5 Data Mining at the Intersection of Psychology and Linguistics

R. Harald Baayen
University of Nijmegen and Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

Large data resources play an increasingly important role in both linguistics and psycholinguistics. The first data resources used by both psychologists and linguists alike were word frequency lists such as Thorndike and Lorge (1944) and Kučera and Francis (1967). Although the Brown corpus on which the frequency counts of Kučera and Francis were based was very large for its time, comprising some one million word forms carefully sampled from different registers of English, many common words did not appear in the frequency lists, while others appeared with counterintuitive frequencies of use.

Gernsbacher (1984) addressed this issue, claiming that subjective frequency estimates would be superior to objective frequency counts. Corpus-based frequency counts would be inherently unreliable due to regression towards the mean. In another corpus, higher frequency words would be less frequent, and lower frequency words would be more frequent. These considerations have led many psychologists to turn away from research directly addressing frequency effects in lexical processing. This distrust in psychology of corpus-based frequency data mirrors the rejection of corpora as a valid source of information about grammar in generative linguistics.

Fortunately, more and larger corpora were developed, driven in part by the needs of commercial lexicography, in part by the research interests of corpus linguistics, and in part by the growing needs for reliable data in computational linguistics and linguistic engineering. These develop-

ments made the creation possible of the CELEX lexical database, an initiative of the psycholinguist Levelt, which is widely used in both the (computational) linguistic and psycholinguistic research communities. For English, this resource provides the frequencies in the Cobuild corpus at the time that this corpus comprised some 18 million words. The British National Corpus (BNC) currently is the largest available tagged corpus of British English, with 100 million words, of which 10 million transcribed spoken English. Thus, linguistics now has at its disposal large data resources, although much remains to be done with respect to annotation and the sampling of everyday spoken language. The largest unstructured source of examples of language in use is, nowadays, the World Wide Web, which combines the advantage of quantity with the disadvantages of the absence of linguistic annotation and the restriction to written language.

The lexical resources developed specifically within psychology are relatively new, scarce, and small compared to linguistic corpora. Perhaps the most important large data resources are WordNet (Miller, 1990; Fellbaum, 1998), the Florida association norms (Nelson, McEvoy, & Schreiber, 1998), and the databases of visual lexical decision latencies, word naming latencies, and subjective frequency ratings of Balota, Cortese, and Pilotti (1999) and Spieler and Balota (1998). These resources provide psycholinguistics with a wealth of data on the behavioral properties for thousands of words. Although here too much remains to be done, especially from a morphological point of view, these behavioral data resources are a tremendous step forwards compared to the small numbers of items typically studied in factorial psycholinguistic experiments.

The aim of this chapter is to show that, when combined, the linguistic and psychological resources become a particularly rich gold mine for the study of the lexicon and lexical processing. I will illustrate the new methodological possibilities for data mining by examining the databases compiled by Balota and colleagues, in combination with CELEX, the BNC, and WordNet. For 1424 monomorphemic and monosyllabic nouns, and 832 monomorphemic and monosyllabic verbs, we study the predictive potential of a range of variables for three behavioral measures: visual lexical decision latencies and word naming latencies in ms, and subjective familiarity ratings on a 7-point scale.

In what follows, I will show that mining these combined resources yields several new insights. Section 1 examines the correlational structure of the predictors, and sheds new light on the nature of word frequency. Section 2 shows that subjective frequency ratings are an independent variable in their own right, just as response latencies in, for instance,

visual lexical decision or word naming. Section 3 illustrates how the information carried by response latencies can be mined by means of lexical covariates, and calls attention to the methodological advantages of regression above factorial designs and the importance of relaxing the linearity assumption.

DATA MINING THE PREDICTORS

Lexical variables that are regularly considered in studies of lexical processing are frequency, length, number of neighbors, and bigram frequency. Frequency of use is well-studied and highly robust predictor. In this study, we will estimate a word's frequency of occurrence by means of the token frequency of its orthographic form in the subcorpus of written English that is part of the BNC, a subcorpus comprising 89.7 million words.

Although this subcorpus has the advantage of being large, it need not be the case that its frequency estimates are the best predictors for lexical processing. Frequency estimates based on spoken language are likely candidates of having superior predictivity, as speech is more fundamental to normal day-to-day communication than writing. Fortunately, the BNC also contains two subcorpora of spoken English. The demographic subcorpus (4.2 million words) provides transcriptions for spontaneous conversations of speakers sampled across England recorded with portable tape recorders. The context-governed subcorpus (6.2 million words) provides transcriptions of oral language in more formal contexts, often requiring preparation, such as speeches, sermons, and lessons. As the three subcorpora of the BNC differ in size, we scale the frequencies to a corpus size of 1 million words.

Word length is a second variable that is often taken into consideration. In what follows, we consider word length measured in letters. A third variable that has received widespread attention (e.g., Andrews, 1989; Grainger & Jacobs, 1996) is the density of a word's orthographic similarity neighborhood. Orthographic (or phonological) similarity is generally quantified in terms the number of words of the same length that are identical to a given target word except for one letter. The neighborhood density can then be estimated in terms of the count of such orthographic neighbors. A fourth variable is a word's mean bigram frequency. In this chapter, the mean bigram frequency is calculated as the mean of the logarithms of the frequencies of the pairwise letter pairs (including the initial and final spaces as letters). As the bigram consisting of the first two (non-space) characters might be predictive for word naming, it is included as well. Note that word length, neighborhood density, bigram

frequency, initial bigram frequency, and also frequency of occurrence, are all measures of a word's form.

More recently, various lexical measures for a word's meaning have been explored. Best studied is the morphological family size measure, the number of complex words in which a given word occurs as a constituent (see e.g., Schreuder & Baayen, 1997). Following Moscoso del Prado (2003), we also consider two related measures, the derivational and inflectional entropy. The entropy of a probability distribution is defined as $\Sigma\ p\ \log(1/p)$, and quantifies the amount of information of that distribution. Applied to the probabilities (estimated by relative frequencies) of a word's morphological family members, one obtains the derivational entropy, which can be viewed as a variant of the family size measure in which the family members are weighted for their token frequency. The entropy can also be calculated for a word's inflectional variants, in which case it estimates the information complexity of that word's inflectional paradigm.

Two other semantic measures first explored in Baayen, Feldman, and Schreuder (2004) address a word's number of meanings by means of the synsets listed in the WordNet resource. WordNet (Miller, 1990) is a lexical database in which words are organized in synonym sets, known as synsets. For *hand,* WordNet lists several synsets, for instance, {*hand, manus, hook, mauler, mitt, paw*}, {*handwriting, hand, script*}, {*hand, deal*}, {'*hired hand*', *hand,* '*hired man*'}, {*pass, hand, reach,* '*pass on*', '*turn over*', *give*}. By counting the number of different synsets in which a word is listed in WordNet, we can gauge how many different meanings a word has. In what follows, I consider two complementary measures. The first measure counts the number of different synsets in which the word is listed as such. I will refer to this measure as the simple synset count. The second measure counts the number of synsets in which the word is part of a compound or phrasal unit (such as *hired hand* in the *hand* example). I will refer to this count as the complex synset count.

Many of these predictors are known to be intercorrelated. For the system of correlations of frequency, word length, number of meanings, and dispersion (the number of different texts in which a word occurs), the reader is referred to Köhler (1986). The correlational structure of family size, derivational entropy, word frequency, and word length is investigated in Moscoso del Prado (2003). In the following analyses, we logarithmically transformed all measures with skewed distributions (frequency, family size, derivational entropy, and the synset measures) in order to reduce potential atypical effects of high-valued outliers.

Figure 5.1 provides a summary of the correlational structure of the numerical predictors by means of a hierarchical cluster analysis using Spearman's ρ^2 as a nonparametric distance measure. Interestingly, the BNC frequency measures cluster with the WordNet synset measures and also with the paradigmatic morphological measures family size and derivational entropy. The measures of word form, word length, mean bigram frequency, and number of neighbors, appear in a different branch of the dendrogram together with the frequency of the initial bigram. Although all numerical predictors are correlated, word frequency emerges from the distributional statistics of English primarily as a semantic measure, and not as a measure of form-related lexical properties. Baayen et al. (2004) came to similar conclusions using a different technique, principal components orthogonalization. This distributional observation supports the hypothesis of Balota and Chumbley (1984) that the word frequency effect has a strong post-access component, and argues against the idea that frequency effects would arise primarily or exclusively at the access level.

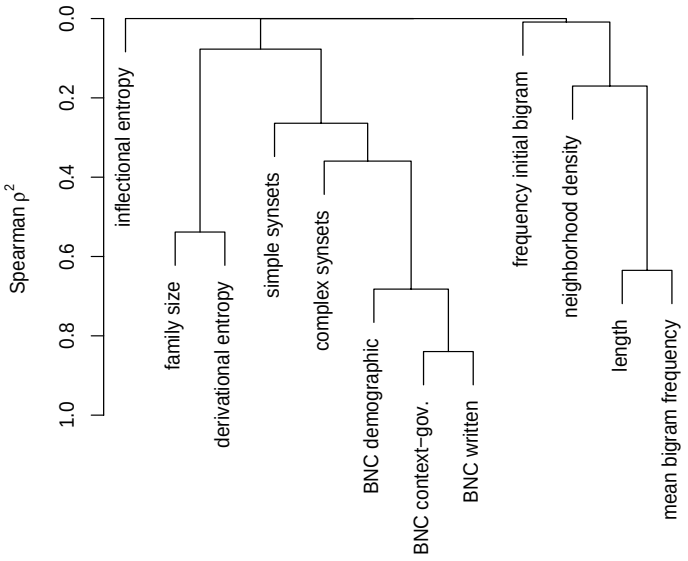

FIG. 5.1. Hierarchical clustering of the predictors for the ratings and response latencies in the Balota database.

Note, finally, that the written frequencies reveal a tighter correlation with the context-governed frequencies than with the demographic frequencies, the frequencies recorded for the spontaneous conversations. This tighter correlation is in line with the more formal character of both the context-governed samples and the written samples in the British National Corpus. In the next sections, we see that this clustering of the frequency measures is reflected in behavioral measures of lexical processing.

DATA MINING FAMILIARITY RATINGS

Are familiarity ratings an alternative, perhaps better frequency measure than corpus-based frequency counts, as suggested by Gernsbacher (1984)? Although this is commonly believed, we can ask ourselves whether ratings measure only frequency of occurrence. Would subjects be able to tap into frequency without being influenced by the many other variables that are known to affect, for instance, lexical decisions?

To address this question, let's fit a multiple regression model to the ratings in the database of Balota and colleagues. Before doing so, two preliminary questions need to be addressed. First, the substantial correlational structure characterizing our set op predictors points to a collinearity problem. When the predictors are highly collinear, as in this data set, it is difficult to tease the effects of the individual predictors apart (see Baayen et al., 2004, for detailed discussion). For the present purpose, it suffices to address the high collinearity of the three frequency measures. It makes no sense to include all three in the same regression model. In what follows, I therefore selected the written frequency as primary frequency measure. In order to study the potential predictivity of the other frequency measures, I constructed two new variables, the standardized differences between the written frequency and the two spoken frequencies. These standardized differences are only mildly correlated with the written frequencies ($r = -0.069$ for the demographic standardized differences, $r = -0.19896$ for the context-governed standardized differences).

Second, it is important not to impose a-priori that a predictor enters into a linear relation with the dependent variable. A flexible way of exploring potential non-linearities is to make use of restricted cubic splines (see e.g., Harrell, 2001: 16-24). In construction, a spline is a flexible strip of metal or piece of rubber that is used for drawing the curved parts of objects. In statistics and physics, a spline is a function for fitting nonlinear curves. This function is itself composed of a series of simple cubic polynomial functions defined over a corresponding series of

intervals. These polynomials are constrained to have smooth transitions where they meet, the knots of the spline. The number of intervals is determined by the number of knots. In order to capture more substantial nonlinearities, one will need more knots. In other words, the number of knots is a smoothing parameter. In the following analyses, I have used the minimum number of knots necessary to model non-linearities.

Figure 5.2 shows that subjective familiarity ratings are indeed a dependent variable in their own right. Each panel shows the partial effect of a predictor on the rating scores in the database of Balota and colleagues in the model resulting from a stepwise multiple regression analysis. The first row of panels shows relations that have a significant non-linear component ($p < 0.0001$ for the nonlinear component of word frequency, $p = 0.0002$ for the nonlinear component of the frequency difference, and $p = 0.0409$ for the nonlinear component of family size).

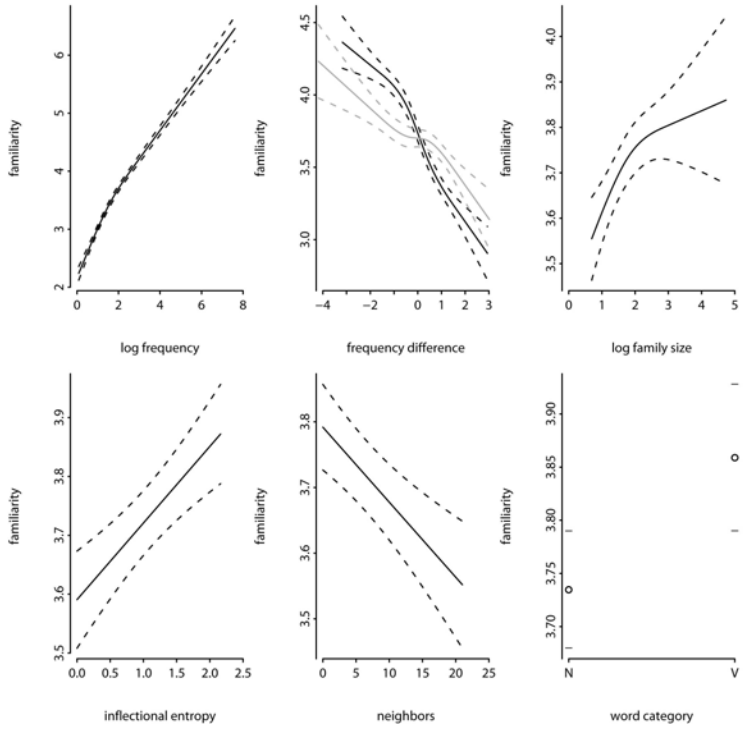

FIG. 5.2. Partial effects of the predictors on familiarity ratings with 95% confidence intervals. Only significant effects (linear and nonlinear) are shown ($\alpha = 0.05$). $R^2 = 0.696$, bootstrap corrected $R^2 = 0.695$.

These non-linear relations were fitted using restricted cubic splines with four, four, and three knots respectively. The second row of panels presents the linear effects of inflectional entropy and neighborhood density, and the factorial effect of word category.

Note, first of all, that there is a strong nearly linear relation between written frequency in the BNC and familiarity ratings, with very narrow 95% confidence intervals. This shows that subjective frequency estimates are good predictors for objective corpus-based (relative) frequencies, as expected.

The second panel in the top row shows that familiarity ratings decrease as the frequency difference increases. The black lines represent the frequency difference with the demographic subcorpus (the spontaneous conversations), the grey lines represent the corresponding difference with the context-governed subcorpus (the more formal oral language). Positive values indicate a word is encountered more in writing than in speech, negative values indicate the word is more typical of speech than of writing. What this panel shows, then, is that words that are typical of speech are rated more highly than words that are predominantly written. Note that this effect is stronger for the demographic subcorpus than for the context-governed subcorpus. This suggests that an optimal frequency measure for predicting ratings should be based on a large corpus of spontaneous conversations. Apparently, subjective frequency estimates tap primarily into frequency of exposure and use in spontaneous everyday spoken discourse.

The remaining panels show that subjective frequency estimates capture more than just frequency of occurrence. Family size predicts the ratings, with higher families leading to higher ratings (cf. Schreuder & Baayen, 1997). The nonlinearity points to a ceiling effect for large families. A large inflectional entropy likewise leads to higher ratings, suggesting that a greater information load of the inflectional paradigm causes a word to be perceived as more familiar. On the other hand, neighborhood density is negatively correlated with the familiarity ratings. Note that even word category differences (Wcat) are reflected in the ratings, with verbs eliciting higher ratings than nouns.

The observation that familiarity ratings are an independent variable in their own right, just as response latencies or eye fixation durations, has important methodological consequences. Ratings should not be used as a substitute for corpus-based frequency counts. Matching for familiarity ratings, for instance, implies at least partial matching for a series of other variables, potentially including variables of interest, and reduces the likelihood of finding significant effects. Likewise, familiarity ratings should not be included along with frequency counts in a regression

analysis of, for instance, lexical decision, just as one would not normally include lexical decision latencies as a predictor for, for example, eye fixation durations.

DATA MINING RESPONSE LATENCIES

The databases compiled by Balota and colleagues also make available response latencies for visual lexical decision and word naming. Again, we make use of restricted cubic splines in order to trace potential nonlinearities. In order to reduce the skewness of the distributions of latencies, the latencies in both tasks were logarithmically transformed (using the natural logarithm). Figure 5.3 shows the partial effects of the predictors on word naming, Figure 5.4 is the corresponding plot for visual lexical decision. Only significant predictors are shown, and nonlinearities are shown only when significant.

As in the analysis of the ratings, frequency of use as gauged by the BNC counts of written English is a strong predictor for both tasks. Note that the confidence intervals are quite narrow, and more so for the low frequency words than for the high-frequency words. The wider confidence intervals for the higher frequencies are a consequence of the relative data sparseness in the higher frequency ranges, even after the logarithmic transform. The narrow confidence intervals even for the lowest frequency ranges show that there is no reason to be particularly concerned about the reliability of corpus-based estimates of the frequencies (probabilities) of low-frequency words.

In both lexical decision and word naming, the frequency difference measure comparing the written frequency with the frequency in the spontaneous conversations (the demographic subcorpus) is positively correlated with RT. No such correlation is present for the comparison with the frequencies in the context-governed subcorpus. Whereas the ratings revealed a reduced effect for the context-governed counts, the latency measures restrict the effect to truly spontaneous, unprepared speech. This supports the hypothesis that the word frequency effect is grounded in casual day-to-day verbal interaction.

What is striking is the large number of predictors that enter into nonlinear relations with the response latencies. Some of these non-linearities are readily interpretable. For instance, there seems to be a floor effect for word frequency in both tasks for the higher-frequency words. The U-shaped curves for word length might reflect response optimization for the most frequently occurring word length (the median word length in the data is 4 letters). However, for the U-shaped curves for the family size measure and for the simple synset counts I do not have

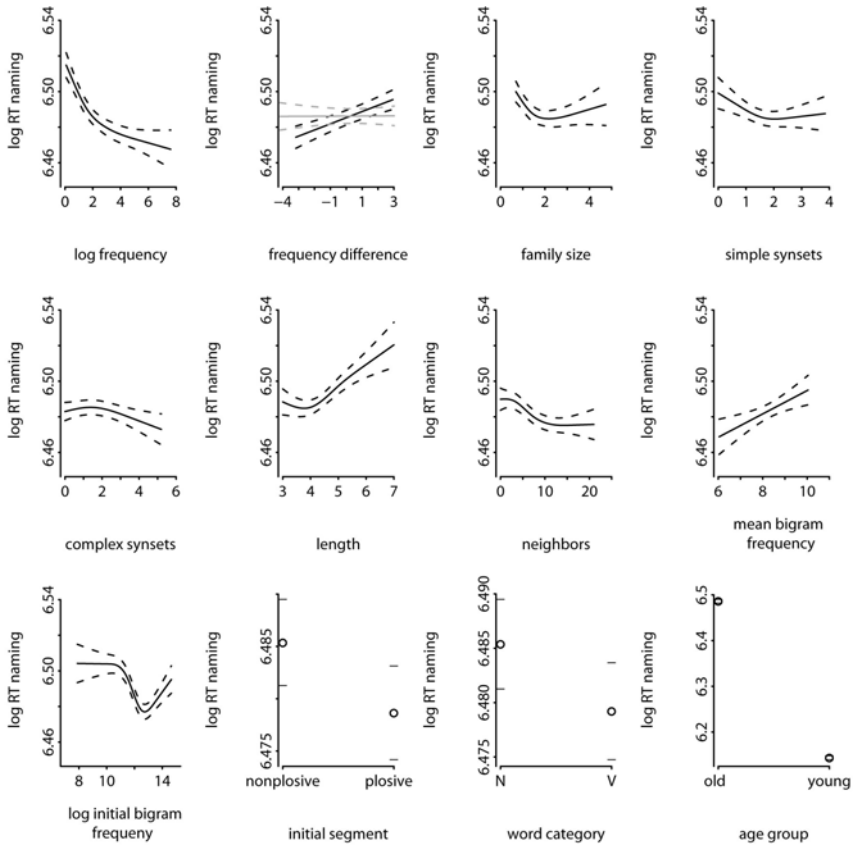

FIG. 5.3. Partial effects of the predictors on word naming latencies with 95% confidence intervals. Only significant effects (linear and nonlinear) are shown (α = 0.05). R2 = 0.942, bootstrap corrected R2 = 0.941.

an explanation. Apparently, the advantage of having, for example, more morphological family members turns into a disadvantage when the family size becomes very large. More research and modeling is required here.

Note that in both tasks, the age group of the participants is a very strong predictor of the response latencies, as illustrated by the relevant panels on the third rows of Figures 5.3 and 5.4. The 95% confidence intervals are so narrow that they are indistinguishable form the circles representing the group means.

There are a number of differences between the two tasks. In visual lexical decision, the effect of word category did not reach significance, in

word naming, it did. The log initial bigram frequency, included as a covariate for word naming, turned out to be a significant predictor in both tasks, a simple facilitatory linear predictor in visual lexical decision, and a strangely shaped non-linear predictor in word naming. In the word naming study, the nature of the initial phoneme (plosive vs. non-plosive) was included as a control variable for the voicekey. Plosives elicited shorter latencies than non-plosives: The voice key is especially sensitive to the burst of the plosive.

Another interesting difference between the two tasks concerns the effect of neighborhood density, non-linear but facilitating in word naming, but U-shaped in visual lexical decision. In Figure 5.4, the panel for neighborhood density on the second row, with the scale on the Y-axis fixed to the range of the frequency effect, is repeated on the third row, with the scale on the Y-axis set to the range of the effect of neighborhood density itself. Although the effect of neighborhood density is relatively small, the U-shaped form of the graph in the visual modality is especially interesting, as it suggests an inhibitory component for larger neighborhood sizes, as reported for French (Grainger & Jacobs, 1996) but not for English (Andrews, 1989). An inhibitory effect for neighborhood density in reading was also observed by Baayen et al. (2004) after addressing the problem of collinearity, once the effect of semantic variables has been partialled out.

From a methodological point of view, these nonlinearities bear witness to the importance of exploratory regression analysis without a-priori assumptions about linearity, and to the dangers of factorial designs for the study of numeric variables. Consider again the count of orthographic neighbors in visual lexical decision. A factorial design contrasting high versus low conditions for this variable would fail to observe that it is a relevant predictor. In addition, the arbitrariness of the cutoff points for factorial contrasts increases the risk of inconsistent results across replication studies using different materials. The inconsistent results reported in the literature for neighborhood density might have arisen precisely because of these reasons.

CONCLUDING REMARKS

Data mining the combined large lexical data resources of linguistics and psychology has led to a number of insights. First of all, it suggests that the linguistic variable of 'word frequency' should be rehabilitated in psychology. The present study illustrates the reliability of word frequency as a predictor of behavioral measures, even though Gernsbacher (1984) previously discredited such counts and falsely

accused them of regression toward the mean (see Baayen, Moscoso del Prado, Schreuder, & Wurm, 2003, for technical discussion).

In addition, the measure comparing written frequency in the BNC with spoken frequencies revealed that frequency in spontaneous, unprepared speech is the optimal predictor. As pointed out by Gernsbacher (1984), corpus-based frequency counts are sometimes rather counterintuitive, especially when based on written language sampled from more formal registers. This study provides an example of how this issue can be addressed by bringing appropriate covariates for register variation such as the frequency difference measures into the statistical model.

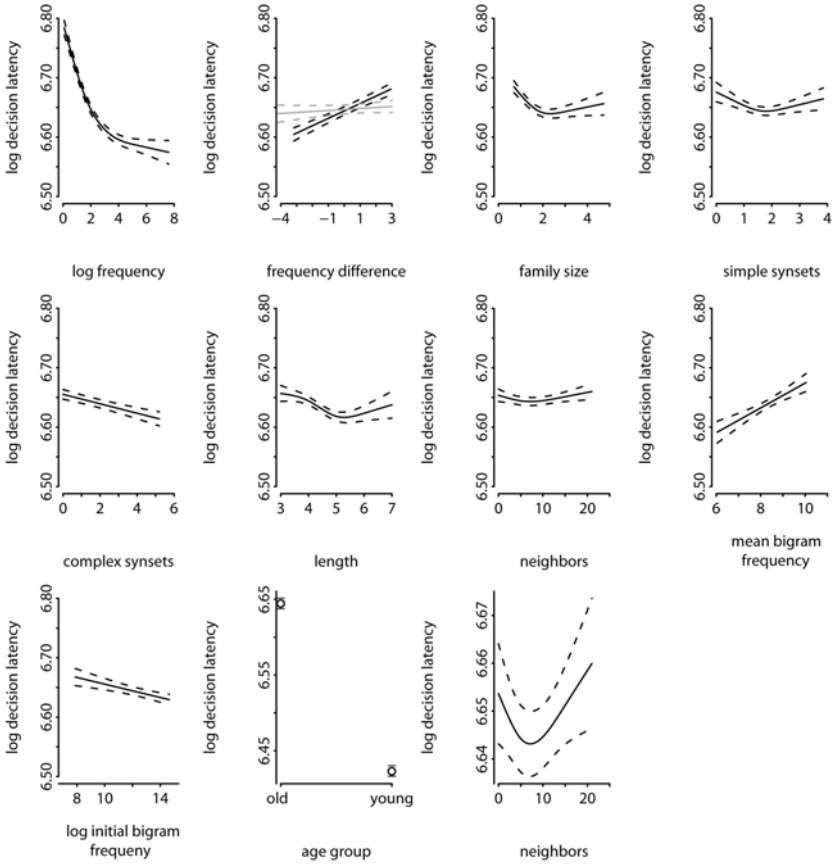

FIG. 5.4. Partial effects of the predictors on visual lexical decision latencies with 95% confidence intervals. Only significant effects (linear and nonlinear) are shown ($\alpha = 0.05$). $R^2 = 0.729$, bootstrap corrected $R^2 = 0.727$.

Second, this exercise in data mining shows that the complexity of subjective frequency ratings has been underestimated. Introspection does not produce pure estimates of frequency of occurrence, but estimates that capture a wide range of other variables in addition to frequency. The methodological consequence of this finding is that matching on familiarity, or including familiarity as a covariate, should be avoided.

Third, this study demonstrates the methodological importance of data mining with tools that are appropriate for detecting the functional relations between predictors and behavioral measures in their full complexity. Current research on lexical processing makes use mostly of factorial designs and occasionally of linear regression. With respect to factorial designs, however, the dichotomization required to transform a numerical variable into a factor brings along a number of disadvantages.

The imposition of factor levels such 'high' and 'low' forces the researcher to impose arbitrary cutoff points, and the gain in power obtained by considering only extreme values is offset by the risk of having studied atypical extremes and comes with the price of having no insight whatsoever into the shape of the regression function, which, as the examples discussed in this study demonstrate, may be highly non-linear. U-shaped curves such as observed for neighborhood density may wreak havoc in a literature based exclusively on factorial experiments. It is important to realize the importance of non-linearity. Straight lines are ubiquitous in man-made environments, but exceptional in the natural world. Note that even the linear relations in Figures 5.3 and 5.4 imply non-linear relations between the untransformed response latencies and the relevant predictors—there is not a single linear predictor for the RTs in milliseconds.

An additional problem with factorial designs is that they require matching on all other potentially relevant variables. For the lexical data such as illustrated in this chapter, we have no less than 9 significant numerical predictors for visual lexical decision as well as for word naming, and more variables are sure to be discovered. It is simply impossible to match a dichotomous contrast in one of these variables on all the others. In other words, even though for many psycholinguists a 'real' experiment is a factorial experiment, this view is misguided for any domain of inquiry in which the dichotomized continuous variable is one of a cluster of correlated variables. In short, only use a factor when no more fine-grained numerical information is available.

Even when the main variable of interest is a true factor (such as word category in the present examples), it is important to include all known potentially relevant covariates in the design, in order to guarantee incrementality in research and to avoid a random walk through the

complex multidimensional parameter space that is under investigation in (psycho)linguistics.

A final insight that this study has to offer is that the tighter correlation of the word frequency measure with measures of word meaning compared to measures of word form sheds new light on the (psycho)linguistic interpretation of lexical frequency. Whereas previous research in quantitative linguistics has addressed the mathematical form of the functions relating frequency to other lexical measures (see e.g., Köhler, 1986), the present study addressed the tightness of these relations. This led to the insight that that word frequency is primarily a measure of conceptual familiarity.

In conclusion, the construction of large data resources, both in linguistics and in psychology, although labor intensive and time consuming, is essential for understanding the more subtle details of linguistic structure and its consequences for language processing.

REFERENCES

Andrews, S. (1989). Frequency and neighborhood size effects on lexical access: Activation or search? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 802-814.

Baayen, R., Feldman, L., & Schreuder, R. (2004). *Collinearity and non-linearity in regression analyses of simple word recognition, word naming, and subjective frequency estimation*. Manuscript submitted for publication.

Baayen, R. H., Moscoso del Prado, F., Schreuder, R., & Wurm, L. (2003). When word frequencies do NOT regress towards the mean. In R. H. Baayen & R. Schreuder (Eds.), *Morphological structure in language processing* (pp. 463-484). Berlin: Mouton de Gruyter.

Balota, D. A., & Chumbley, J. I. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception and Performance, 10*, 340-357.

Balota, D., Cortese, M., & Pilotti, M. (1999). *Visual lexical decision latencies for 2906 words.* [http://www.artsci.wustl.edu/~dbalota/lexical_decision.html]

Fellbaum, C. E. (1998). *WordNet: An electronic database*. Cambridge: MIT Press.

Gernsbacher, M. A. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General, 113*, 256-281.

Grainger, J., & Jacobs, A. M. (1996). Orthographic processing in visual word recognition: A multiple read-out model. *Psychological Review, 103*, 518-565.

Harrell, F. E. (2001). *Regression modeling strategies (Statistics and Computing Series)*. New York: Springer.

Köhler, R. (1986). *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik*. Bochum: Brockmeyer.

Kučera, H., & Francis, W. N. (1967). *Computational analysis of Present-Day American English*. Providence, RI: Brown University Press.

Miller, G. A. (1990). Wordnet: An on-line lexical database. *International Journal of Lexicography, 3*, 235-312.

Moscoso del Prado, F. (2003). *Paradigmatic effects in morphological processing: Computational and cross-linguistic experimental studies*. MPI Series in Psycholinguistics, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). *The University of South Florida word association, rhyme, and word fragment norms*. http://www.usf.edu/FreeAssociation/.

Schreuder, R., & Baayen, R. H. (1997). How complex simplex words can be. *Journal of Memory and Language, 37*, 118-139.

Spieler, D. H., & Balota, D. A. (1998). *Naming latencies for 2820 words*. [http://www.artsci.wustl.edu/~dbalota/naming.html]

Thorndike, E. L., & Lorge, I. (1944). *A teacher's word book of 30.000 words*. New York: Columbia University Press.