A brief introduction to regression designs and mixed-effects modelling by a recent convert[1]

Laura Winther Balling

Abstract

This article discusses the advantages of multiple regression designs over the factorial designs traditionally used in many psycholinguistic experiments. It is shown that regression designs are typically more informative, statistically more powerful and better suited to the analysis of naturalistic tasks. The advantages of including both fixed and random effects are demonstrated with reference to linear mixed-effects models, and problems of collinearity, variable distribution and variable selection are discussed. The advantages of these techniques are exemplified in an analysis of a word recognition experiment in Danish.

1. Introduction

Almost by definition, experimental work and statistics play a central role in psycholinguistics. In the study of word recognition, a wide range of both visual and auditory experimental tasks are used, including naming (reading aloud of single words), eye-tracking of reading of either single words or text, gating (where gradually larger chunks of a word are heard and identified) and lexical decision. In lexical decision, participants are presented with a mixture of words and phonotactically or orthographically legal nonwords and asked to determine whether they recognise each token. Although lexical decision is arguably dominant, there is considerable

[1] I am grateful to Harald Baayen for introducing me to the use of regression designs and mixed effects modelling, and for patiently helping me use and understand these techniques. I am also grateful to Ocke-Schwen Bohn for comments and discussion, and to the editors for their comments on an earlier draft of this article. Sections 2 and 3 of this chapter are adapted from chapter 4 of Balling (2008), section 4 refers extensively to chapter 5 of the same.

variation in the tasks employed to study word recognition. In contrast to the range of variation in task choice, the vast majority of experimental studies have relied on very similar designs and statistical techniques, namely factorial designs and analyses of variance (ANOVAs). There are a number of problems with this statistical approach, and recently, regression designs have emerged as a both more powerful and more flexible alternative to ANOVAs (e.g. Balota *et al.* 2004; Baayen *et al.* 2006; New *et al.* 2007; Kuperman *et al.* in press).

I have recently converted to this approach to experimental design and statistics in my work on word recognition and found it overwhelmingly more informative than the alternative factorial designs. I believe a similar conversion could be useful for many researchers who study translation processes experimentally. More specifically, regression designs allow statistical control of a number of variables which cannot be controlled experimentally; this is particularly important in more naturalistic tasks. Factorial designs require strict control between groups of experimental items and therefore make more naturalistic, less experiment-like approaches difficult, if not impossible.

In the present article, some of the advantages of regression designs in comparison with the traditional factorial approaches will be discussed in section 2. Section 3 presents a few more practical aspects of the analysis of regression design data, focusing on one specific technique, namely linear mixed effects modelling as implemented in the R environment for statistical computing, and discussing both advantages and potential problems. Although concentrating on a specific technique, there are no low-level practical instructions; for that, the reader is referred to, for instance, Dalgaard (2002) for a general introduction to R, or Baayen (2008) for an introduction to R with a specific focus on linguistic data. The discussion of linear mixed-effects models is not exhaustive, but includes some of the more important aspects. Throughout, examples are taken from my own work on word recognition, but section 4 is devoted to exemplifying the advantages of regression in a specific experiment. Section 5 offers conclusions and further perspectives for the field of translation studies.

2. Comparing experimental designs: factorial and regression designs

Factorial designs are based on experimental control between groups of experimental items, so-called conditions. In the simplest case, all potentially relevant variables are controlled except one variable of interest which is manipulated systematically between two conditions. One variable which is often investigated in the word recognition literature is corpus frequency: the frequency with which a word (or other constituent) occurs in a representative corpus is assumed to index participants' familiarity with that word. In a factorial design investigating the effect of frequency, words are chosen to fall into two or more groups according to frequency – typically high vs. low frequency – while all other variables that could affect recognition must be matched between these conditions. If there is a significant difference in the dependent variable between the high- and low-frequency conditions, this is assumed to be caused by the difference in frequency – i.e. if participants recognise the high-frequency words faster than the low-frequency words, this is assumed to be an effect of frequency. However, as an increasing number of variables are found to affect lexical processing, it becomes increasingly difficult to control all the relevant variables (Cutler 1981). This difficulty of controlling all the relevant variables may make it questionable whether a difference between the conditions is in fact caused by the variable which was manipulated. Moreover, very strict control criteria often result in a relatively low number of items, thus considerably reducing the statistical power of many factorial experiments.

By contrast, multiple regression techniques make it possible not only to assess the effect of frequency (if that is the variable of interest), but to investigate multiple correlations of both central and more control-oriented independent variables with the dependent variable at the same time. In this way, the same experiment can assess the effects of frequency, length and semantic variables on word recognition latency at the same time. The same potentially relevant variables need to be considered in a regression design as in a factorial design, but because each variable can be investigated rather than controlled, the number of items is not restricted by matching criteria, and additional variables can be investigated post hoc. The larger number of items increases the statistical power of regression designs relative to typical factorial designs.

This description indicates a primary advantage of regression designs, namely that they are generally much more informative than factorial designs. Two additional reasons are outlined in the remainder of this section.

One major problem with factorial designs is that the between-conditions manipulation of one variable is best suited to factors where the levels are to some degree naturally given, such as differences between different types of words that can be construed as natural categories. However, factorial designs are also often employed to investigate numerical variables such as frequency which must then be dichotomised into a condition of high-frequency items and a condition of low-frequency items. As shown by Cohen (1983), dichotomisation of continuous variables results in a substantial loss of statistical power (see also MacCallum *et al.* 2002), which is already often reduced because of the scarcity of items that fit the control criteria. This makes it difficult to detect significant effects.

Moreover, from a linguistic point of view, the control of everything except one key variable can result in items being unrepresentative. For instance, for complex words such as *book-s* and *read-ing*, the corpus frequencies of the base (*book* or *read*) and of the whole complex word are typically highly correlated: if the base *book* is a frequent word in a language, then the plural form *book-s* is likely also to be frequent. A factorial design that investigates one of these two variables, say base frequency, requires that base frequency is dichotomised – typically into a condition of high base-frequency words and a condition of low base-frequency words – while the other variable, whole-word frequency, must be controlled between the two conditions. This necessitates the choice of items for which the two measures are *not* highly correlated, i.e. items that may not be representative of words in the language. Cases such as *moon-s* (the base of which is much more frequent than the whole word) and *cliff-s* (the base of which is less frequent than the whole word) are often employed in experiments (for more examples, see Taft 2004), but the extent to which the recognition of such words generalises to the whole vocabulary may be questionable.

This problem of dichotomisation is avoided in regression designs. If base frequency is a variable of interest, the base frequency of the experimental items is included as one of the independent variables and the effect of this naturally graded variable is assessed. Moreover, because several variables can be investigated in the same analysis, it is unnecessary to control the whole-word frequency, and unrepresentative items like *moon-s* and *cliff-s* can be avoided. Given sufficient items and a large enough range of values, regression designs also allow the researcher to investigate whether frequency has a non-linear effect. For instance, it has been observed that frequency has a strong effect in the lower frequency ranges while the effect becomes smaller or disappears entirely for the higher frequency ranges (e.g. Baayen *et al.* 2006). Moreover, the same experiment could reveal interactions with other continuous or categorical variables.

Another advantage of regression designs is that they permit statistical control of longitudinal effects in experiments. De Vaan *et al.* (2007) found that reaction times were co-determined by the speed of the preceding responses: in their lexical decision experiment, a slow response was generally preceded by responses that were also slow. Similarly, responses are generally slower if the previous response in a lexical decision task was incorrect (Balling & Baayen in press; Balling 2008). In lexical decision, it also matters whether the previous item was a word or a nonword, though different directions have been observed for this effect (Diependaele *et al.* 2007; Balling 2008). Finally, effects have been found of where in the experimental list an item occurs, either in the shape of learning where response speed increases or as a fatigue effect where responses become slower as the experiment progresses (e.g. Balling 2008; Kuperman *et al.* in press). Such variables cannot be controlled between conditions, because they arise as each experimental session progresses, but they can be straightforwardly included in a regression analysis and thus statistically controlled. The variables used as examples here mainly come from the lexical decision literature, but similar variables are relevant in many experimental tasks, including eye-movement studies (Kuperman *et al.* in press), and regression designs provide a way of statistically controlling these sources of noise.

The possibilities of investigating different lexical and other variables are not unrestricted in regression analyses. More specifically, the problem of collinearity between the independent variables must be addressed, and it is necessary to assess whether effects are driven by outliers. These problems and possible remedies are discussed in section 3.2.

3. Analysing regression design data: linear mixed effects modelling

Various statistical choices exist for the analysis of data from a regression design. Of these, linear mixed effects modelling (Pinheiro & Bates 2000) seems to be the most powerful without being anti-conservative (see Baayen 2008: 282-299).[2] Mixed-effects regression analyses can be performed in the R environment for statistical computing, using the lme4-library of Bates *et al*. (2008). R is an open-source version of the commercial programme S-plus, and freely available on the Internet. R has the advantage of very flexible and powerful graphics, as exemplified in the figures included below. Some general advantages of mixed-effects analyses are discussed here; some are also valid for other regression techniques but discussed here with reference to mixed-effects models.

3.1 Fixed and random effects

Mixed-effects models include both random and fixed effects. The variables that are included as fixed effects in the models are either co-variates or factors. Co-variates are numerical variables such as frequency; factors are categorical variables with a fixed and low number of levels which exhaust the levels in the sampled population. A factor whose levels exhaust the population of morphologically complex words, for instance, could be called

[2] A statistical technique is conservative if it is likely to result in Type 2 error, not rejecting a false null hypothesis, while it is anti-conservative if it is likely to result in Type 1 error, rejecting a null hypothesis which is in fact correct. Type 2 errors are cases where an effect that is in fact significant comes out as non-significant in the analysis and is therefore not interpreted; therefore, Type 2 errors are the least serious. Type 1 errors, where a non-significant effect is shown to be significant and therefore interpreted as a real result, are far more problematic. Therefore, while a statistical technique should not be too conservative, it is more important that it is not anti-conservative. The linear mixed effects technique discussed here fulfils these requirements.

‘Complexity Type’ and have the levels ‘Inflected’, ‘Derived’ and ‘Compound’ (see Balling 2008). The fixed effects are repeatable: for instance, word frequency has a fixed number of values between the 0 and *n* for a corpus of size *n*, and several words can, in principle, be chosen for each of these values, making the variable repeatable. Similarly, the complexity type factor can be repeated since a practically unlimited number of words can be chosen for each category.

The variables included as random effects are not repeatable and do not have a fixed number of levels. Typical random effects in psycholinguistic studies are participant and item: both participants and items are in principle sampled randomly from the relevant populations, and each participant or item corresponds to a level of the variable which is not repeatable. Linear mixed effects models can include random variation in two ways: as random intercepts and as random slopes which are explained next.

In a regression model, the intercept is the value of the dependent variable (e.g. reaction time or gaze duration) in the hypothetical situation where all independent variables have a value of 0. In terms of the typical illustration of effects in regression analyses, the intercept is where the regression line crosses the y-axis. The fixed effects of the independent variables specify the change to the main intercept for the given factor or co-variate. Random intercepts are adjustments to the main intercept of the regression model for each level of the random factor. For a random factor such as ‘Participant’, the random intercepts adjust the main intercept for each participant, such that the model takes into account that relatively fast participants have lower intercepts (if the y-axis represents reaction time) while relatively slow participants have higher intercepts. A substantial amount of variation between participants is accounted for in this way, making the evaluation of the fixed effects in the regression analysis more reliable.

Random slopes are adjustments to the slope (the distance that the regression line covers on the x-axis for a given change on the y-axis) of a given co-variate for each level of the random effect. For instance, an analysis may include a main effect of frequency as a co-variate, but participants may differ in the effect of corpus frequency on their performance. This difference between participants can be included in the form of random slopes, adjustments to the general effect of word frequency for each participant (i.e. for every level of the random factor ‘Participant’).

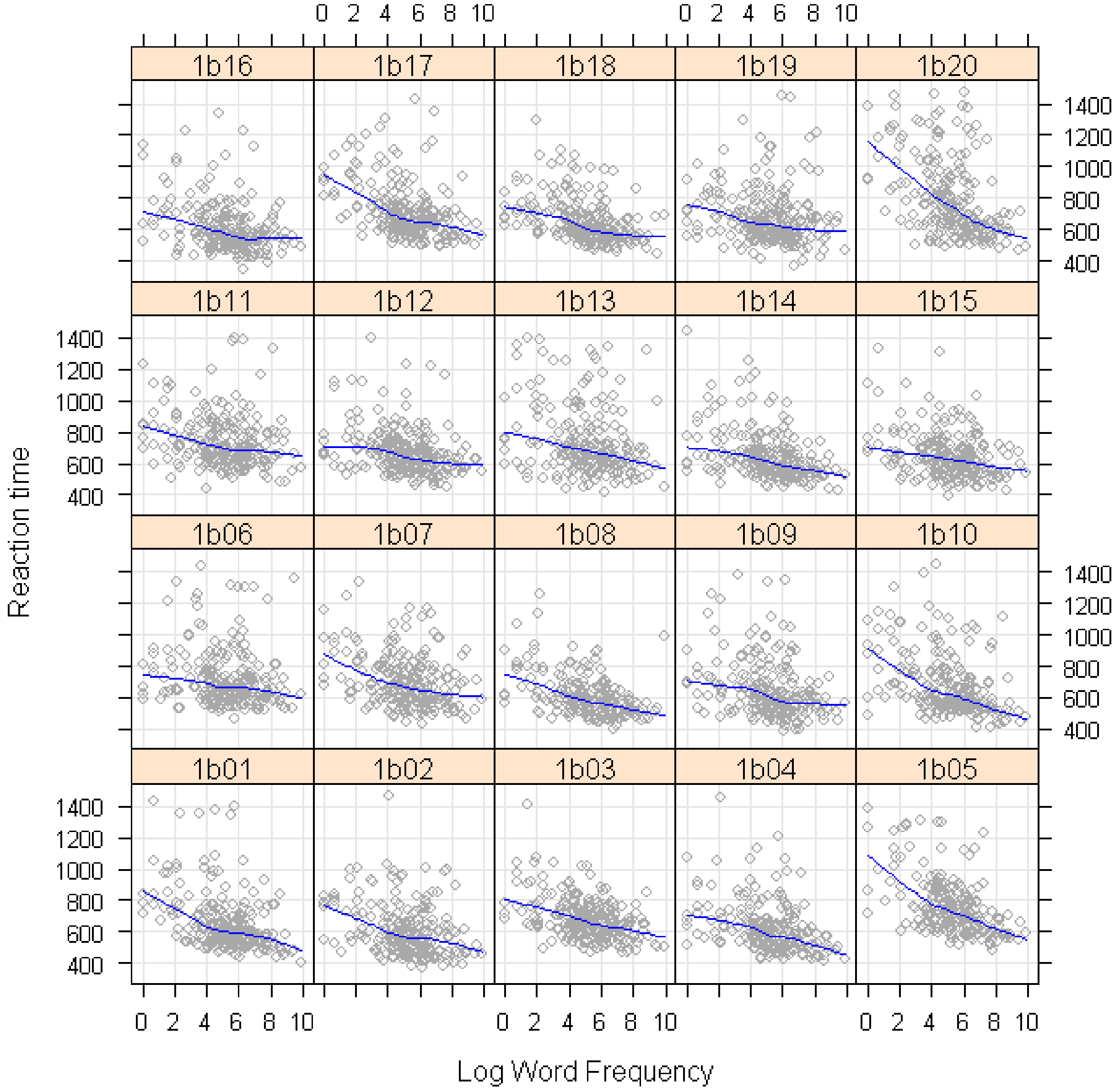

Figure 1. Conditioning plots of reaction time as a function of word frequency for each participant in the visual lexical decision experiment reported in chapter 5 of Balling (2008).

Figure 1 illustrates both random intercepts and slopes. The figure plots reaction time (the dependent variable) on the y-axis as a function of word frequency on the x-axis, for each participant in a visual lexical decision experiment reported in Balling (2008). The lines in each panel are regression lines which illustrate the correlation between RT and frequency for each individual participant. The different places in which the lines cross the y-axis correspond to the different intercepts for the participants, which are accounted for in the statistical model as random intercepts for each level of the non-repeatable factor ‘Participant’. The mean of the random intercepts should be 0 because the mean of the adjustments for all participants should correspond to the general intercept of the model.

Though the slopes of all the regression lines in Figure 1 are negative, corresponding to faster recognition of higher frequency words, the participants show clearly different profiles, with for instance participants 1b05 and, even more so, 1b20 showing much stronger frequency effects than the others. These differences can be modelled by random slopes for frequency for the participants.

The experiment on which Figure 1 is based is a particularly interesting example because the initial analysis suggested an overall difference in frequency effects between male and female participants, when the random slopes for frequency were not taken into account. Once this participant variation was accounted for by the inclusion of the random slopes, the sex difference disappeared. It seems that the sex difference appeared mainly because participant number 1b20, who showed the strongest frequency effect, happened to be male. In this way, some significant effects may turn out to be caused by individual differences. By contrast, other fixed effects may only emerge as significant if the difference between participants is accounted for.

This example shows a major advantage of being able to include both fixed and random effects, namely that it becomes possible to assess whether group differences are significant over and above differences between individual participants. Another advantage is that a single analysis including random effects of participant and item can replace the usual separate ANOVAs by participant and by item.

Whether or not a given random effect is included in a given mixed effects model is determined with the help of likelihood ratio tests. The model with the random effect in question is compared to a corresponding model without that random effect, in order to determine whether there is an increase in the goodness of the fit of the model to the data when the random effect is included in the model. This also determines whether such an increase in goodness of fit justifies the loss of degrees of freedom caused by the inclusion of the random effect. The p-value of the likelihood ratio test indicates whether the difference in goodness of fit between the models is significant; if the p-value is small – indicating reason for surprise in the increase in goodness of fit given that only one parameter has been added – the random effect is justified.

3.2 Collinearity and variable distribution

As mentioned, collinearity between the fixed effects in regression models can be a severe problem. If predictors such as various frequencies and length measures are highly correlated, collinearity is said to be high, and it becomes difficult or impossible to assess which of the variables have significant effects on the dependent variable and which do not. As a consequence, models become unstable if collinearity is very high.

In order to assess the collinearity, the condition number (kappa or κ) is calculated. The condition number is a measure of how close the correlation matrix is to being singular, i.e. completely collinear. High and potentially harmful collinearity is indicated by κ-values of 30 or above (Baayen 2008).

Many of the variables that are significant in word recognition studies tend to be highly correlated. This is one reason why items in factorial designs may become unrepresentative if one variable is dichotomised and the other variables must be controlled, as exemplified by the correlation between base and word frequency above. The same problem emerges as collinearity in regression analyses, including mixed-effects models. There are various remedies against harmful collinearity.

A first useful example is the reaction times on the previous trials which co-determine reaction times (four previous trials in the case of the experiment discussed in section 4). These previous-trial reaction times are necessarily highly correlated, leading to collinearity between the predictors. Importantly, the individual contributions of each of the four preceding reaction times do not matter greatly; what matters is that this variance in the data is controlled. When this is the case, a useful tool is Principal Components Analysis (PCA, see Baayen 2008). PCA transforms the variance of the original variables into principal components (PCs) which account for the same variance but are orthogonal to each other, i.e.

completely uncorrelated. PCA is particularly useful when the individual variables are not crucial, but the contribution of each variable can be roughly assessed by inspecting the so-called factor loadings of the original variables on the PCs, which show the contribution of each original variable to each PC.

A second useful tool for reducing collinearity is to take two highly correlated variables and construct simple linear regression models with one variable as a function of the other. For instance, in the experiment discussed below, word frequency and length in letters are correlated (longer words tend to be less frequent). In order to separate the effects of these two variables, a regression model is constructed with length in letters as a function of word frequency. For the overall analysis, the dependent variable of the linear regression model, in this case length, is replaced with the residuals of the regression model, i.e. the variance in the dependent variable length that is not accounted for by the independent variable word frequency in the simple model. The correlation between the original dependent variable and the new residualised one is generally large, indicating that the residuals are a reasonable replacement. The new residualised length variable is uncorrelated with the independent variable word frequency, so the two correlated variables in the original data set are replaced with two uncorrelated variables, and collinearity is reduced.

A third way of reducing collinearity, which is not used in the experiment discussed below, is the calculation of ratios of two correlated variables: for instance Baayen *et al.* (2006) operate with the ratio of spoken-to-written corpus frequency for their experimental items, thus including both frequencies without increasing collinearity.

In addition to making sure that collinearity between predictors is not harmful, the experimenter must consider the distribution of the variables. It is important to ascertain that the significance of the effects is not caused by outliers. In order to rule this out, the distributions of all significant variables should be examined visually. If there are clear outliers, for instance two or three words that are much longer or much more frequent than the remaining words, it should be investigated whether effects of length or frequency are only significant when these variables are included, by constructing a new model that excludes these outliers. Skewed distributions, which are typical of word frequency, can be normalised by logarithmic transformation.

3.3 Variable selection

In regression analyses, a large number of variables can enter into the analysis, some of which may turn out to be significant predictors of the observed data and some of which may not. It is debated whether non-significant predictors should be retained in the final analysis or not: On the one hand, the statistical model becomes simpler if the number of variables is reduced and the results become much clearer, both because the model is simpler and because the inclusion of non-significant variables can sometimes obscure the effects that are really significant and vice versa. On the other hand, there are potential theoretical problems with the validity of the final statistical model if a large number of models are constructed and one chosen (Harrell 2001). In exploratory data analysis, with no very clear-cut dichotomous hypotheses, the consensus seems to be that reducing the statistical model is acceptable (Baayen 2008), but this should be done in a careful and systematic way and the removal of variables should preferably be motivated by factors that are external to the statistical model, i.e. linguistic or psychological factors. In the experiment outlined in the next section, a model including all available predictors was constructed and then reduced in step-wise fashion, reaching a model which only included significant predictors.

4. An example experiment

In this section, I present results from a visual lexical decision experiment in Danish, to further exemplify some of the advantages of regression designs. Only broad results are presented, while technicalities and more detailed interpretation of the results are reported in chapter 5 of Balling (2008).

There were 20 adult participants (12 women and 8 men) in the experiment. The participants read a mixture of words and nonwords. For each token, participants were asked to determine as quickly and accurately as possible whether the token was a word they recognised. The words were a mixture of 109 morphologically simple words and 120 prefixed and suffixed derived forms in Danish (after the removal of six items that had error rates over 30%). Each nonword was based on a real word in the experiment, constructed by changing 1 to 3 letters in the real word but retaining the affixes on the complex words. The analysis includes only reaction times to the real words, i.e. the time it took participants to make a correct yes-decision when hearing a word.

A number of lexical variables were determined for the words; those that turned out to be significant are illustrated in Figure 2. Only fixed effects are depicted in Figure 2, but the model also included random intercepts for participant and item and random slopes for frequency by participant (as discussed in section 3.1 and illustrated in Figure 1). The number of significant effects illustrated in Figure 2 clearly shows the large amount of information that can be extracted from a single data set with the help of linear mixed-effects models. Each panel in Figure 2 shows the effect of a fixed-effect predictor while all other significant predictors are held constant at their medians, i.e. the effect of the given predictor when the other predictors are statistically controlled.

Turning first to the effects of experimental context, illustrated in the top row of Figure 2, there were significant effects of two out of four Principal Components (PCs) based on reaction times on the four previous items. Both PCs correlate negatively with the original variables, so the first two panels of Figure 2 reflect inhibitory effects of previous reaction times: reaction times on each trial were longer when the reaction times on previous trials were long. Reaction times were also longer when the previous response was an error: it seems that participants were aware of having made an error and slowed down subsequent responses as a consequence, as shown in the top right panel of Figure 2.

The next row shows the effect of more semantic variables, which were all extracted from a combination of the two Danish corpora *Korpus90* and *Korpus2000*. The word frequency in the first panel of the middle row is a logged frequency of the whole word, whether simple or complex, in the corpora; the effect is a standard facilitatory effect of word frequency.

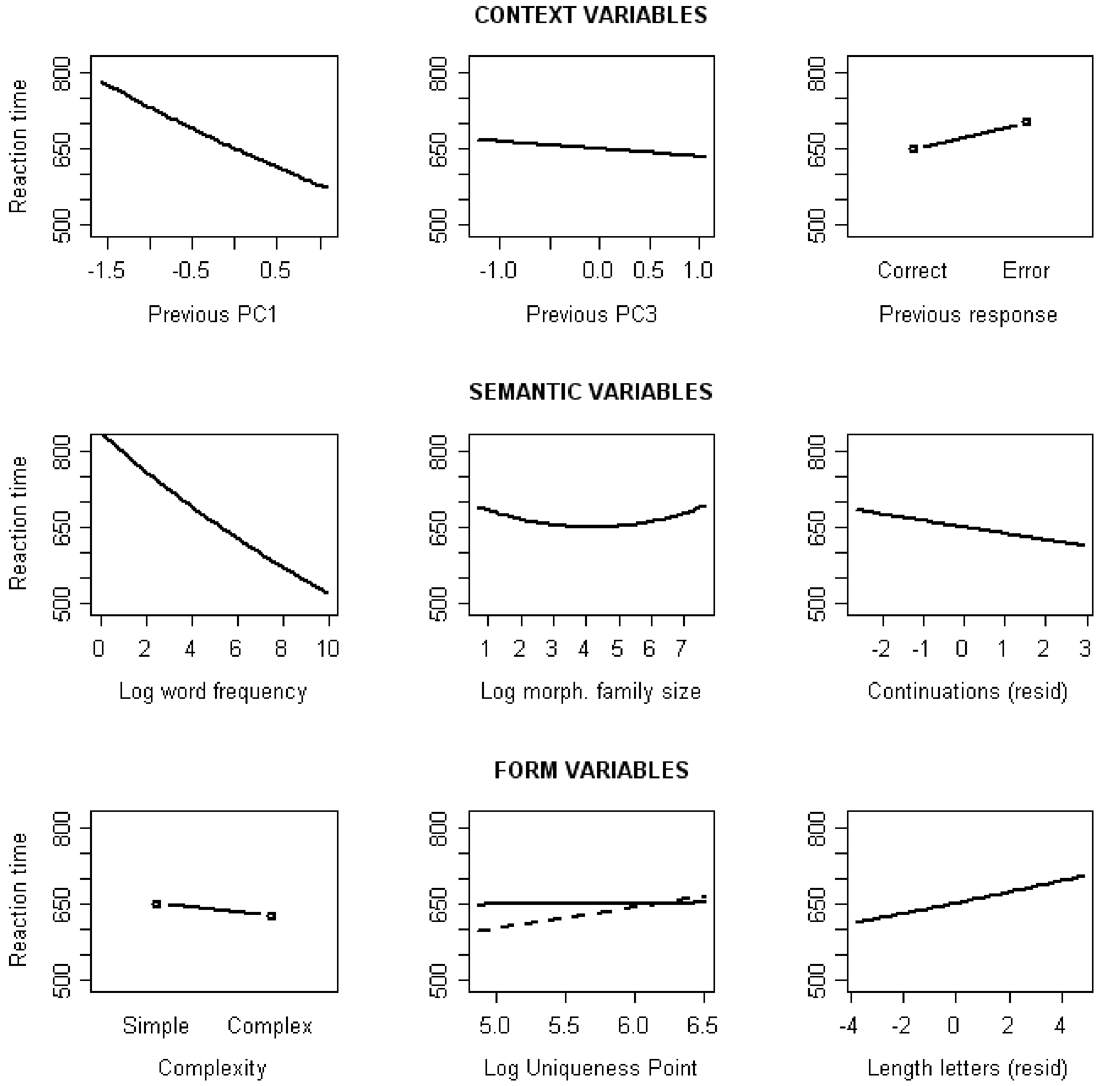

Figure 2. Results of a Danish visual lexical decision experiment, with the dependent variable reaction time on the y-axes and the significant independent variables on the x-axes. The first two panels in the top row show effects of Previous Components based on reaction times on the four previous trials. In the middle panel of the bottom row, the solid line represents simple words and the dashed line morphologically complex words. Continuations in the right panel of the middle row and length in letters in the final panel are both residualised from word frequency.

The next panel shows a non-linear effect which can also be included in the mixed-effects analysis. This is the effect of morphological family size, the number of derived forms and compounds in the corpus that contain the same base as the simple or complex word in question. For a word like *natur-lig* ('natural'), the morphological family consists of 724 words in *Korpus90* and *Korpus2000* that contain the stem *natur*. The effect illustrated is facilitatory up to about 35 family members (log value 3.6); this reflects that similar form-meaning mappings of the family members aid the recognition of the target word. However, if families are large, more than about 150 members (log value 5), family size becomes inhibitory, either because too many words are co-activated and compete with the target or because the large families are less semantically coherent than the smaller ones.

The right panel of the middle row shows a related effect, of the number of continuation forms. The continuation forms are morphologically related words that overlap with the target until word offset, i.e. for *naturlig* (‘natural’), words such as *naturligvis* (‘naturally’). The existence of many such continuation forms supports the lexicality and semantics of the target, making recognition easier.

The first panel of the bottom row illustrates an important result of the experiment, namely that, all other things being equal, morphologically complex words are recognised faster than simple words. The panel illustrates the difference between simple and complex words, when all the other factors are held constant at their medians. The complexity type factor interacted with UP1 as shown by the next panel; this means that the complexity advantage was not found for all UP1-values, as discussed below. Generally, complex words are longer and less frequent than simple words, but when these variables are statistically controlled, complex words tend to be recognised faster than simple words. It seems that the availability of both whole-word and morphemic sources of information support the recognition of the complex words.

Such comparisons between simple and complex words can also be included in factorial designs (e.g. Bertram *et al.* 1999), but in that case the simple and complex words must be matched between conditions, which means that the items chosen for the categories may not be typical of words in the language. In the regression design, by contrast, the differences in frequency and length can be statistically controlled, allowing the choice of typical examples of simple and complex words, which makes the results more generalisable.

The final variables are form variables: UP1 is the Uniqueness Point of the first constituent, the point in the word where it deviates from all other words in the language that share part of the onset of the word but do not share the same first constituent. This measure is distance in log msec from word onset to UP1 in the stimuli of a parallel auditory experiment. Surprisingly, this auditory measure had an effect on the response times to visually presented words, indicating some element of auditory recoding of the visual stimuli. The panel illustrates an interaction included in the regression model, between the Complexity and UP1: UP1 was only significant for the complex words, illustrated by the dotted line, not for the simple words, shown by the solid line. This interaction also shows that the complex words have a processing advantage only when UP1 is relatively early, as indicated by the fact that the lines meet at the end of the UP1-range. This means that the complexity advantage only emerges when the competition from unrelated lemmas is resolved relatively early.

Finally, length in letters (residualised from word frequency) is shown in the bottom right panel. This shows a straightforward inhibitory effect of length, with longer words being recognised more slowly than shorter words.

5. Summary and conclusions

The advantages of regression designs can be said to form two main clusters: Firstly, as discussed in section 2 and exemplified in section 4, regression designs are more informative and more powerful than most factorial designs, including more items and more representative items. Secondly, regression designs allow the statistical control of a wide variety of variables, in mixed-effects models including both fixed and random effects. This makes it possible to use more naturalistic tasks to study various language processes, including translation. The control of context variables should be particularly useful with experimental techniques such as eye-tracking and keystroke logging.

Although the discussion has focused on the way that I have used regression techniques to study word recognition, I hope – with the zeal of the recent convert – to have conveyed the usefulness of these techniques also to experimental linguists working on translation.

6. References

Baayen, R.H. 2008. *Exploratory Data Analysis: An Introduction to R for the Language Sciences*. Cambridge: Cambridge University Press.

Baayen, R.H., Feldman, L.B. & Schreuder, R. 2006. Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language* 55: 290-313.

Balling, L.W. 2008. *Morphology and its Functionality in the Danish Mental Lexicon*. PhD thesis, University of Aarhus.

Balling, L.W. & Baayen, R.H. in press. Morphological effects in auditory word recognition: evidence from Danish. *Language and Cognitive Processes*.

Balota, D., Cortese, M., Sergent-Marshall, S., Spieler, D. & Yap, M. 2004. Visual Word Recognition of Single-Syllable Words. *Journal of Experimental Psychology: General* 133: 283-316.

Bates, D.M. Maechler, M. & Dai, B. lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-20. http://lme4.r-forge.r-project.org/. Accessed June 27, 2008.

Bertram, R., Laine, M. & Karvinen, K. 1999. The Interplay of Word Formation Type, Affixal Homonymy, and Productivity in Lexical Processing: Evidence from a Morphologically Rich Language. *Journal of Psycholinguistic Research* 28: 213-226.

Cohen, J. 1983. The cost of dichotomization. *Applied Psychological Measurement* 7: 249-254.

Cutler, A. 1981. Making up materials is a confounded nuisance, or: will we be able to run any psycholinguistic experiments at all in 1990? *Cognition* 10: 65-70.

Dalgaard, P. 2002. *Introductory Statistics with R*. New York: Springer.

De Vaan, L., Schreuder, R. & Baayen, R.H. 2007. Regular morphologically complex neologisms leave detectable traces in the mental lexicon. *The Mental Lexicon* 2: 1-24.

Diependaele, K., Sandra, D. & Grainger, J. 2007. Masked priming with Dutch compounds: is the whole greater than the sum of the parts? Paper presented at the 5th International Workshop on Morphological Processing, Marseille, June 2007.

Harrell, F.E. 2001. *Regression Modeling Strategies With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York: Springer.

Kuperman, V., Bertram, R. & Baayen, R.H. in press. Morphological Dynamics in Compound Processing. *Language and Cognitive Processing*.

MacCallum, R., Zhang, S., Preacher, K. & Rucker, D. 2002. On the Practice of Dichotomization of Quantitative Variables. *Psychological Methods* 7: 19-40.

New, B., Brysbaert, M., Veronis, J. & Pallier, C. 2007. The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics* 28: 661-677.

Pinheiro, J. & Bates, D.M. 2000. *Mixed-Effects Models in S and S-PLUS*. New York: Springer.

Taft, M. 2004. Morphological decomposition and the reverse base frequency effect. *Quarterly Journal of Experimental Psychology* 57: 745-765.