysis.

The chapter covers basic information that helps researchers identify how they can use *ATA* in their online research. To maximize the chapter's utility, we focus on two specific *ATA* tools: Linguistic Inquiry and Word Count (LIWC; Pennebaker, Francis, & Booth, 2001) and Wmatrix (Rayson, 2008). We selected these tools because they (a) cover a range of *ATA* needs, (b) are user friendly, and operate fully automatically, and (c) are maintained by researcher groups with a track record in the field. We also selected them because we have used them extensively in our own online research (Cohn, Mehl, & Pennebaker, 2004; Gill, French, Gergle, & Oberlander, 2008; Lyons, Mehl, & Pennebaker, 2006; Oberlander & Gill, 2006).

Broader reviews of ATA strategies are provided by Mehl (2006), Krippendorff (2004), Neundorff (2002), Poppinga (2000), and West (2001).

### *What Is the Value of Automatic Text Analysis for Online Behavioral Research?*

ATA is a valuable method for online research for at least three reasons. First, textual data, the input for ATA, is abundant on the Internet. World knowledge is increasingly available online. Collaborative enterprises such as the Google Library Project or the Open Content Alliance are creating full-text online indices of millions of digitized documents. Furthermore, within only a few years, the Internet has become an indispensable means of daily communication. People routinely interact with others through e-mail, instant messages, chat rooms, blogs, and social networking sites. From a researcher's perspective, such text-based Internet data are informative and can be used to study psychosocial phenomena without running actual participants (for ethical considerations around the use of Internet data in research, see chap. 16, this volume).

For example, national newspaper coverage can be analyzed to compare the prevalence of psychological themes across cultures. Similarly, people's responses to disasters can be studied through the tracking of public postings on social sharing Web sites (Stone & Pennebaker, 2002). The global and archival nature of the Internet has made it possible to simulate the virtual equivalent of a multisite, longitudinal study conveniently and retroactively from the investigator's office computer—with the opportunity to obtain extensive baseline information on unpredictable events such as disasters after the fact (Cohn et al., 2004).

Second, textual data are often collected as part of online research anyway. ATA, then, can provide additional, low-cost means for exploratory data analysis. Researchers routinely include open-ended questions in their online surveys. Because of the burden of manual coding, however, participants' answers to these questions often remain unanalyzed. ATA can efficiently content analyze free responses. For example, in a Web-based survey of responses to the attacks of September 11, 2001, many participants responded to the final open-ended question "Is there anything else you would like to add?" in considerable detail and provided their accounts of the events (Skitka, L. J., personal communication, August 15, 2006). An ATA of cognitive complexity in such stories could help reveal individual differences in the processing of traumatic life events.

Third, data derived from ATA have some unique psychometrically desirable features: (a) They share zero method variance with the most

common method in the social sciences, the self-report rating scale; (b) they are objective in the sense that they ensure measurement equivalence across studies and labs using the same tool; (c) they are expressed in a nonarbitrary, naturally meaningful metric, the number or percentage of words in a text that fall into a certain category (e.g., positive emotion words, adverbs). These psychometric features positively affect the generalizability and ecological validity of text-analytically derived findings.

### *What Are the Potential Limitations of Automatic Text Analysis for Online Behavioral Research?*

ATA has the following potential limitations. First, it can be somewhat inflexible in its application. Whereas questionnaires can be constructed to measure any construct, ATA is generally constrained by the variables that the programs provide. LWC (Pennebaker et al., 2001), for example, has standard categories for positive and negative emotion words but does not extract information about specific emotions, such as pride, shame, or guilt. Similarly, Wmatrix (Rayson, 2008) identifies different types of verbs but not the ones suggested by the Linguistic Category Model (Sermin & Fiedler, 1988). However, some programs do provide users with limited freedom over the analysis. LWC, for instance, can search for lists of target words through user-defined dictionaries. Cohn et al. (2004) used this option to count how often participants used words such as *Osama*, *terrorist*, or *hijack* in their blogs after September 11, 2001.

Second, ATA applications are not always designed with the needs of the average behavioral scientist in mind. Some of the more powerful tools have been developed within computational linguistics and artificial intelligence and have their primary application in these fields (e.g., Cohn-Metrix [Graesser, McNamara, Louwrese & Cai, 2004]; Latent Semantic Analysis [Laudauer, Foltz, & Laham, 1998]). Yet these tools can be successfully used to answer behavioral research questions, and because of their computational advantages, they often extract critical language information that is lost with a simple word count (e.g., Campbell & Pennebaker, 2003). The consequence of using these tools outside of their original domain tends to be a loss of user friendliness. For example, Wmatrix, an application developed in corpus linguistics, does not have the "each-participant-a-line-each-variable-a-column" setup that psychologists are used to. Instead, it operates by comparing two text corpora (Oberlander & Gill, 2006; Rayson, 2008).

Finally, (word-count-based) ATA is sometimes viewed as simplistic in its approach. Mehl (2006) provided an in-depth discussion of (and rebuttal to) this concern. In essence, most word-count-based ATA tools (e.g., LIWC) neglect grammar (e.g., they do not distinguish between “the mother yelled at her child” and “the child yelled at her mother”); confuse context-specific word meanings (e.g., “What you did made me *mad*” vs. “I am *mad* about your cute curls”); and take metaphors (e.g., “I am on cloud nine”), irony (e.g., “It was as pleasant as getting a root canal”), and sarcasm (e.g., “Thanks a lot for blaming me for this”) for their literal meanings. More sophisticated tools, such as Wmatrix, are beginning to address these issues. However, it is important to note that it is not the computational sophistication of an ATA tool that determines the validity of a text-analytically derived finding; it is the degree to which the extracted linguistic information unambiguously answers a research question. For example, for the question of whether self-focused attention in depression manifests itself in an elevated use of first-person singular, the specific context in which people with depression use *I, me,* and *my* is not immediately relevant (Chung & Pennebaker, 2007).

In the remainder of the chapter, we provide a user guide for analyzing text-based Internet data with ATA. Because of space constraints, we limit this user guide to two ATA tools: LIWC (Pennebaker et al., 2001), as a word-count-based program that has gained considerable popularity within psychology; and Wmatrix (Rayson, 2008), as a more complex, Web-based ATA application developed in corpus linguistics. We illustrate the steps involved in a LIWC and Wmatrix analysis using (slightly modified) excerpts from four daily blogs. The blogs were selected from a larger data set collected by Nowson (2006). Two of the four blogs were written by female students (Blogs A and B) and two, by male students (Blogs C and D). The excerpts of the sample blogs are shown in Exhibit 8.1. In our user guide, we aim at providing sufficient detail to allow researchers to start analyzing their own textual data after reading the chapter. We also supplement our step-by-step guide with recommendations based on our own experiences working with the two tools.

## Word-Count–Based Psychological Text Analysis: Linguistic Inquiry and Word Count

LIWC was developed in the 1990s in the context of research on the salutary effects of writing about traumatic experiences. Over time, LIWC has been used more broadly to study the psychological implica-

### EXHIBIT 8.1

#### Excerpts From Four Sample Blogs

##### Blog A (Female Author 1)

Imagine you are happy and life is good, but it wasn't always like that, you once had a love and, at that point you knew, you felt that this was the one. But he wasn't. He broke you, he took everything you had and more from you, but you gave to him because you couldn't have enough of him. Then one day you realized how bad he was and you broke away. It took you a long time to get away from him, he had tied up your emotions and controlled you physically, you knew that he didn't want it to end.

##### Blog B (Female Author 2)

I find that I just don't have the stamina or brain power to write about war or politics at night. Which is fine, because you can't be all vitriolic and indignant 24 hours a day. Well you can, but I have better things to think about. I was thinking about writing more. No, not blogging more. Writing more. I was thinking about finding an agent. I was thinking of how I wanted to be nothing but a writer since I was about seven, how holding a pencil in my hand at that young age made me feel alive and important.

##### Blog C (Male Author 1)

This day fucking blows. I mean, I'm really happy for all the seniors, they're finally free, but this sucks for me. Everyone who I could really talk to is gone. Bye C\*\*, bye N\*\*, bye even M\*\*, bye everyone else. I'll probably stay in touch with most of them, so that's not even the worst part. If you know me you probably know what other horrible connotation this day has for me. I always knew it was coming, but could never really believe it. Right now I'm completely crushed. Then all these other stupid things popped up to make the day worse.

##### Blog D (Male Author 2)

I can't take it anymore, I'm going absolutely crazy. I'm locked up in this house and haven't been out besides school in a week. I'm questioning myself. And I'm totally fucking obsessed. Even had a falling out with someone today, and I didn't want that to happen at all. And I just don't know what to do. There's nothing I can do, it's just building inside me and has no way to escape. I feel like I'm going to explode. The only time I feel good is right after a run, and I fucking suck and can't even run a lot anymore.

*Note:* The excerpts were spell checked; the original punctuation was maintained; asterisks were used to preserve the authors' confidentiality. Data from *The Language of Weblogs: A Study of Genre and Individual Differences*, by S. Nowson, 2006, Unpublished Doctoral Dissertation, Scotland: University of Edinburgh. We are grateful to Scott Nowson for kindly making these data available.

tions of language use (Pennebaker, Mehl, & Niederhoffer, 2003). LIWC is a word-count–based ATA tool that operates by comparing each word of a given text with an internal dictionary consisting of 2,300 words. The default LIWC dictionary comprises 74 grammatical and psychological dimensions. The LIWC program was revised in 2001 and recently underwent a second major conceptual and technical revision (Pennebaker, Booth, & Francis, 2007). Because of the high popularity of, and

our extensive experience with LIWC 2001, we provide our user guide for this version. More information on LIWC 2007 is available at <http://www.liwc.net>.

LIWC is one of the most widely used ATA tools in psychology (Mehl, 2006). Its popularity is in part due to its effectiveness in meeting the needs of behavioral scientists: (a) it analyzes basic grammatical features of texts but also provides information about important psychological processes; (b) its categories have been psychometrically tested (c); the software is extremely user friendly; (d) it is available in several languages, and translated dictionaries with demonstrated equivalence to the original English dictionary are available in German (Wolf et al., 2008), Spanish (Ramírez-Esparaza, Pennebaker, García, & Suriá, 2007), and Dutch (Zijlstra, van Meerfeld, van Middendorp, Pennebaker, & Geenen, 2004) (psychometrically untested translations exist for Italian, Norwegian, and Portuguese, and new translations are being developed for Chinese, Hungarian, Korean, Polish, Russian, and Turkish); and (e) numerous studies have successfully used LIWC and thereby contributed to the construct validity of its categories.

## PREPARING THE DATA FOR LINGUISTIC INQUIRY AND WORD COUNT ANALYSIS

Data collection generally starts with sampling online text from participants (e.g., e-mails, instant messages) or directly from the Internet (e.g., blogs, chat rooms, discussion boards). We recommend the following steps to prepare the data for LIWC analysis:

1. Save the collected texts as plain text files (.txt); even though LIWC can analyze delimited text segments, we recommend creating a separate text file for each unit of analysis (i.e., each personal home page or each daily blog in a nested design).
2. Clean the texts by applying the “what you see is what LIWC gets” rule: remove any word that does not reflect the author’s language (e.g., e-mail histories, signatures, system information, advertising, buttons).
3. To maximize word recognition by the dictionary, submit the texts to an automatic spell-checker (LIWC analyses are case insensitive). Note, however, that Wolf et al. (2008) recently demonstrated that LIWC analyses of longer texts (more than 400 words) are fairly robust against regular amounts of typos and misspellings.
4. Render the texts consistent with LIWC typing conventions (documented in the program’s “Help” menu). These conventions regulate the handling of colloquialisms (e.g., *gatta*) and abbrevi-

ations (e.g., *w/*). Common verb contractions are included in the dictionary and need not be changed (e.g., *I’m, we’re, don’t, isn’t*). The use of slang in e-mails, chats, and instant messages (e.g., *LOL, CUL*) can challenge LIWC. If use of slang is of interest, user-defined dictionaries should be created to capture the relevant words or abbreviations. Otherwise, slang or abbreviations should be spelled out.

5. Manually tag fillers in natural language (see also the program’s “Help” menu). To avoid misclassifications, change *well, like, you know, I mean*, and *I don’t know to rwell, rlike, youknow, lmean*, and *I don’t like rush to lmean, I didn’t rlike rush*. The tagging is facilitated by using a “Search and Replace” function. However, it is critical to search for each individual occurrence and not to use “Replace All.”
6. For confidentiality reasons, remove personally identifying information: using asterisks (e.g., \*\*\*) helps to keep the word count accurate.

## RUNNING THE LINGUISTIC INQUIRY AND WORD COUNT ANALYSIS

Running the data through LIWC is straightforward. Clicking “Process text” in the “File” menu opens the “Select file(s) to process” window. Clicking “Select” after the text files to be analyzed have been marked opens the “LIWC results file” window where the output file is specified; clicking “Save” runs the analysis; and the output file (“LIWC results.dat”) opens after all files are processed. As a hands-on example, we submitted the four sample blogs depicted in Exhibit 8.1 to a LIWC analysis (see Table 8.1). Each blog was saved as a separate text file, cleaned, and spell checked. *I mean* in Blog C was tagged as a filler (*lmean*) and names were identified (e.g., *C\*\**). No other manual changes were made. Figure 8.1 shows a screenshot of the LIWC analysis.

We recommend processing text files from a folder that is located at the level below the hard drive (e.g., D:\LIWCtemp). Because of a bug in some versions of the software, LIWC 2001 sometimes crashes (i.e., freezes) when it processes files that are stored lower in the data hierarchy.

## INTERPRETING THE LINGUISTIC INQUIRY AND WORD COUNT OUTPUT

The LIWC output is a tab-delimited text file that can be imported into statistical software packages. Each column contains one LIWC variable; each row, the language information for one text file. The first column

TABLE 8.1

Results of the Linguistic Inquiry and Word Count (LIWC) Analysis of the Four Sample Blogs

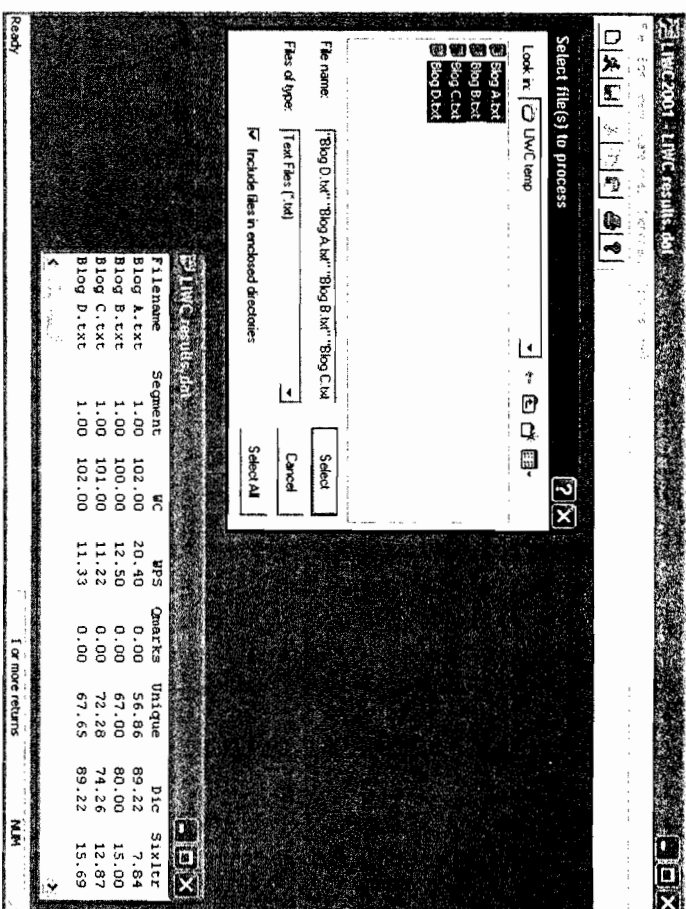
LIWC variable	Blog A	Blog B	Blog C	Blog D
Raw word count	102.0	100.0	101.0	102.0
Words captured by the dictionary	89.2	80.0	74.3	89.2
Emotional processes				
Emotion words	4.9	1.0	7.9	5.9
Positive	3.9	1.0	2.0	1.0
Negative	1.0	0.0	5.9	4.9
Cognitive processes				
Words of more than six letters	7.8	15.0	12.9	15.7
Cognitive mechanism words	12.8	14.0	9.9	3.9
Causation words	2.0	4.0	0.0	0.0
Insight words	3.9	6.0	4.0	2.0
Interpersonal processes				
First person singular pronouns	0.0	10.0	7.9	13.7
First person plural pronouns	0.0	0.0	0.0	0.0
Second person pronouns	14.7	2.0	2.0	0.0
Third person pronouns	8.8	0.0	2.0	0.0
Social words	25.5	2.0	7.9	2.0

Note. All LIWC variables except raw word count are expressed in percentages of total words; for Blog C, *I* mean was manually changed to *I*mean to tag it as a filler word; for heuristic purposes, the selected LIWC variables have been arranged into three important psychological domains.

shows the file name; the first row, the LIWC variable names. All variables (except the raw word count) are expressed in percentages of total words and are thus controlled for text length. LIWC by default provides language information along 74 dimensions.

The four blogs were comparable in length (see Table 8.1); in consideration of the reliability of low base-rate categories, we generally recommend using texts of at least 100 words. Across the four blogs, LIWC recognized around 80% of the words, which is typical for nontechnical language. Because statistical analyses with an *N* of 4 are not meaningful, we compare the blogs descriptively with regard to selected variables that have repeatedly been found to be implicated in psychological processes (Chung & Pennebaker, 2007). Heuristically, these variables can be thought of as capturing three important psychological domains: emotional processes (positive and negative emotion words), cognitive processes (cognitive mechanism words, words of more than six letters), and interpersonal processes (personal pronouns, social words). This is not to suggest that the LIWC categories that fall outside of these three domains are not important. Naturally, it is the research question that determines the relevance of a specific LIWC variable (e.g., sexual words,

FIGURE 8.1



Screenshot of the Linguistic Inquiry Word Count (LIWC) analysis of the four sample blogs; clicking "Select" in the "Select file(s) to process" window opens the "LIWC results file" window, where the name of the output file is specified; clicking "Save" runs the analysis, and the tab-delimited output file ("LIWC results.dat") opens after all files have been processed. From *LIWC* [Computer Software], by W. Pennebaker, Roger J. Booth, and Martha E. Francis, Pennebaker Conglomerates. Copyright 2007, LIWC.net, Pennebaker Conglomerates. Reprinted with permission.

numbers): within our own research, however, these categories have repeatedly emerged as important (Mehl, 2006).

With regard to emotional processes, Blog A emerged as relatively positive in emotional tone (3.9% positive vs. 1.0% negative emotion words), Blog B as neutral (1.0% vs. 0%), and Blogs C and D as quite negative (2.0% vs. 5.9%, and 1.0% vs. 4.9%, respectively); this is consistent with our impression after reading the blogs. Note that, in our

opinion, the labels of the LIWC emotion categories can be slightly misleading: “Positive emotion words” includes words such as *careful* and *perfect*; “negative emotion words” includes words such as *doubt* and *fail*. In the original context of writing about a trauma, these words likely adequately indicated the experience of positive or negative emotions. Yet when LIWC is used on a wider range of genres, the categories seem to tap more generally into the emotional tone of a text rather than the specific verbal expression of emotions.

With regard to cognitive processes, Blog A was less complex in its language than the other three blogs (only 7.8% of the words were longer than six letters). Blogs A and B, however, contained considerably more cognitive mechanism words (12.8% and 14.0%, respectively) than Blogs C (9.9%) and D (3.9%), suggesting a higher degree of cognitive (self-)reflection. It is interesting to note that Blog B, with its existential concerns and slightly esoteric word choice, also emerged as highly cognitive overall.

Finally, and maybe most important, the four blogs differed in the interpersonal processes they referenced. Beyond the general use of social words such as *talk* or *share*, interpersonal processes tend to be encoded in language through personal pronouns. Whereas Blog A was written from a detached second person perspective (*you*: 14.7%), Blogs B and D used first person singular at a high rate (10.0% and 13.7%, respectively). The frequent use of *I*, *me*, or *my* indicates personal involvement, with attention being on the self as social reference point. Psychologically, use of first person singular is correlated among other things with vulnerability for depression, low self-esteem, and the experience of stress (Pennebaker et al., 2003). Consistent with the psychological urgency it conveys, Blog D came out highest in the use of first person singular.

In sum, our analyses show that LIWC extracts language information at a psychologically meaningful level. This information often converges with ad hoc impressions derived from reading a text but also goes beyond what is noticed by a human observer (Mehl, 2006).

#### ADVANCED LINGUISTIC INQUIRY AND WORD COUNT FEATURES

LIWC has a few advanced features, such as the option of loading other dictionaries (e.g., foreign language dictionaries, special pronoun or particle dictionaries), creating user-defined dictionaries, and analyzing text in segments. Because of space limitations, we refer the interested reader to the detailed information provided in the program’s “Help” menu and the LIWC manual.

### Advanced Grammatical and Semantic Category Analysis: Wmatrix

Wmatrix is a powerful, Web-based ATA tool that uses corpus linguistics methods (Rayson, 2008; note that our description is based on Wmatrix2, the most recent version of the program). Corpus linguistics studies language using a large (usually electronic) collection of texts (i.e., a *corpus*; McEnery & Wilson, 1996). The research question often determines which texts are included in the corpus, whether they represent the whole of the English language (e.g., the British National Corpus; BNC; Burnard, 1995), or a particular group of interest (e.g., English language learners, International Corpus of Learner English; ICLE; Granger, Dagneaux, & Mcunier, 2002). In addition to working with existing corpora, researchers also use texts that were collected under specific conditions (e.g., for a study on personality and language use; Oberlander & Gill, 2006).

Analytically, Wmatrix uses a *corpus comparison* approach, which compares a corpus of interest, the *Research Corpus*, to a second corpus, the *Reference Corpus*. This process identifies ways in which the research corpus differs from the reference corpus. For example, to examine characteristics of second language learners, a research corpus such as the ICLE may be compared against a general collection of English language, for example, the BNC. Alternatively, corpora derived from two comparable groups (e.g., male and female authors) may be compared.

Wmatrix and LIWC differ in two main ways. First, Wmatrix does not impose a set of relevant language features (defined by the dictionary). Instead, it extracts comprehensive word use, grammatical, and semantic information on a set of texts. Data-driven approaches like Wmatrix operate bottom-up because they allow characteristic language features to emerge from the data. Dictionary-based ATA tools like LIWC, in contrast, operate top-down by focusing on predefined, theoretically derived dictionaries. Second, Wmatrix is more sophisticated than most dictionary-based ATA tools. Wmatrix uses an automatic part-of-speech tagger (Constituent-Likelihood Automatic Word-Tagging System [CLAWS] tagger) to disambiguate and classify the syntactic function of words in a sentence (e.g., in “She is a mine worker,” *mine* is a noun, not a pronoun). Similarly, Wmatrix uses a semantic tagger (UCREL Semantic Analysis System [USAS] tagger) to automatically disambiguate and classify the semantic function of words in a sentence (e.g., Wmatrix codes the example of *land mine* as “happy,” rather than “meteorological”).

Yet Wmatrix can also create challenges for behavioral researchers. First, the detailed information at the individual word level can be difficult

to interpret. What does it mean if women overuse the words *about* and *was* relative to men? This problem can be reduced by focusing on Wmatrix's analysis of broader grammatical and semantic features. With this type of analysis, Wmatrix results (e.g., use of "superlatives") can be interpreted similar to LWC results (e.g., use of "emotion words") with the difference that they are based on more comprehensive linguistic information.

Second, as noted before, Wmatrix analyzes text at the corpus level, not at the level of the individual author. Therefore, data collected from participants need to be clustered into text corpora. Although clustering texts is straightforward with discrete variables (e.g., gender, disease diagnosis, experimental group), it is more complex with continuous variables (e.g., age, extraversion). Then, the data need to be categorized, for example, by splitting a variable on the median or by forming extreme groups (e.g., participants low vs. high in extraversion; Oberlander & Gill, 2006). In this way, Wmatrix has been used to examine attitudes toward fashion (Wilson & Moudraia, 2006) and to code judgments of language used by science learners (Forsyth, Ainsworth, Clarke, Brundell, & O'Malley, 2006).

## PREPARING THE DATA FOR WMATRIX ANALYSIS

We now demonstrate a Wmatrix analysis step-by-step using our four sample blogs. Steps 1 through 4 of the data preparation are identical to those described in the section "Preparing the Data for Linguistic Inquiry and Word Count Analysis" (save as plain text files; clean up extraneous text not written by the author; check spelling; carefully consider the inclusion of abbreviations or slang). The following steps, however, are unique to preparing texts for analysis with Wmatrix:

- Wmatrix recognizes normal alphanumeric characters (A-Z, a-z, 0-9), but special characters (e.g., %, @, \*, ', ") require care and should be encoded using Standard Generalized Markup Language (SGML; e.g., ampersand, @, is encoded as "&amp;"). The use of punctuation also needs consideration; for details, see <http://www.comp.lancs.ac.uk/ucrel/claws/format.html>.
- For Wmatrix, it is recommended to render data anonymous by substituting alternative names for the original ones because, for example, asterisks as a means of deidentification ("\*\*") are not recognized by the program.
- Because text corpora are built from the original texts by merging the individual text files into larger files, we recommend including a unique identifier at the start and end of each text within a corpus; for example the start of a single text can be marked as "<text=filename>" and the end as "</text=filename>" (with "filename" as the identifier).

In addition to plain text (.txt) files, Wmatrix also supports HTML format as input files, which is particularly useful when dealing with online content (and which we recommend when the data contain SGML characters). A useful feature of Wmatrix is that it lists words it could not classify; we recommend checking this list for spelling or typographical errors. Note that filters (e.g., *well, like*) need not be manually tagged because Wmatrix detects them automatically.

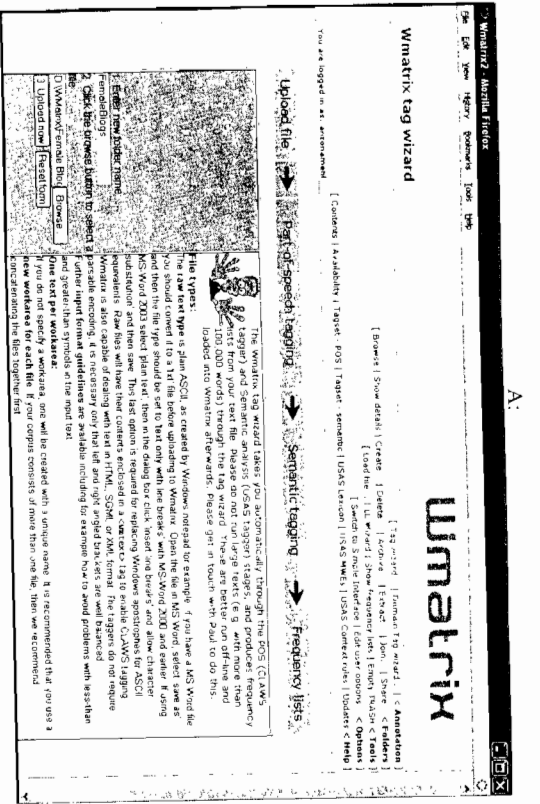
Finally, corpora need to be built from the individual texts. In our example, we cluster the four blogs into blogs by female authors (Blogs A and B) and male authors (Blogs C and D) and merge the four text files accordingly into two corpora, "FemaleBlogs.txt" and "MaleBlogs.txt." Note that our corpora are too small for answering research questions and are used for illustrative purpose only.

## RUNNING THE WMATRIX ANALYSIS

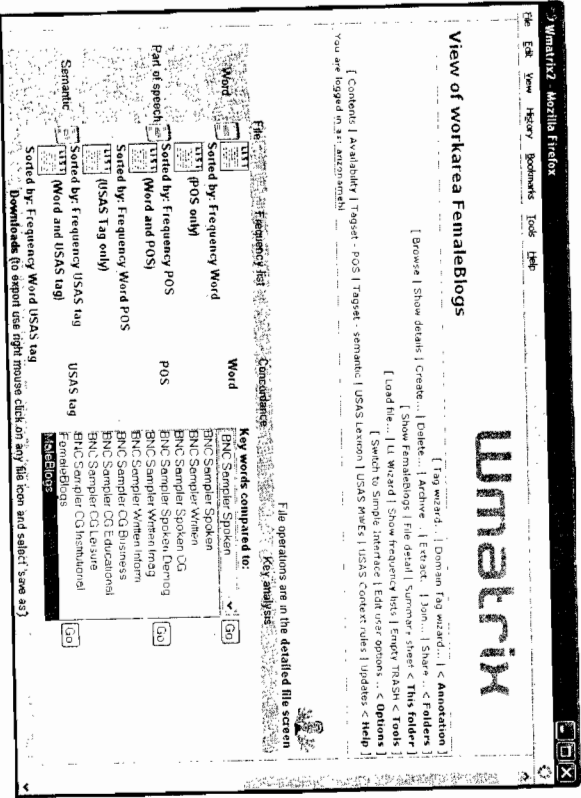
Wmatrix analyses consist of uploading the data, part-of-speech tagging, semantic tagging, and a word frequency analysis:

- Researchers log on to <http://ucrel.lancs.ac.uk/wmatrix2.html> with their unique username and password; free 1-month trial accounts are available for academic use.
- The tag wizard guides users through the automatic analysis of the data (see of Figure 8.2A). We recommend switching to the "Advanced Interface" by clicking on the respective icon in the "Options" menu. The user then (a) names a work area (e.g., "FemaleBlogs"), (b) specifies a data file to be uploaded and tagged (e.g., "FemaleBlogs.txt"), and (c) starts the process by pressing "Upload now."
- Processing the text can take between a few seconds and several minutes, depending on the complexity of the corpus. Once the analyses are finished, Wmatrix jumps to the view of the work area (e.g., "FemaleBlogs," see Figure 8.2B). The work area has links to the raw output of the word frequency, part-of-speech, and semantic tagging, as well as pull-down menus for specifying the comparison corpus.
- The final step involves the comparison of two corpora with regard to word, part-of-speech, and semantic frequencies. This can be done using one of the built-in reference corpora (e.g., BNC) or a specific comparison corpus collected by the researcher (e.g., "MaleBlogs"). Pressing "Go" runs the corpus comparison and opens up the output. The output for the word frequency (Figure 8.3A) and the semantic (Figure 8.3B) comparison of our corpora of female and male bloggers are shown in Figure 8.3.

FIGURE 8.2

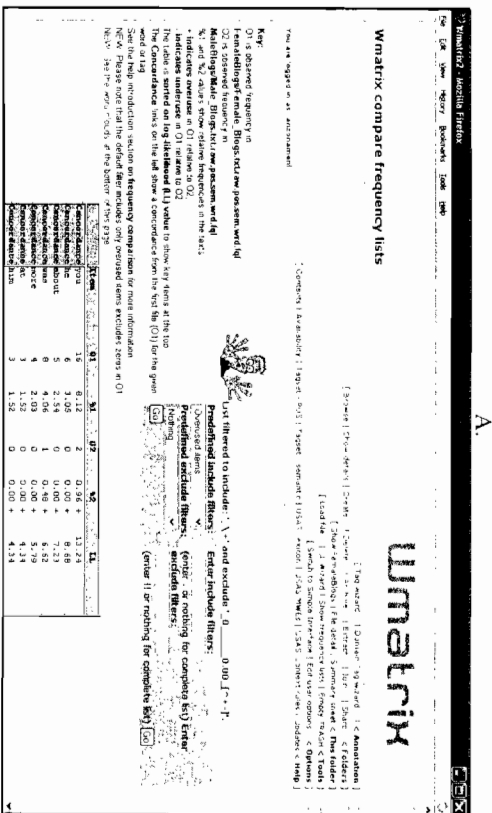


B:

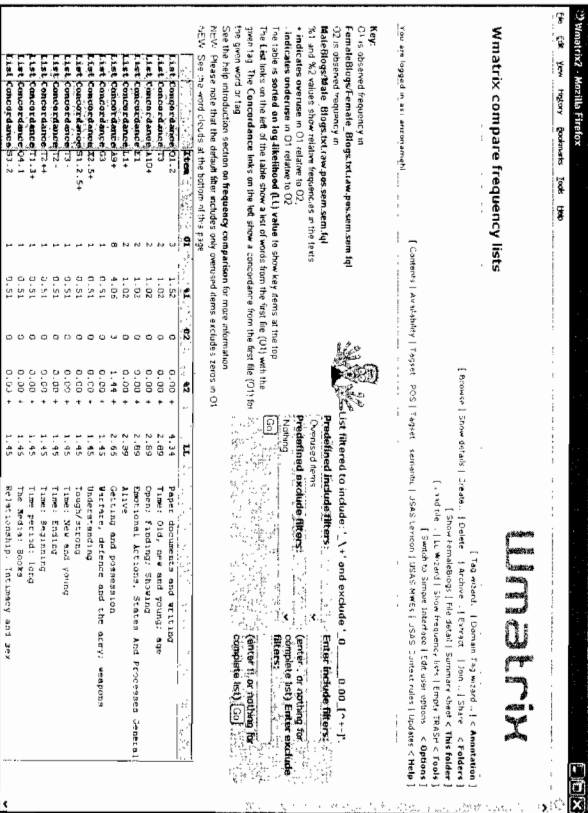


Wmatrix screenshots (advanced interface): A: Uploading the text files (FemaleBlogs.txt); B: "FemaleBlogs" work area with word, part-of-speech, and semantic tag files; "MaleBlogs" is specified as the comparison corpus. From *Wmatrix*. Copyright 2000–2009 by UCREL. Reprinted with permission.

FIGURE 8.3



B:



Wmatrix screenshots (advanced interface) for the "FemaleBlogs" versus "MaleBlogs" comparison. A: Output for the comparison of word frequencies; B: Output for the comparison of semantic tag frequencies. Results filtered for items overused in female blogs. From *Wmatrix* by Paul Rayson, Computer Software, Lancaster, UK: UCCEL, Lancaster University. Copyright 2000–2009 by UCCEL. Reprinted with permission.

## INTERPRETING THE WMATRIX OUTPUT

The Wmatrix corpus comparison output provides statistical information on the overuse and underuse of individual words, part-of-speech (i.e., grammatical category), and semantic features of one text corpus relative to another. In the advanced interface (shown in the screenshots and described in our example), it supplies log-likelihood (LL) values to estimate the reliability of the between-corpora differences in words and text features; in the simple interface, it uses graphical “cloud” images to represent the most significant features. Rayson (2003) recommended 15.13 as a critical LL value ( $p < .0001$ ) to minimize capitalization on chance due to the number of tests; however, a value of 6.63 could be used with care. It is not surprising that, given our small amount of textual data, no language differences passed this threshold. Yet, female relative to male bloggers tended to overuse the words *you*, *he*, and *about* (see Figure 8.3A). Consistent with these findings for single words, female bloggers also grammatically overused second person personal pronouns (*you*), third person singular subjective pronouns (*she*), and past tense verbs. Wmatrix did not reveal any semantic features that females overused greatly (all LL values  $< 6.63$ ; see Figure 8.3B). Male bloggers, on the other hand, tended to overuse the word contraction *'m*, grammatically general adverbs (*truly*, *even*), the auxiliary verb *am*, and semantically nonspecific quantifiers (*even*).

As our sample comparison of male and female blog language use reveals, Wmatrix can provide linguistic information at a very fine-grained level. The researcher's challenge then lies in conceptually interpreting the identified characteristic word-based, grammatical, or semantic group differences (Oberlander & Gill, 2006). Yet it is important to note that the extensive—and potentially overwhelming—Wmatrix output also offers unique potentials: For the first time, it is possible to automatically assess almost any grammatical feature and a wide spectrum of semantic language use features. Many of these features have escaped other, simpler ATA tools, and many of them are inherently important to behavioral scientists (e.g., use of comparatives and superlatives; personality traits; referents to emotional states, health, and disease). It is because of its unique blend of computational power (automatic grammatical and semantic disambiguation), linguistic sophistication, and user friendliness that we decided to introduce it to behavioral scientists. As an easy-to-use ATA tool, Wmatrix has wide applicability and unique potentials for revealing the natural interactions among psychological and linguistic processes.

## Summary and Conclusion

In this chapter, we have highlighted some of the possibilities that ATA offers for working with text-based Internet data and provided a user guide for two ATA approaches. Which of the two tools, then, should researchers

use? In general, for psychologically complex (e.g., involving several continuous measures) but linguistically relatively simple phenomena (e.g., focusing on pronoun use only), a dictionary-based approach like LWC seems optimally suited. In the contrary situation, that is, for rich linguistic data and dichotomous psychological variables, Wmatrix seems best suited. It is our conviction, though, that unique insights result from a synergy of both approaches—the joint use of the psychological LWC categories and the linguistic Wmatrix categories (Oberlander & Gill, 2006). Additional techniques that provide a good balance between linguistic sophistication and psychological complexity include Coh-Metrix and Latent Semantic Analysis (e.g., Gill et al., 2008). More information about these tools can be found at <http://cohmetrix.memphis.edu> and <http://lsa.colorado.edu>.

## Additional Resources

Chung, C. K., & Pennebaker, J. W. (2007). The psychological function of function words. In K. Fiedler (Ed.), *Social communication: Frontiers of social psychology* (pp. 343–359). New York: Psychology Press.

This chapter provides a comprehensive summary of research on psychological aspects of natural word use with a focus on variables such as gender, age, culture, personality, depression, and deception. Mehl, M. R. (2006). Quantitative text analysis. In M. Eid & E. Diener (Eds.), *Handbook of multivariate measurement in psychology* (pp. 141–156). Washington, DC: American Psychological Association.

This chapter discusses quantitative text analysis in the context of multimethod measurement; it reviews nine text analysis strategies in psychology and classifies them on four dimensions.

Neuendorf, K. A. (2002). *The content analysis guidebook*. Thousand Oaks, CA: Sage.

This comprehensive book on content analysis as a scientific method (manual and computerized) discusses measurement issues and describes various text analysis programs. A helpful online companion is provided at <http://ATAdemic.cs.ohio.edu/kneuendorf/content>

## References

- Burnard, L. (Ed.). (1995). *Users' reference guide for the British National Corpus Version 1.0*. Oxford, England: Oxford University Computing Services.
- Campbell, R. S., & Pennebaker, J. W. (2003). The secret life of pronouns: Flexibility in writing style and physical health. *Psychological Science*, 14, 60–65.

- Chung, C. K., & Pennebaker, J. W. (2007). The psychological function of function words. In K. Fiedler (Ed.), *Social communication: Frontiers of social psychology* (pp. 343–359). New York: Psychology Press.
- Cohn, M. A., Mehl, M. R., & Pennebaker, J. W. (2004). Linguistic indicators of psychological change after September 11, 2001. *Psychological Science*, 15, 687–693.
- Forsyth, R., Ainsworth, S., Clarke, D., Brundell, P., & O'Malley, C. (2006, June). Linguistic computing methods for analysing digital records of learning. *Proceedings of the 2nd International Conference on e-Social Science*, Manchester, England.
- Gill, A. J., French, R. M., Gergle, D., & Oberlander, J. (2008). The language of emotion in short blog texts. *Proceedings of the Association for Computing Machinery Conference on Computer Supported Cooperative Work (CSCW 2008)* (pp. 299–302). New York: ACM Press.
- Graesser, A., McNamara, D. S., Louwerse, M., & Cai, Z. (2004). Coh-Matrix: Analysis of text on cohesion and language. *Behavioral Research Methods, Instruments, & Computers*, 36, 193–202.
- Granger S., Dagneaux E., & Meunier F. (2002). *The International Corpus of Learner English* [Handbook and CD-ROM]. Louvain-la-Neuve, France: Presses Universitaires de Louvain.
- Krippendorff, K. (2004). *Content analysis: An introduction to its methodology*. Beverly Hills, CA: Sage.
- Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to Latent Semantic Analysis. *Discourse Processes*, 25, 259–284.
- Lyons, E. J., Mehl, M. R., & Pennebaker, J. W. (2006). Pro-anorexics and recovering anorexics differ in their linguistic Internet self-presentation. *Journal of Psychosomatic Research*, 60, 253–256.
- McEneaney, T., & Wilson, A. (1996). *Corpus linguistics*. Edinburgh, Scotland: Edinburgh University Press.
- Mehl, M. R. (2006). Quantitative text analysis. In M. Eid & E. Diener (Eds.), *Handbook of multivariate measurement in psychology* (pp. 141–156). Washington, DC: American Psychological Association.
- Neuendorf, K. A. (2002). *The content analysis guidebook*. Thousand Oaks, CA: Sage.
- Nowson, S. (2006). *The language of weblogs: A study of genre and individual differences*. Unpublished Doctoral Dissertation, University of Edinburgh, Scotland.
- Oberlander, J., & Gill, A. J. (2006). Language with character: A corpus-based study of individual differences in e-mail communication. *Discourse Processes*, 42, 239–270.
- Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). *Linguistic Inquiry and Word Count: LIWC 2007*. Austin, TX: LIWC available from <http://www.liwc.net>
- Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). *Linguistic Inquiry and Word Count (LIWC) 2001*. Mahwah, NJ: Erlbaum. Available from <http://www.liwc.net>
- Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, 54, 547–577.
- Popping, R. (2000). *Computer-assisted text analysis*. London: Sage.
- Ramírez-Esparza, N., Pennebaker, J. W., García, A. F., & Surriá, R. (2007). La psicología del uso de las palabras: Un programa de computadora que analiza textos en Español [The psychology of word use: A computer program that analyzes texts in Spanish]. *Revista Mexicana de Psicología*, 24, 85–99.
- Rayson, P. (2003). *Matrix: A statistical method and software tool for linguistic analysis through corpus comparison* (doctoral thesis). Lancaster University.
- Rayson, P. (2009). *Wmatrix: A Web-based corpus processing environment*. Computing Department, Lancaster University. Available from <http://ucreclan.ac.uk/wmatrix/>
- Semin, G. R., & Fiedler, K. (1988). The cognitive functions of linguistic categories in describing persons: Social cognition and language. *Journal of Personality and Social Psychology*, 54, 558–568.
- Shapiro, G., & Markoff, J. (1997). A matter of definition. In C. W. Roberts (Ed.), *Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts* (pp. 8–31). Mahwah, NJ: Erlbaum.
- Stone, L. D., & Pennebaker, J. W. (2002). Trauma in real time: Talking and avoiding online conversations about the death of Princess Diana. *Basic and Applied Social Psychology*, 24, 172–182.
- West, M. D. (Ed.). (2001). *Theory, method, and practice in computer content analysis*. New York: Ablex.
- Wilson, A., & Moudraia, O. (2006). Quantitative or qualitative content analysis? Experiences from a cross-cultural comparison of female students' attitudes to shoe fashions in Germany, Poland, and Russia. In A. Wilson, P. Rayson, & D. Archer (Eds.), *Corpus linguistics around the world* (pp. 203–217). Amsterdam: Rodopi.
- Wolff, M., Horn, A. B., Mehl, M. R., Haug, S., Kordy, H., & Pennebaker, J. W. (2008). Computergestützte quantitative Textanalyse: Äquivalenz und Robustheit der deutschen Version des Linguistic Inquiry and Word Count [Computerized quantitative text analysis: Equivalence and robustness of the German adaptation of Linguistic Inquiry and Word Count]. *Diagnostica*, 54, 85–98.
- Zijlstra, H., van Meerweld, T., van Middendorp, H., Pennebaker, J. W., & Geenen, R. (2004). De Nederlandse versie van de "Linguistic Inquiry and Word Count" (LIWC): Een gecomputeriseerd tekstanalyseprogramma. [Dutch version of Linguistic Inquiry and Word Count (LIWC), a computerized text analysis program]. *Gedrag & Gezondheid*, 32, 271–272.

**Advanced Methods  
for Conducting**

# **Online Behavioral Research**

---

**Edited by**

**Samuel D. Gosling**

**John A. Johnson**