Constituent family size and compound stress assignment in English

Ingo Plag & Gero Kunter

June 28, 2009

Abstract

There have been claims in the literature (e.g. Marchand 1969, Bell 2008) that variable stress assignment to English compounds is influenced by the size of the constituent families, i.e. the number of compounds that share the same left or the same right constituent. This paper tests this claim empirically on the basis of a large amount of data from three different corpora. The expected effects can be found in some form (and to varying degrees) in all three sources, alongside of other effects that have been held to be responsible for compound stress assignment. The results can be interpreted as evidence against deterministic rule-based approaches to compound stress and lend independent evidence to a model in which compound stress assignment emerges from the lexicon.[1]

1 Introduction

It has often been claimed that English compounds tend to have a stress pattern that is different from that of phrases. This is especially true for nominal compounds, which is the class of compounds that is most productive. While phrases tend to be stressed phrase-finally, compounds tend to be stressed on the first element. This systematic difference is captured in the so-called nuclear stress rule and compound stress rule (Chomsky & Halle 1968:17). While the compound stress rule apparently makes correct predictions for a large proportion of nominal compounds, it has been pointed out that there are also numerous exceptions to the proposed rule (cf. Jespersen 1909:153ff, Kingdon 1958, Schmerling 1971, Fudge 1984, Ladd 1984, Liberman & Sproat 1992, Sproat 1994, Bauer 1998, Olsen 2000, 2001, Giegerich 2004). In other words, there are structures that are stressed on the right-hand side in spite of the fact that these structures are regarded as compounds by most

[1]The authors are especially grateful to Melanie Bell for raising the issue of family sizes, and for stimulating discussions with her. We also thank Sabine Arndt-Lappe, Kristina Kösling, Mareile Schramm, Linda Zirkel and the editor of this special issue, Susan Olsen, for their feedback on an earlier version. Material from this paper was presented at the University of Edinburgh and the University of Aaarhus in February and March 2009. Many thanks go to these audiences for their constructive questions and comments. Special thanks also to Heinz Giegerich for critical discussion, and to Harald Baayen for the discussion of family matters (linguistic and non-linguistic) and for his latest R code. This work was made possible by two grants from the *Deutsche Forschungsgemeinschaft* (PL151/5-1, PL 151/5-3), which we gratefully acknowledge.

analysts. Some of these forms are listed in (1). The most prominent syllable is marked by an acute accent on the vowel.

(1) Examples of rightward-stressed compounds
geologist-astrónomer, apple píe, scholar-áctivist, apricot crúmble, Michigan hóspital, Madison Ávenue, Boston márathon, Penny Láne, summer níght, aluminum fóil, spring bréak, silk tíe

In view of this situation, the obvious question is how we can account for this variability in stress assignment to noun-noun constructs. The literature provides numerous hypotheses (see, for example, Plag et al. 2008 for an overview), but until a few years ago systematic empirical work on the problem was lacking. Recent experimental and corpus studies have shown that a variety of factors influence compound stress assigment, with lexicalization, the distinction between argument-head and modifier-head compounds, the morphology of the head, and the semantic properties of the compounds having significant effects (e.g. Plag 2006, Plag et al. 2007, 2008). The results of these studies seriously challenge traditional rule-based approaches to compound stress à la Chomsky & Halle's, which has led researchers to investigate an alternative approach, in which stress is assigned on the basis of analogy to similar compounds in the lexicon.

This idea has been around for quite some time (see, for example, Schmerling 1971) and in its simplest form says that compounds with the same right or left constituent tend to exhibit the same type of stress. In other words, stress assignment should be largely due to the effect of the 'constituent family', i.e. the set of compounds that share the first, or the second, constituent with a given compound. If there is the tendency of a given constituent family to favor a particular kind of stress, for example rightward stress, then the compound in question will also tend to have that kind of stress. This approach has recently been tested empirically using exemplar-based modeling (Lappe & Plag 2008) and regression analysis (Plag 2009). In both types of analysis family bias emerged as a strong predictor. While Lappe & Plag's (2008) models perform best with constituent information as the only predictor, Plag (2009) shows that the stress bias in the constituent families is significant alongside other significant predictors (semantics and lexicalization in particular).

There is yet another, related hypothesis about compound stress assignment around, namely one that focuses on the size of the constituent families. Bell (2008) has recently proposed that constituent family size has an influence on compound stress assignment (see also Marchand 1969 for an earlier, similar approach). She puts forward the idea that there is a negative correlation between the family size of a compound constituent and the proportion of stress on this constituent. The larger the right constituent family, the smaller the proportion of right-stressed compounds among the compounds with that right constituent. The larger the left family, the smaller the number of left-stressed compounds among the compounds with this left constituent. The underlying reason for the relationship between family size and stress would be the fact that with increasing type frequency, the given constituent becomes more predictable, and hence less informative (vis-à-vis the other constituent), which then leads to stress on the more informative constituent.

This approach is of special theoretical interest, since, unlike rule-based approaches to compound stress, it essentially assumes that compound stress emerges from the lexicon. Relevant information for the assignment of stress to a given compound is retrieved from

related forms in the mental lexicon, and is not computed by some abstract rule mechanism in the grammar. From numerous psycholinguistic studies, we know that lexical processing depends in part on the amount of information carried by words, which are defined by the accumulated knowledge of words and their paradigmatic and syntagmatic connectivity in the mental lexicon. Part of that connectivity are morphological families, i.e. the sets of words that contain the same morphological constituents (see, e.g., Baayen et al. 2006, Moscoso del Prado Martín et al. 2004, Kuperman et al. 2009, Milin et al. 2009 for detailed studies). The frequency with which a given constituent occurs in combinations with other constituents is thus a measure of the informativeness of the constituents. A constituent that occurs infrequently carries more information than one that occurs more frequently. And if we follow the assumption that informativeness has influence on the distribution of stress in multi-word sequences (see, for example, Ladd 1984 for discussion and examples), with more informative constituents tending to attract stress, there should be a relation in compounds between constituent family sizes and stress assignment. Thus, from a theoretical perspective, finding an effect of constituent family size on stress assignment would provide independent evidence for an approach in which compound stress assignment emerges from the lexicon, and against a deterministic rule-based approach.

After a more detailed discussion of the different approaches to compound stress assignment, and an explanation of our methodology, we will first test the family size effect in a regression analysis with only the family sizes as predictors of leftward and rightward stress. This will be followed by a multivariate analysis that also takes analogical, semantic, structural and lexicalization effects into account, to see whether an effect of family size survives in a more complex model.

2 Hypotheses about compound stress assignment

Roughly speaking, four types of approach have been taken to account for the puzzling facts of variable noun-noun stress. The first one is what Plag (2006) has called the 'structural hypothesis'. In its most recent formulation, Giegerich (2004) proposes that, due to the order of elements, complement-head structures like *trúck driver* cannot be syntactic phrases, hence must be compounds, hence are left-stressed. Modifier-head structures such as *steel brídge* display the same word order as corresponding modifier-head phrases (cf. *wooden brídge*), hence are syntactic structures and regularly rightward-stressed. This means, however, that many existing modifier-head structures are in fact not stressed in the predicted way, since they are left-stressed (e.g. *ópera glasses, táble cloth*). Such aberrant behavior, is, according to Giegerich, the result of lexicalization. Recent large-scale empirical studies investigating the predictions of the structural hypothesis have all provided evidence for either a weak effect of argument structure (Plag 2006, Plag et al. 2007), or for no effect at all, if other variables are taken into account (Plag et al. 2008, Lappe & Plag 2008, Plag 2009). Plag et al. (2007, 2008) also found (weak) lexicalization effects in the expected direction.

The second approach makes use of the semantic characteristics of compounds. It has been argued that words with rightward stress such as those in (1) above are systematic exceptions to the compound stress rule (e.g. Sampson 1980, Fudge 1984, Ladd 1984, Liberman & Sproat 1992, Sproat 1994, Olsen 2000, 2001, Spencer 2003). Although these

authors differ slightly in details of their respective approaches, they all argue that rightward stress is restricted to only a limited number of more or less well-defined types of meaning categories and relationships. Pertinent examples are copulative compounds like *geologist-astrónomer* and *scholar-áctivist* (cf. Plag 2003:146), which are uncontroversially considered to be regularly rightward-stressed. Other meaning relationships that are often, if not typically, accompanied by rightward stress are temporal or locative (e.g. *summer níght, Boston márathon*), or causative, usually paraphrased as 'made of' (as in *aluminum fóil, silk tíe*) or 'created by' (as in *Shakespeare sónnet, a Mahler sýmphony*). However, there are only a few systematic empirical studies available that investigate the role of semantics in variable compound stress assignment. While Sproat (1994) and Plag (2006) do not find the predicted effects, Plag et al. (2007) tested many more semantic relations and found many effects, some of them new, and some of them predicted by the literature. However, not all of the effects predicted by the literature were manifest in their data, and large parts of the data were ill-behaved. A similar picture emerges from the study of Plag et al. (2008). Although a number of robust significant semantic effects were found, these effects were far from categorical and large parts of the data were unaccounted for.

Under the third type of approach, the analogical one, stress assignment is generally based on analogy to existing NN constructions in the mental lexicon. Plag (2003:139) mentions the textbook examples of *street* vs. *avenue* compounds as a clear case of analogy. All street names involving *street* as their right-hand constituent, pattern alike in having leftward stress (e.g. *Óxford Street, Máin Street, Fóurth Street*), while all combinations with, for example, *avenue* as right-hand member pattern alike in having rightward stress (e.g. *Fifth Ávenue, Madison Ávenue*). Along similar lines, Spencer (2003:331) proposes that "stress patterns are in many cases determined by (admittedly vague) semantic 'constructions' defined over collections of similar lexical entries." In a similar vein, Ladd (1984) proposes a destressing account of compound stress which would explain the analogical effects triggered by the same rightward members as basically semantico-pragmatic effects. Schmerling (1971:56) is an early advocate of an analogical approach, arguing that many compounds choose their stress pattern in analogy to combinations that have the same head, i.e. rightward member. Liberman & Sproat (1992) extend this proposal to both constituents of the compound. Overall, all the above authors leave it unclear how far such an analogical approach can reach.

The effect of analogy in stress assignment has been tested empirically in some very recent studies. In his experimental investigation using novel compounds, Plag (2006) found a very robust effect of the right constituent on the stress pattern of a given compound. In particular, compounds with *symphony* as right constituent behave consistently differently from compounds with *sonata* or *opera* as right constituents, irrespective of the semantic relation expressed by the compound. While this study did provide evidence for an effect of the right constituent family, the potential effect of the left constituent family was not tested. The effects of analogy were more thoroughly investigated in three corpus-based studies: Plag et al. (2007) looked at data from CELEX, Lappe & Plag (2008) present exemplar-based models for data from CELEX and from the Boston University Radio Speech Corpus. Plag (2009) is a regression study using data from Teschner & Whitley (2004), CELEX and from the Boston University Radio Speech Corpus. All of these studies provide robust evidence for a constituent family effect in compound stress assignment.

A fourth approach to compound stress assignment makes reference to the number of

compounds in a given constituent family. Marchand (1969:23-4) already claimed that "the frequent occurrence of a word as a second constituent is apt to give compound character [i.e. left stress] to combinations with such words". In other words, compounds with a large right family should be left-stressed. Bell (2008) recently extended Marchand's hypothesis to compounds with large left constituent families, making the additional prediction that a large left constituent family should go together with rightward stress. Overall, Bell's and Marchand's hypotheses boil down to a negative correlation between family size and stress. The larger the right family, the smaller the proportion of right-stressed compounds among the compounds with that right constituent. The larger the left family, the smaller the number of left-stressed compounds among the compounds with this left constituent.

Ladd (1984) applies a related kind of reasoning to explain the constrast between left-stressed compounds headed by *street*, and right-stressed compounds headed by *avenue*, *boulevard*, or *road*. Ladd (1984:260) argues that "we do get less information about the category of things being named from *Street* than from any of the others, and hence more from the attribute [i.e. the left constituent]; this is more typical of ordinary compounds, and is exactly what is signalled by the stress pattern." In essence, then, the underlying reason for the negative correlation between family size and stress would be the fact that with increasing type frequency, the given constituent becomes more predictable, and hence less informative (vis-à-vis the other constituent), which then leads to stress on the more informative constituent.

In a production experiment with native speakers, Bell (2008) finds evidence in favor of her hypothesis. For example, in her data (taken from the BNC Demographic Corpus) there is a large left family for *world* with a majority of right-stressed compounds (as in *world chámpion, world cóuncil, world cúp, world léader*), and the opposite effect for the very frequent right constituent *line*, as in *clóthes line, hélp line, prodúction line, trável line*. Bell also points out that there are clear counterexamples, such as the right constituent *pie*, which has a large family, but all pertinent compounds (with the exception of lexicalized and opaque *hóney-pie*) are right-stressed (cf. *apple píe, fish píe, lemon píe, meringue píe, mince píe*, etc.). Obviously, there seem to be competing forces at work, in this case perhaps the constituent family stress bias, or the semantic relation ('N1 is an ingredient of N2'), which is constant across the family and usually goes together with rightward stress.

It is thus unclear how far the family size approach can take us in explaining variable compound stress in English. Furthermore, it is unclear how the supposed effect would interact with other factors that influence compound stress assignment. Is the family size effect stronger, weaker, or not found at all? In the rest of the paper we will test Bell's hypothesis with the help of regression analyses, using a large amount of independently gathered data from three corpora.

We will first test the family size hypothesis with FAMILY SIZE as the only predictor variable and then factor in all other variables that have been found to influence compound stress assignment, to see if the family size effect survives as an independent predictor among other significant predictors. We use multiple regression as a statistical technique because it is especially well suited to test the influence of many variables at a time, namely by calculating the effect of one variable while holding all other variables constant (see, for example, Baayen (2008) for an introduction to multiple regression in linguistic studies).

3 Methodology

3.1 The corpora

We took the data from three different sources, to be described in more detail below: Teschner & Whitley (2004), the English part of the CELEX lexical database and the Boston University Radio Speech Corpus. The latter two sources have been employed in previous studies of compound stress (Plag et al. 2007, 2008, Lappe & Plag 2007, 2008, Plag 2009). We used the same data sets as those authors, with the Boston Corpus contributing an initial set of 4353 tokens of noun–noun constructs, representing 2450 word types, and CELEX providing 4491 types. The data in Teschner & Whitley (2004) amount to 2583 types overall. For illustration of our data sets, a random sample of 100 compounds from each data set is given in appendix 1.

For the Teschner & Whitley (2004) compounds, stress position and constituents were the only types of information available to us. Hence for this data set, we will only be able to test the constituent family bias effect and the constituent family size effect, but no other potential effects. For the other two corpora we also had at our disposal the codings of the semantic and structural categories, as used in the above-mentioned studies by Plag and colleagues, enabling us to look at the simultaneous effects of other variables.

Teschner & Whitley (2004) is a textbook for teaching pronunciation, and it comes with a CD-ROM on which there are, among other things, lists of words and phrases with their respective stress patterns, as gleaned from a Spanish-English dictionary (Carvajal & Horwood 1996). From these lists we manually extracted all items that consisted of two (and only two) adjacent nouns. Teschner & Whitley use three categories of compound stress, i.e. left, right, and level stress. There is some confusion in the literature about how many different stress patterns should be assumed, and whether, when more than two patterns are used, these levels refer to the phonetic or the phonological level. In recent work on the phonetic implementation of compound stress in English (e.g. Kunter & Plag 2007, Kunter 2009), it was shown that rightward stress manifests itselfs mostly in a more or less level pitch and intensity. It is this level pitch and intensity that gives rise to descriptions of (phonologically) rightward stress as 'level' or 'even'. We have therefore collapsed Teschner & Whitley's 396 level-stressed items and the 36 right-stressed items into one category, with the stress value `right`. We will refer to this database as 'T&W' for short.

The English part of CELEX has been compiled on the basis of dictionary data and text corpus data. The dictionary data come from the *Oxford Advanced Learner's Dictionary* (41,000 lemmata) and from the *Longman Dictionary of Contemporary English* (53,000 lemmata). The text corpus data come from the COBUILD corpus, which contains 17.9 million word tokens. 92 percent of the word types attested in COBUILD were incorporated into CELEX. The frequency information given in CELEX is based on the COBUILD frequencies. Overall, CELEX contains lexical information about 52,446 lemmata, which represent 160,594 word forms. From the set of lemmata all words were selected that had two (and only two) nouns as their immediate morphological constituents. This gave us a set of 4491 NN compounds, the vast majority of which come from the two dictionaries (see Plag et al. 2007 for detailed discussion). Each of these compounds was coded for the pertinent semantic and structural categories.

The Boston University Radio Speech Corpus was collected primarily to support research in text-to-speech synthesis, particularly the generation of prosodic patterns. The corpus consists of professionally read radio news data and includes speech from seven (four male, three female) FM radio news announcers associated with WBUR, a public radio station. The main radio news portion of the corpus consists of over seven hours of news stories recorded in the WBUR radio studio during broadcasts over a two-year period. In addition, the announcers were also recorded in a laboratory at Boston University. For the latter recordings (the so-called 'lab news'), the announcers read a total of 4 stories from the radio news portion. The announcers were first asked to read the stories in their non-radio style and then, 30 minutes later, to read the same stories in their radio style. Each story read by an announcer was digitized in paragraph size units, which typically include several sentences. The orthographic transcripts were generated by hand by the corpus compilers.

The Boston Corpus is especially well suited for testing hypotheses on compound stress assignment for at least three reasons. First, due to the topics covered in the news texts a large number of compounds are present in the corpus. Second, the corpus provides high-quality recordings, which is very useful for perceptual and acoustic analyses. Third, given that the speakers were trained news announcers they produce relatively standard, error-free speech. From all texts Plag and colleagues manually extracted all sequences consisting of two (and only two) adjacent nouns, one of which, or which together, functioned as the head of a noun phrase. From this set proper names such as *Barney Frank* and constructions with an appositive modifier, such as *Governor Dukakis* were eliminated. The final set of of noun–noun constructs obtained in this way contains 4353 tokens, representing 2450 word types. Each of these compounds was coded for the pertinent semantic and structural categories.

The data from the Boston Corpus present us with two different options. One can analyze tokens, or one can generalize over tokens and provide a type-based analysis. For the present paper we resorted to a type-based analysis to be better able to compare the results across corpora.

While T&W and CELEX give us (type-based) categorical stress information (either 'left' or 'right'), the data from the Boston Corpus are speech data for which categorical stress information is not provided. Although it has been shown that it is possible to model the perception of stress for this data set based on acoustic parameters (see Kunter & Plag 2007, Plag et al. 2008, Kunter 2009), preliminary explorations using automatic classification showed that such an automatic procedure still had an error margin that runs the danger of being detrimental for the present analyses. It was therefore decided to have two trained listeners rate all tokens on the basis of their acoustic impression. Both listeners had phonetic training and held a degree in English linguistics. Only those compounds entered the analysis on which both raters agreed.

A type-based analysis presents the additional problem that in those cases where different tokens of the same type vary in their stress pattern, a decision in one or the other direction had to be taken for this type. In such cases majority decisions were taken in order to decide how a given type would be stressed. If the number of tokens with rightward stress was equal to the number of tokens with leftward stress, this compound was excluded from the analysis (this happened only once).[2]

[2]Note that all type-based analyses ignore the problem of variability within types (see Bauer 1983:103,

3.2 Determining constituent family sizes

In order to test the effect of constituent family size in compound stress assignment, one first has to determine the left and right constituent families for each compound. To do so, we proceeded as follows. For each compound we first established two sets of compounds as they occur in its respective database. The first set, the so-called left constituent family, is the set of compounds that share the left constituent with the given compound. The second set of compounds, the so-called right constituent family, contains all compounds from the respective corpus that share the right constituent with the compound in question. Since we are interested in the effect of the right or left constituent family, we selected for further analysis only those compounds that had at least one other member in each of their two families. This led to a considerable reduction in the size of the data, but the remaining data setsare still large enough to allow serious testing (T&W: $N = 782$ types, CELEX: $N = 2638$ types, Boston Corpus: $N = 536$ types). Appendix 2 illustrates some constituent families (listed with their stress biases, as discussed in section 5).

Table 1 gives the distributions of leftward and rightward stresses for all corpora, with the proportion of left-stressed items in the last row. The proportion of leftward stresses varies across corpora. For dictionary data the proportion of leftward stresses seems generally higher than for news texts. For example, Sproat (1994:88) counts 70 percent leftward stresses in his Associated Press newswire corpus, which is almost the same proportion of left stresses as in the Boston Corpus news texts.

Table 1: Distribution of stresses across corpora.

	T&W	CELEX	Boston Corpus (types)
leftward stress	700	2483	359
rightward stress	82	155	176
percent leftward stresses	89.5	94.1	67.1

For each compound in each corpus the size of its left constituent family and the size of its right constituent family was computed. To give an example from the Boston Corpus, consider the compound *advertising business*, which has a left family with six other members (*advertising agency, advertising battle, advertising commentator, advertising costs, advertising days, advertising dollars*), i.e. 7 members overall, and a right family with two other members (*biotechnology business, computer business*), i.e. 3 members overall. Overall, and across corpora, most families are quite small. For example, in the Teschner & Whitley corpus 60.2 percent of the 782 compounds have left constituent families with only one or two other members, and 63.6 percent have right constituent families with only one or two other members. However, one also finds compounds with families of up to eleven members. For the two other corpora a similar preponderance of small families can be observed, but we also find larger, and sometimes even extraordinarily large, families in our data sets (up to 76 members in CELEX, up to 31 members in the

Plag et al. 2008, Kunter 2009 for discussion), but a proper treatment of this kind of variation is beyond the scope of the present paper.

Boston Corpus).[3]

For the statistical analysis we used logistic regression models to estimate the effect of the two predictor variables (i.e. LEFT FAMILY SIZE and RIGHT FAMILY SIZE). We applied a log transformation to the family size in order to transposed the heavily skewed distributions to a more or less normal distribution. Furthermore, for the models that featured interactions between family size and family bias (see section 5), the family sizes were scaled after log-transformation to reduce the danger of collinearity (cf. Jaccard et al. 1990). To return to our predictions, if family sizes play a role, one should find significant effects of family sizes in our regression models.

For the statistical analysis the statistical package R (R Development Core Team, 2007) was used. The final models to be presented have been obtained using the standard simplification procedures, according to which insignificant predictors are eliminated in a step-wise evaluation process (e.g. Baayen 2008). To answer the question of whether several different factors have independent effects, it is essential to control potential collinearity effects. All the models presented in this paper have been tested for collinearity using variance inflation factors (VIFs). Variance inflation factors indicate the extent to which the correlation of a given variable with other variables in the model inflates the standard error of the regression coefficient of that variable (e.g. Stine 1995, Allison & Allison 1999, O'Brien 2007, Kutner et al. 2005). The final models presented below generally show no danger of collinearity, with almost all VIFs having values below 2.5, and only very few between 6 and 7. A maximum value in excess of 10 is normally taken as an indication that multicollinearity may be unduely influencing the model (Kutner et al. 2005:409). We nevertheless flag out all VIFs that exceed the very conservative threshold of 2.5. To check whether our models overfit the data, and to substantiate the robustness of our predictors, we also ran bootstrap validations for all final models (e.g. Baayen 2008:193-195). In all simulations all predictors remained in the models, and only very small corrections of R^2 occured.

3.3 Hypotheses and predictions

The family size hypothesis makes the following predictions:

(2)
- •Prediction 1: The larger the left constituent family of a given compound, the smaller the chances of leftward stress.
- •Prediction 2: The larger the right constituent family of a given compound, the smaller the chances of rightward stress.
- •Prediction 3: The family size is an independent predictor of compound stress, alongside other predictors.

[3]The largest family of the Boston Corpus, for example, is the one with the left constituent *state*: *state administration, state aid, state authority, state benefit, state budget, state college, state company, state constitution, state court, state firm, state fund, state funding, state house, state job, state law, state legislator, state money, state office, state official, state park, state policy, state prison, state program, state property, state revenue, state road, state senator, state service, state spending, state university, state worker*. In accordance with the predictions put forward at the end of section 3, this family has a strong bias towards rightward stress, with only 3 out of the 31 compounds having leftward stress (*state company, state house, state official*).

Given that our regression models explicitly predict the probability of only one outcome (i.e. either leftward or rightward stress), we need to translate these predictions into predictions that make reference to only one type of stress. Using the probability of rightward stress as the value to be predicted, we can reformulate the predictions as follows:

(3)
- Prediction 1: The larger the left constituent family of a given compound, the higher the probability of rightward stress in that family.
- Prediction 2: The larger the right constituent family of a given compound, the lower the probability of rightward stress in that family.
- Prediction 3: FAMILY SIZE is an independent predictor of compound stress, alongside other predictors.

In order to test prediction 3, we present models that include other known significant predictors (i.e. structural, semantic, and analogical ones). If the prediction is correct, family size should emerge as significant even in those models that incorporate also other factors that influence the distribution of stress in English compounds.

4 Results 1: The constituent family size effect

4.1 Teschner & Whitley (2004): Constituent family size alone

The barplot in figure 1 shows the distribution of leftward and rightward stresses according to the family size of the left and right constituent families.[4]

[4]Missing bars indicate family sizes that do not occur in the data. We have included those a empty bars to give full documentation about the distribution of family sizes.

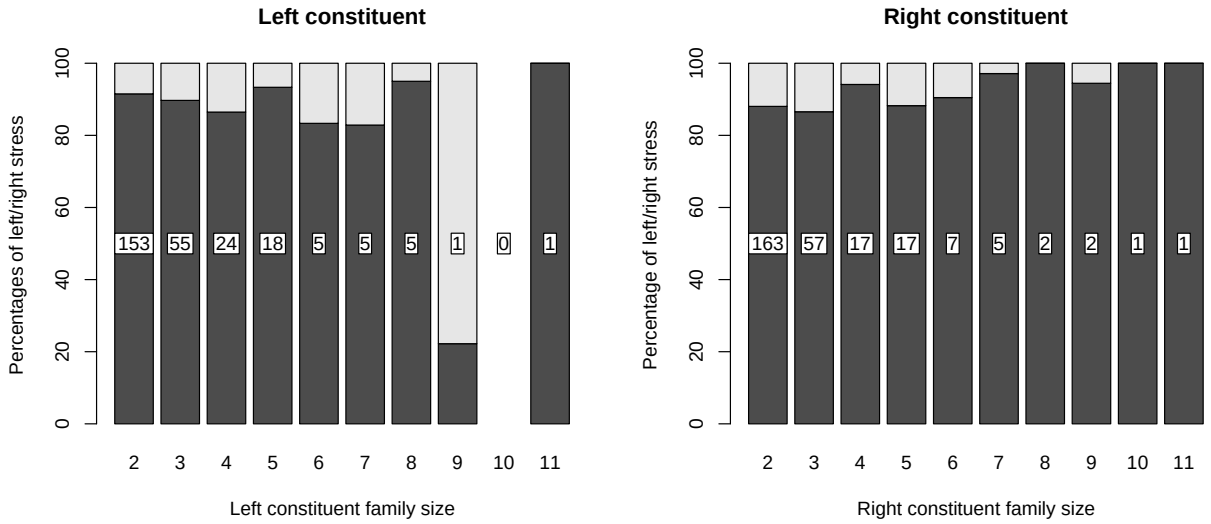

Figure 1: Stress patterns by left and right constituent family size, T&W data. The light portions of the bars indicate right stresses, the black portions left stresses. The figures inside the bars give the number of observations, i.e. the number of compounds with that familiy size.

We can see from the two distributions that there seem indeed to be tendencies in the predicted directions, i.e. a growing proportion of right stresses for growing left family size and a decreasing proportion of right stresses with increasing right family size. There are, however, also some family sizes that do not follow the general trend (e.g. left families with 5, 8 or 11 members, or right families with 4 or 9 members). Let us see whether the observed trends are statistically significant.

We fitted a logistic regression model with STRESS POSITION as the dependent variable and LEFT CONSTITUENT FAMILY SIZE and RIGHT CONSTITUENT FAMILY SIZE as the two predictor variables. The result is documented in table 2. Only the effect for the right family size reaches significance, while LEFT FAMILY SIZE is only marginally significant. A look at the coefficients of the regression models reveals that the two effects work in the expected directions. Negative coefficients in the model indicate a tendency towards left stress (as shown by the negative coefficient of the intercept, which represents the baseline, i.e. left stress). The positive coefficient for the left family size means that with increasing left family size, the tendency towards right stress becomes stronger. The opposite is true for the right family size. With increasing right family size, the tendency towards right stress becomes weaker. There was no significant interaction.

The final model, from which LEFT FAMILY SIZE has been removed in the usual model simplification process, is documented in table 3.

Table 2: Logistic regression model with left and right constituent family sizes as predictors, T&W data, N = 782.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.9788	0.4211	-4.70	0.0000
left family size	0.4253	0.2320	1.83	0.0668
right family size	-0.6189	0.2635	-2.35	0.0188

Table 3: Final logistic regression model with only family size as predictors, T&W data, N = 782.

	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	-1.4546	0.2999	-4.85	0.0000
right family size	-0.6235	0.2626	-2.37	0.0176

There is a negative coefficient for RIGHT FAMILY SIZE, which indicates an effect towards leftward stress, i.e. in the expected direction. The predictive power of the model with only the right family is very small ($C = 0.568$, model $p = 0.0133$)

Although these results show a trend according to the above predictions, we have to state that family size does not have a very strong influence on compound stress assignment in this data set. First, one of the two family sizes is only marginally significant, and second, the final model, which has the right family as the only remaining significant predictor, does not perform well as a classifier.

4.2 CELEX: Constituent family size alone

The barplot in figure 2 shows the distribution of leftward and rightward stresses according to the family size of the left and right constituent families. There is only one family which is not included in the graph, namely the right family of *man*, which has 71 members. This family has been removed from the data set as an outlier.

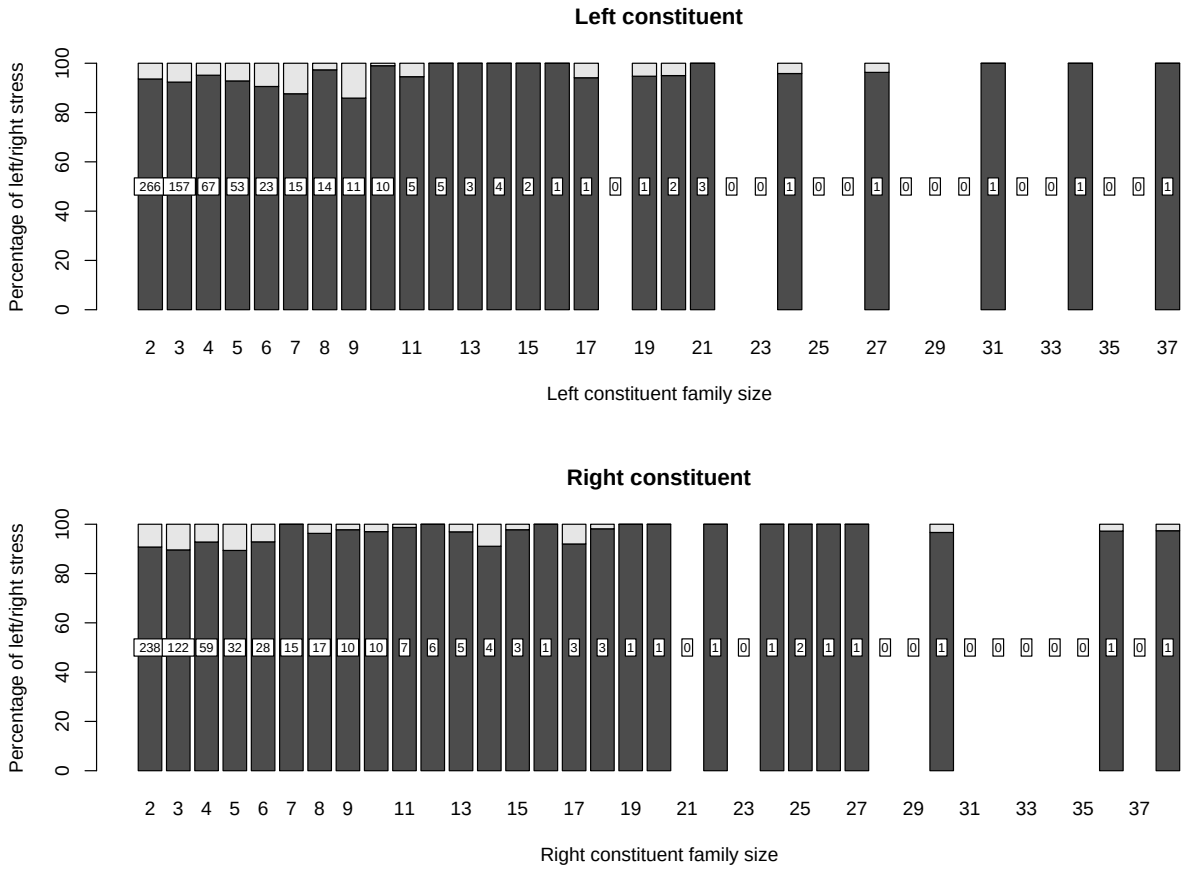

Figure 2: Stress patterns by left and right constituent family size, CELEX data. The light portions of the bars indicate right stresses, the black portions left stresses. The figures inside the bars give the number of observations, i.e. the number of compounds with that familiy size.

With regard to the left constituent, shown in the upper panel, there seems to be the expected increase in the proportion of right stresses up to a family size of 9 members, but then this effect quickly disappears and turns into its opposite. Larger left families with 12 or more members almost exclusively have only left stresses, instead of larger proportions of right stresses. In contrast, the lower panel shows that, in accordance with the prediction, larger right families have a smaller proportion of right stresses, but we also find some variation in the wrong direction.

In the regression model documented in table 4 we see, however, that both family size effects work in the same direction (i.e. less rightward stresses), indicated by the two negative coefficients. The performance of the model is better than with the T & W data, but not impressive ($C = 0.684$).

Table 4: Logistic regression model with left and right constituent family as predictors, CELEX data, N = 2562.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.8911	0.2544	-3.50	0.0005
left family size	-0.4311	0.1162	-3.71	0.0002
right family size	-0.7361	0.1159	-6.35	0.0000

Overall, the CELEX data provides mixed evidence. While for the right constituent, the family size approach makes the correct predictions (in line with what Marchand claimed to be the case), the left family size shows an effect that is exactly opposite to what was predicted.

4.3 Boston Corpus: Constituent family size alone

Figure 3 gives the distribution of stresses according to family size for this corpus. There is only one family that is not included, namely that of the left constituent *state* (as in, for example, *state official*, which has 31 members. This family was removed from the data set as an outlier, since all other families do not have more than 17 members. The distribution of stresses as shown in figure 3 indicates for the left constituent, shown in the left panel, that there seems to be an overall trend in the expected direction. However, especially the four rightmost bars show considerable variation. Notably they contain only a single family each, so that generalizations in these regions are almost impossible to draw anyway. A similar picture holds for the right constituent (right panel), with the predicted effect clearly discernible, but again only up to the family sizes where only one family is contained in each bin.

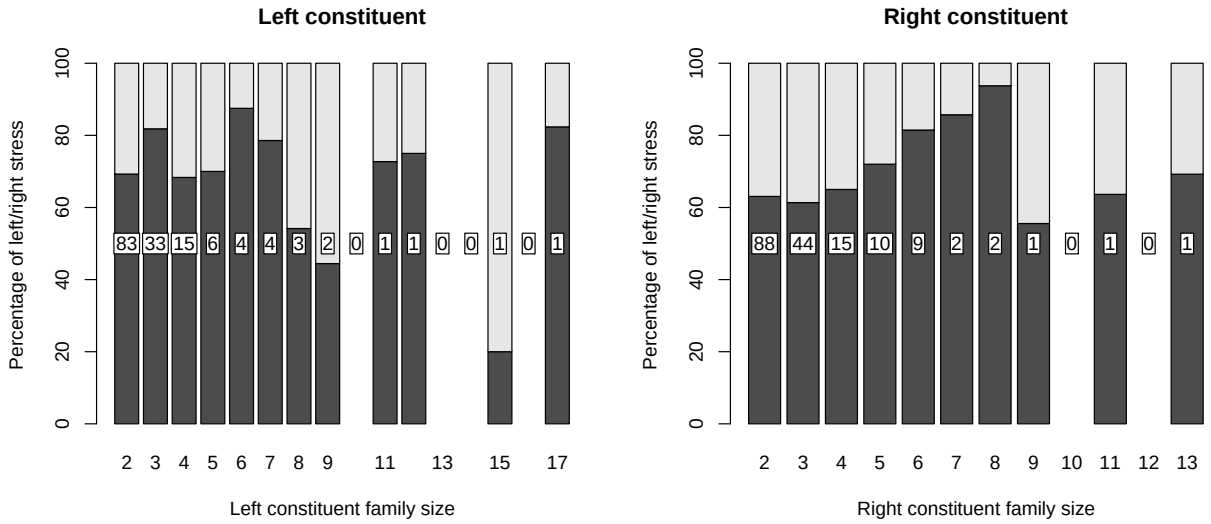

Figure 3: Stress patterns by left and right constituent family size, Boston Corpus. The light portions of the bars indicate right stresses, the black portions left stresses. The figures inside the bars give the number of observations, i.e. the number of compounds with that familiy size.

The regression analysis does not yield both of the expected effects. As shown in table 5, there is a main effect in the expected direction for the right constituent, but only a marginally significant effect in the predicted direction for the left constituent, with no interaction. The coefficients show that the effects work in opposite directions, as predicted. Again, the negative coefficient indicates an effect towards leftward stress, the positive coefficient towards rightward stress.

Table 5: Logistic regression model with left and right constituent family size as predictors, Boston Corpus, N = 504.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.6717	0.3266	-2.06	0.0397
left family size	0.2691	0.1490	1.81	0.0709
right family size	-0.4751	0.1992	-2.38	0.0171

The final model, from which LEFT FAMILY SIZE has been removed, is documented in table 6. The explanatory power of the model is rather weak ($C = 0.57$), which means that family size alone is not a very successful predictor of compound stress.

Table 6: Final logistic regression model with only family size as predictors, Boston Corpus, N = 504.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.3100	0.2561	-1.21	0.2260
right family size	-0.4660	0.1984	-2.35	0.0189

The data from the Boston corpus provide only a very small piece of evidence for an effect of family size, and hence of the informativeness of a given constituent for the assignment of stress to compounds containing this constituent. The predicted effect is very weak and reaches significance only for the right constituent.

To summarize, we have found somewhat mixed evidence concerning the hypothesis that family size plays a significant role in compound stress assignment. All three corpora show the predicted effect of right family size: the probability of right stress decreases with increasing right family size. The left family, however, never behaves as predicted by Bell (2008). It is either an insignificant predictor of stress (T & W, Boston Corpus) or, in CELEX, has an effect in the opposite direction: contrary to the prediction, the propability of left stress increases with increasing left family size. In general, the effect sizes are very small, which means that family size alone is not a good predictor of compound stress. In the following section we will investigate how the family sizes behave in models that also take other factors into account.

5 Results 2: Taking other factors into account

In this section we will include the effects of predictors other than constituent family size into our analysis to see whether constituent family survives as a significant predictor in the presence of the other independent variables. For the CELEX and Boston Corpus compounds Plag et al. (2007, 2008) coded each compound according to the structural and semantic categories held to be responsible for stress assignment in the literature (and some more), and we will use these codings in the following analyses. In addition, we used the stress bias of the constituent families, as coded in Plag (2009), as a means for factoring in analogical effects (see below for discussion). For the T & W data, no additional codings were available apart from the constituent family stress bias. In the next subsection we will describe in more detail which properties were coded, and how.

5.1 The coding: Other factors influencing compound stress assignment

With regard to argument structure, each compound is coded as to whether it is an argument-head structure or a modifier-head structure. In addition, the morphology of the head is also coded.[5] Furthermore, the factor SPELLING is coded as a proxy of lexicalization

[5]Both Plag et al. (2007) and Plag et al. (2008) found a significant effect of the affix of the head noun. In both studies, only those ending in the agentive suffix *–er* showed an effect of the argument-head vs. modifier-head distinction.

(with the values `1` for one-word, `h` for hyphenated, and `2` for two-word spellings).[6] To factor in semantic properties, each compound is coded with regard to following categories shown in (4), all of which are mentioned in the literature to trigger rigthward stress (e.g. Fudge 1984:144ff, Gussenhoven & Broeders 1981, Liberman and Sproat 1992, Zwicky 1986):

(4) N1 refers to a period or point in time (e.g. *night bird*)
N2 is a geographical term (e.g. *lee shore*)
N2 is a type of thoroughfare (e.g. *chain bridge*)
The compound is a proper noun (e.g. *Union Jack*)
N1 is a proper noun (e.g. *Achilles tendon*)

In addition, Plag et al. (2007, 2008) used a set of 18 semantic relations that are more or less established as useful in studies of compound interpretation. We also included these codings in the present study. The bulk of these relations come from Levi (1978), a seminal work on compound semantics, whose relations have since been employed in many linguistic (e.g. Liberman & Sproat 1992) and, more recently, psycholinguistic studies of compound structure, stress and meaning (cf., for example, Gagné & Shoben 1997, Gagné 2001). Levi's catalogue contains fewer than 18 relations, but some additions were made to ensure the possibility of reciprocal relations. Furthermore, a few categories were added, such as N2 IS NAMED AFTER N1. The relations are expressed by supposedly language-independent predicates that link the concepts denoted by the two constituents (see Levi 1978 for discussion). Table 7 gives the 18 semantic relations coded. A subset of these, as given in table 8, have been claimed to trigger rightward stress (e.g. Fudge 1984:144ff, Zwicky 1986, Liberman and Sproat 1992). All semantic predictors have been coded as binary factors with the values `yes` and `no`, to allow for multiple interpretations of a given compound.

[6] Although the spelling of compounds varies among speakers, it is uncontroversial that a more intricate spelling, i.e. as one word or hyphenated, is an indication of a more word-like, i.e. lexicalized, status of that combination. Both Plag et al. (2007) and Plag et al. (2008) found a significant effect of spelling, in that compounds with one-word spelling have a very strong tendency towards leftward stress, while compounds spelled as two words are much more variable in their stress pattern.

Table 7: List of semantic relations coded, illustrated with one example from CELEX each.

	Semantic relation	example
1.	N2 CAUSES N1	*teargas*
2.	N1 CAUSES N2	*heat rash*
3.	N2 HAS N1	*stock market*
4.	N1 HAS N2	*lung power*
5.	N2 MAKES N1	*silkworm*
6.	N1 MAKES N2	*steam-heat*
7.	N2 IS MADE OF N1	*milk pudding*
8.	N2 USES N1	*water mill*
9.	N1 USES N2	*handbrake*
10.	N1 IS N2	*child prodigy*
11.	N1 IS LIKE N2	*kettle drum*
12.	N2 FOR N1	*travel agency*
13.	N2 ABOUT N1	*mortality table*
14.	N2 IS LOCATED AT/IN/... N1	*garden party*
15.	N1 IS LOCATED AT/IN/... N2	*taxi stand*
16.	N2 DURING N1	*night watch*
17.	N2 IS NAMED AFTER N1	*Wellington boot*
18.	OTHER	*schoolfellow*

Table 8: List of semantic relations held to trigger rightward stress.

	Semantic relation	example
6.	N1 MAKES N2	*firelight*
7.	N2 IS MADE OF N1	*potato crisp*
14.	N2 IS LOCATED AT/IN/... N1	*garden party*
16.	N2 DURING N1	*night watch*

With regard to analogical effects, Plag (2009) showed that the constituent family stress bias plays a significant role in stress assignment, with a generally greater effect size than semantic predictors. What is this stress bias? The constituent family stress bias is a measure of the tendency within a left or right constituent family to favor a particular kind of stress. If, for example, all words with a particular right constituent (e.g. all words that have *street* as their right constituent) have leftward stress, we would expect new compounds with that constituent to also show leftward stress. Conversely, if there is a bias towards right stress in the family, as would be the case for the right constituent *avenue*, we would expect new compounds in that family to have rightward stress. This is the kind of analogical effect that has been hypothesized to exist with compounds involving *street* or *avenue* as right constituents. As Plag (2009) shows, the stress bias effect is significant for both left and right constituent families, with left family bias having an even stronger effect.

Plag (2009) computed the stress bias of a constituent family by calculating the proportion of left stresses within each constituent family of all compounds in each corpus

and transformed the resulting proportion into a categorical bias (i.e. with the values `left bias`, `right bias`, and `neutral`). We will use the same procedures to calculate the stress biases, but we will not transform the resulting proportions into categorical biases, in order not to lose statistical power. The way we compute the stress bias means that this variable can be defined as the probability that any member of a given family has left stress. The resulting proportions were standardized to reduce the danger of collinearity.

To illustrate our procedure with an example from the Boston Corpus, consider the compound *advertising business*, which has a left family with six other members (*advertising agency, advertising battle, advertising commentator, advertising costs, advertising days, advertising dollars*), and a right family with two other members (*biotechnology business, computer business*). Of the six other compounds with the left constituent *advertising*, five are left-stressed, one is rightward-stress, which amounts to a probability of 5/6, i.e. 0.83, for compounds of this family to be left-stressed. Of the right constituent family of *advertising business*, one compound (*biotechnology business*) is attested with leftward stress, the other compound (*computer business*) with rightward stress. This amounts to a right constituent family bias for rightward stress of 0.5, i.e. rightward stress and leftward stress are, on average, equally likely for compounds with this right constituent. Note that by using this procedure, the stress of the compound in question is not taken into account when computing the family bias for this compound. This is done in order to avoid the problem of predicting the stress of an item on the basis of stress information gleaned also from that very item. Appendix 2 illustrates a some constituent families with their respective stress biases.

In the following subsections we discuss for each corpus how well the overall 28 different predictors can predict compound stress assignment.

5.2 Teschner & Whitley (2004): Constituent family size and constituent family stress bias

As mentioned above, for this data set only constituent family size and constituent family stress bias were available as predictors. A logistic regression model with the four predictors was fitted to the data, including interactions of family size and family bias. There was no significant interaction term for right family size and right family bias, so that this interaction was removed during model simplification. The final model is documented in table 9.

Table 9: Final logistic regression model, based on family size and family bias, T & W data, N = 782.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.1941	0.2219	-14.39	0.0000
left family size	-0.1676	0.2120	-0.79	0.4291
left family bias	-1.3127	0.1230	-10.67	0.0000
right family size	-0.3977	0.1914	-2.08	0.0377
right family bias	-0.7113	0.1163	-6.12	0.0000
left family size : left family bias	-0.4020	0.1179	-3.41	0.0007

,

There are three main effects, one of right family size, one of right family bias, and one of left family bias. In addition there is a significant interaction of left family size and left family bias. How can we interpret the coefficients? Our model wants to predict rightward stress, so positive coefficients indicate that a predictor works in the direction of rightward stress, while negative predictors work in the direction of lefteward stress. For example, the negative intercept means that on average, our compounds would receive leftward stress. Both family biases have highly significant negative coefficients, which means that with increasing family bias, we get more left stresses. Recall that 'increasing family bias' means an increasing proportion of compounds with left stress in the family. So these family bias effects are exactly what the analogical hypothesis predicts. Let us look at the effects of the family sizes. LEFT FAMILY SIZE by itself is insignificant, but is kept in the model since it enters into a significant interaction with LEFT FAMILY BIAS.

In order to understand better the nature of the interaction and the coefficients of the model, we have plotted all effects in figure 4. The y-axis in the six panels shows the probability of right stress as predicted by our regression model, the x-axis shows the effect of the respective predictor, holding all other predictors constant at their medians in the case of continuous variables, and at the most frequent factor level in the case of categorical variables.[7] To understand the scale of the x-axis, recall that we are dealing with transformed (i.e. standardized) biases here. This means that for the left family, a strictly rightward bias of 0.0 corresponds to a transformed value of -3.20, a neutral bias of 0.5 corresponds to -1.41, and a strictly leftward bias of 1.0 corresponds to 0.37. For the right family, the corresponding transformed values are -3.44, -1.52, and 0.40, respectively. These transformed values for strictly rightward bias and strictly leftward bias delimit the display ranges of the regression lines for the partial effects of left and right family bias, as shown in the middle upper panel and lower left panel, respectively.

[7]The plots show individual dots instead of regression lines if the predictor is either a categorical variable, or if the number of different values for a continuous predictor is not large enough to warrant representation by a line. Broken lines surrounding a regression line, as in the middle upper panel of figure 4, give the 95 percent confidence interval.

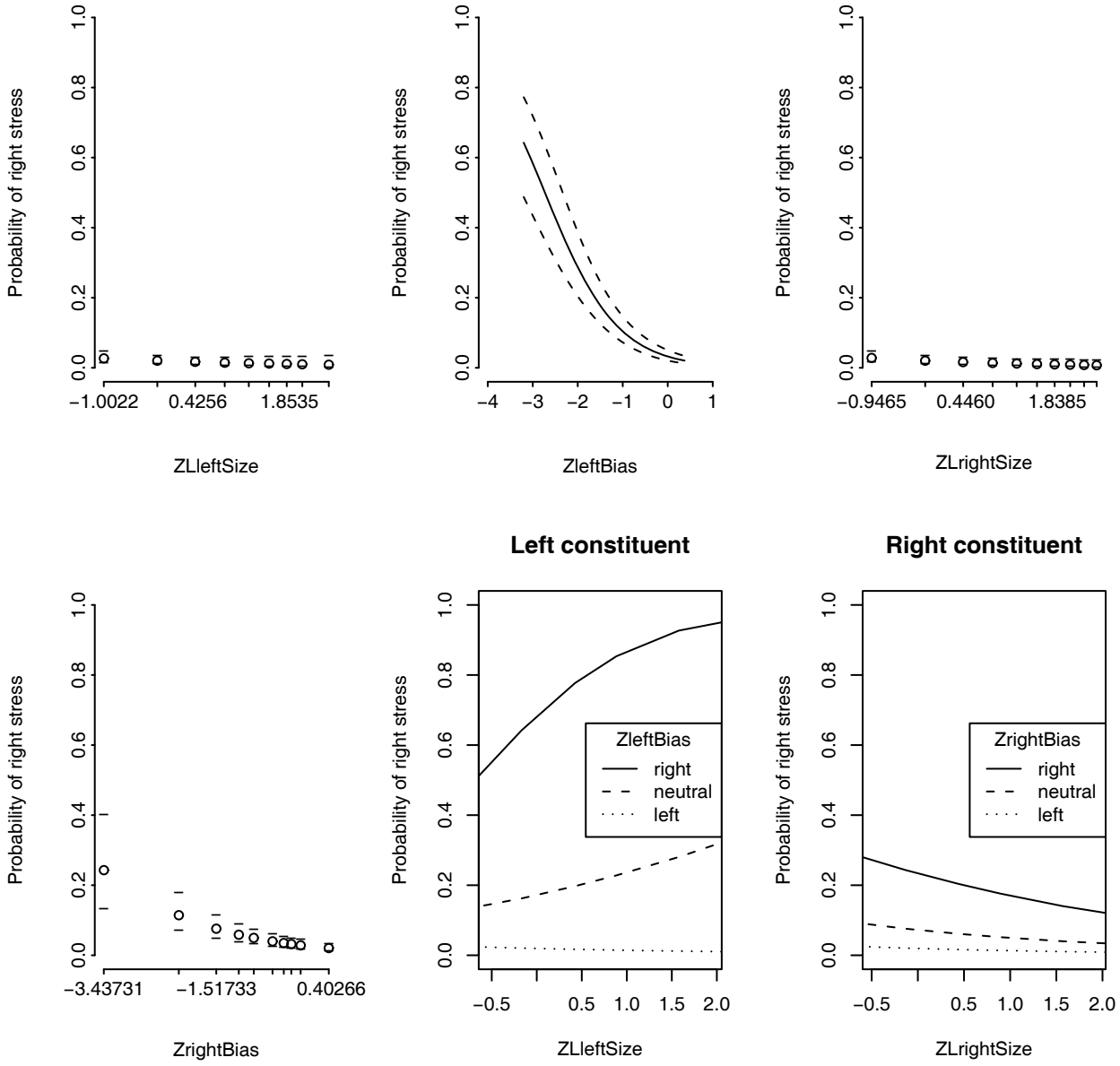

Figure 4: Partial effects of left family size ('ZleftSize'), left family bias ('ZleftBias'), right family size ('ZrightSize'), right family bias ('ZrightBias'), and the interactions of left family bias ('ZleftBias') with left family size('ZleftSize'), and of right family bias ('ZrightBias') with right family size('ZrightSize'), T & W data, N = 782.

The upper left panel shows that increasing the left family size has practically no effect on the proportion of left and right stresses, contra to the hypothesis. In contrast, the upper middle panel shows that an increasing left bias goes together with an decreasing proportion of right stresses. Increasing the right family size (upper right panel) has a significant effect, as shown in table 9, but this effect is very small, and indeed hardly

visible in the plot. The effect of the right family bias is clearly shown in the lower left panel of the plot. Particularly interesting are the final two lower plots, which illustrate the interaction between family sizes and family biases. The different lines represent different types of biases. The solid line is the regression line for compounds with a strict bias towards right stress, the broken line represents compounds with a neutral bias, and the dotted line those with a strict bias for left stress. In other words, the three regression lines in each of the two panels display the effect of family size for three kinds of compounds: those with a strictly rightward family bias (solid line), those with a neutral family bias (dashed line), and those with a strictly leftward family bias (dotted line). The effect of family size for compounds with other family biases falls between these three compound types, and plotting regression lines for compounds with biases ranging between these three special biases would yield additional lines ranging accordingly between the three lines given in the plot.

Looking at the left constituent effects (shown in the middle lower panel), we see that an increase in family size has different effects on the probability of right stress, depending on the kinds of stress biases. Compounds with a full bias towards rightward stress (solid line) have an increasing probability of showing rightward stress with increasing family size. For compounds with a left bias, the probabilty of right stress is largely unaffected by family size (dotted line). Compounds with a neutral bias are in between.

For right constituents, the interaction between size and bias is insignificant, which means that the effect of family size on the probability seems to be the same, regardless of the family bias. In the plot, there is a decrease of probability for right stress with increasing right family sizes, but this decrease does not differ significantly according to the types of biases (the slopes of the three lines are not different enough from each other). The absence of an interaction in the presence of a main effect of the size of the right constituent family may be seen as an argument in favor of the existence of an independent family size effect, but not a very strong one.

5.3 CELEX: Constituent family size and other predictors

A logistic regression model with all predictors was fitted to the data. Table 10 summarizes the final model. The estimates of the categorial predictors indicate the change in the response variable from baseline level to the level to the right of the '=' sign. For example, the baseline level for the semantic relations is `no`, which means that the respective coefficient represents the difference (given in logits) that emerges if we change the factor level from `no` (i.e. 'not showing this relation') to `yes` (i.e. 'showing this relation'). For orthography, the baseline level is `1` (for 'one word').[8]

[8] `h` means 'hyphenated', `2` means 'two words'.

Table 10: Final logistic regression model based on all predictors, CELEX data, N = 2562.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.4616	0.2700	-16.53	0.0000
orth=2	2.2884	0.2889	7.92	0.0000
orth=h	1.1001	0.3054	3.60	0.0003
semRel4=yes	0.7585	0.2845	2.67	0.0077
semRel7=yes	1.1345	0.2780	4.08	0.0000
semRel12=yes	-1.3390	0.2780	-4.82	0.0000
semRel16=yes	1.2607	0.4146	3.04	0.0024
left family size	-0.0272	0.1211	-0.22	0.8222
left family bias	-0.7439	0.1011	-7.36	0.0000
right family size	-0.4754	0.1445	-3.29	0.0010
right family bias	-0.8696	0.1421	-6.12	0.0000
left family size : left family bias	-0.2794	0.1014	-2.75	0.0059
right family size : right family bias	-0.5441	0.1356	-4.01	0.0001

The regression model shows very interesting results. If other predictors are taken into account, the main effect of the left family size in the simpler model (see again table 4), which went in the wrong direction, disappears. It only stays in the model due to its interaction with family bias (to be discussed shortly). The predicted effect of the right family size survives in the model. The overall discrimative power of the model is very good ($C = 0.906$). In addition to the family size effect we find a lexicalization effect (via SPELLING), an analogical effect via the two family stress biases, and an effect for four semantic relations. SEMREL4, SEMREL7 and SEMREL16 (i.e. N1 HAS N2, N2 IS MADE OF N1, N2 FOR N1, and N2 DURING N1, respectively) work in the direction of rightward stress, SEMREL12 (N2 FOR N1) in the direction of leftward stress.

Figure 5 displays the partial main effect of each predictor in the model, holding again all other predictors constant (either at their medians, in the case of continuous variables, or at the most frequent factor level, in the case of categorical variables). As before, display ranges for the transformed family biases are delimited by the values corresponding to strictly rightward and strictly leftward biases, with neutral bias at the mid-point of the range. A closer inspection of the partial main effects reveals that the effect of family bias is much greater than the effect of all other factors. Apparently,these other factors are rather uninformative in comparison to the effect of family bias information.

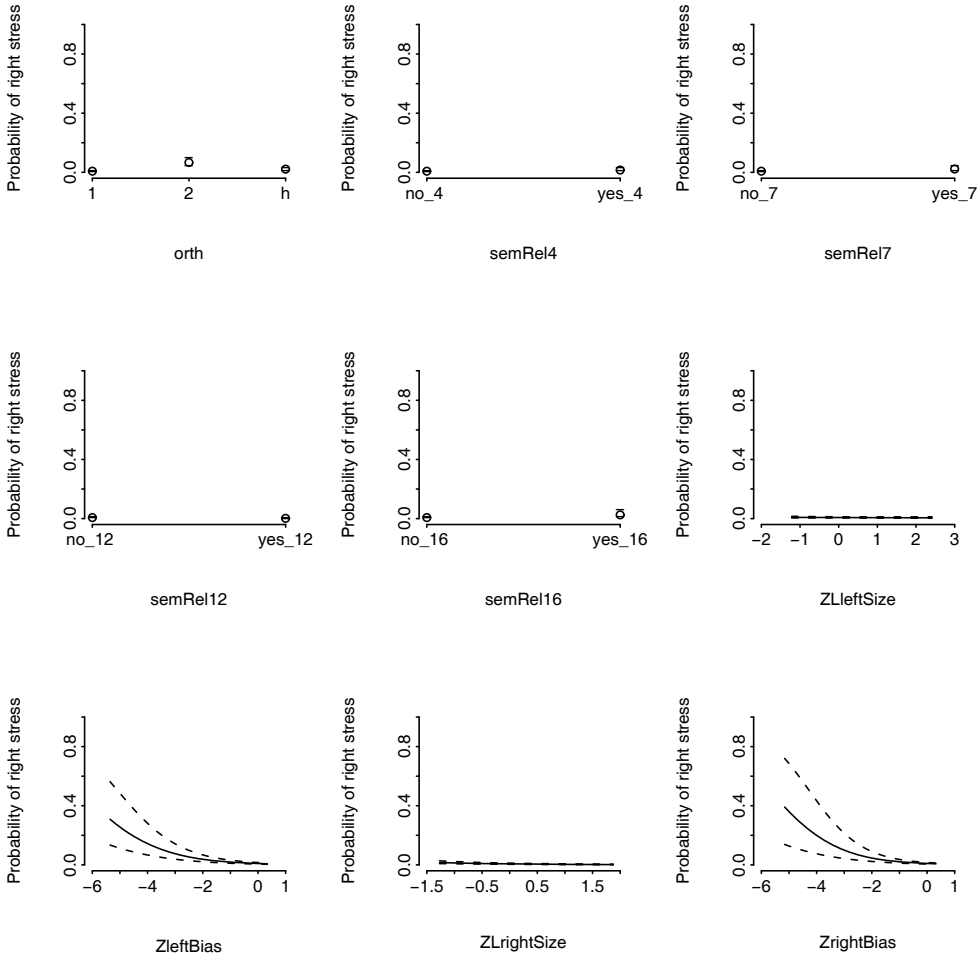

Figure 5: Partial main effects of final regression model, CELEX data, N = 2562.

Let us turn to the interactions. The interaction plots in figure 6 show us how to interpret the interactions.[9]

[9]These interactions bring in collinearities exceeding the VIF value of 2.5 for the following variables: 2.81 for `ZleftBias`, 6.09 for `ZrightBias`, 2.77 for the interaction of left bias and left size, and 6.16 for the interaction of right size and right bias. In order to see whether the same main effects emerged in a model without the interactions, we devised an alternative model with no interactions. In this model all VIFs were below 2.5., and the main effects were identical to those of the model with interactions. This is an indication that we do not run into collinearity problems with the full model including the interactions, in spite of some of the VIFs of the full model being slightly higher than the very conservative threshold level of 2.5.

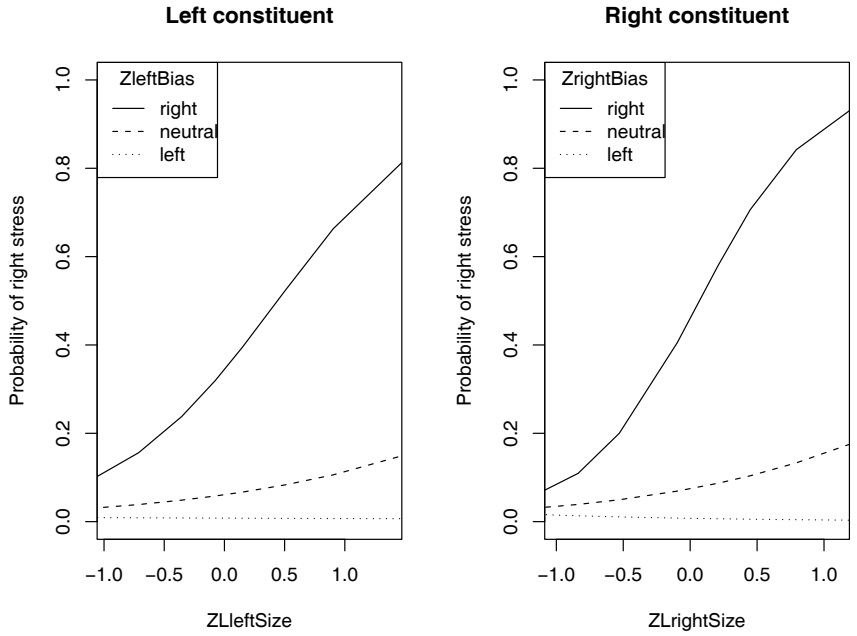

Figure 6: Interactions of left family bias ('ZleftBias') with left family size('ZleftSize'), and right family bias ('ZrightBias') with right family size ('ZrightSize'), CELEX data, N = 2562.

The interaction between family size and bias works in the same way for both left and right constituents. An increase of family size has an effect on the probability of right stress only for families with a strong bias towards right stress (solid lines). For these families, the effect of family bias becomes stronger with increasing family size. For families with a bias towards left stress (dotted lines), family size has no visible effect: The very strong influence of the left family bias is completely independent of family size.

5.4 Boston Corpus: Constituent family size and other predictors

In the logistic regression with all variables (including pertinent interactions, see below), only one semantic predictor survives (SEMREL13: N2 ABOUT N1). In addition, the constituent family stress biases and the right family size are also still significant, with effects in the predicted directions. We again find a significant interaction between the left size and the left bias and an interaction between the right size and the rigth bias. Table 11 is a summary of the model. The fit of the model is quite satisfactory ($C = 0.79$), which means that the model could be used as an automatic classifier with considerable success.

The plots in figure 7 show the partial main effects of the surviving predictors. One can see that the effects of the two biases are the strongest (upper and lower rightmost panels). Increasing the value of the bias means more left stresses for both left and right constituent families. The effects of the two family sizes are much less pronounced (upper middle panel and lower left panel). The semantic relation N2 ABOUT N1 goes together with an increased probability of right stresses.

Table 11: Final logistic regression model based on all predictors, Boston Corpus data, N = 504.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.0870	0.1317	-8.25	0.0000
semRel13=yes_13	0.8170	0.3311	2.47	0.0136
left family size	-0.0297	0.1449	-0.20	0.8376
left family bias	-0.9902	0.1426	-6.94	0.0000
right family size	-0.2372	0.1180	-2.01	0.0443
right family bias	-0.7857	0.1386	-5.67	0.0000
left family size : left family bias	-0.8413	0.1654	-5.09	0.0000
right family size : right family bias	-0.3856	0.1549	-2.49	0.0128

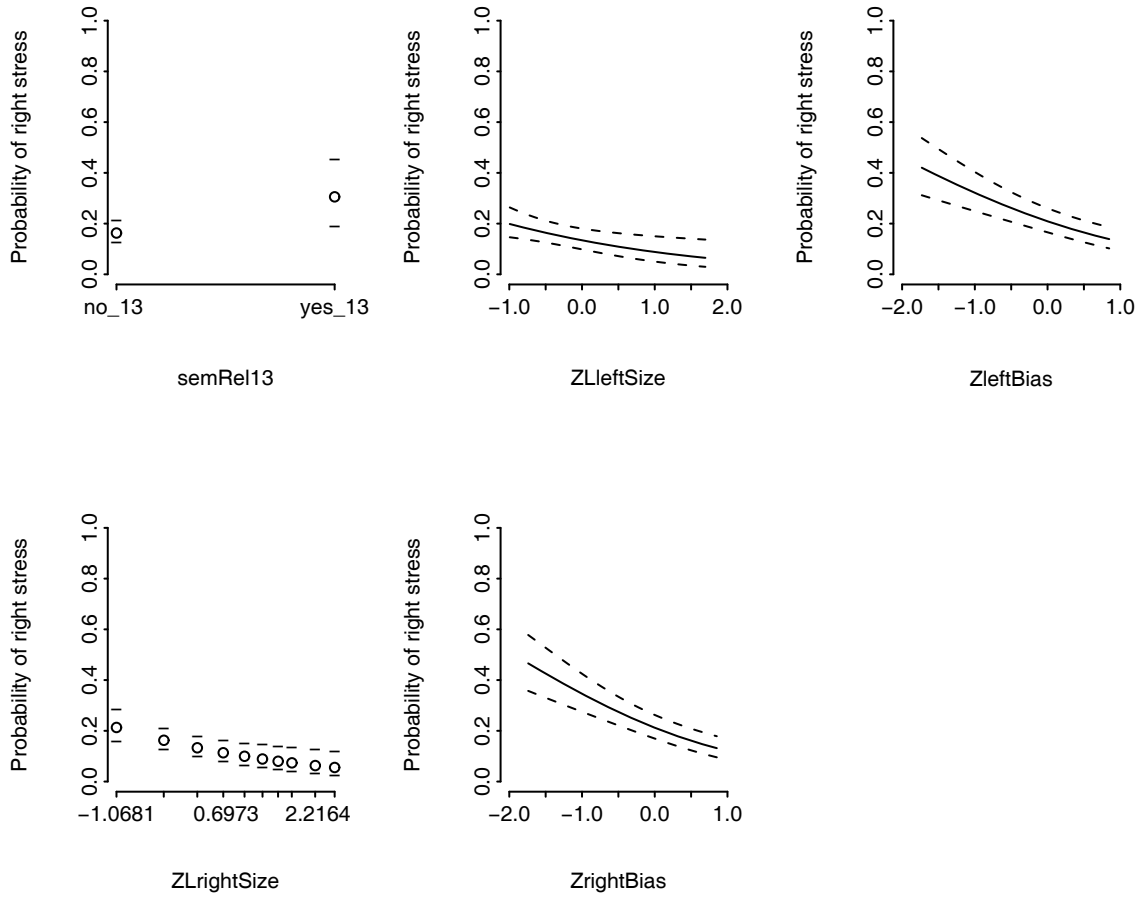

Figure 7: Partial main effects of final regression model, Boston Corpus data, N = 504.

Figure 8 illustrates the interactions. As with CELEX, the effect of the interaction between family size and family bias is very similar for both constituents: Families with a

bias towards left stress (represented by the dotted lines) show a low probability of right stress. This probability decreases slightly with increasing family sizes. The probability of right stress increases with increasing family size for those families that have a bias towards right stress (solid lines). Overall this means that family size in fact does not influence stress assignment directly, but rather modulates the strength of the family bias effect. For both constituents, increasing family sizes increase the influence of the family biases.

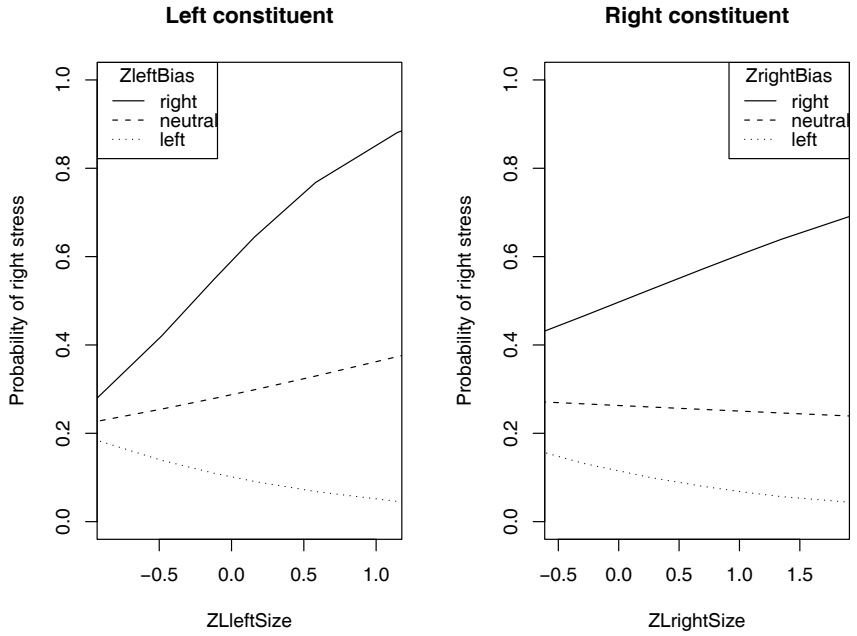

Figure 8: Interactions of left family bias ('ZleftBias') with left family size('ZleftSize'), and right family bias ('ZrightBias') with right family size('ZrightSize'), Boston Corpus data, N = 504.

6 Summary and discussion

This paper investigated the effect of constituent family size (as a proxy for the informativeness of a given compound constituent) on compound stress assignment. In the analysis of the two family sizes as the only predictors it turned out that there is somewhat mixed evidence. The T & W data showed the predicted tendencies, but only the size effect of the right family was significant. The same pattern was found for the Boston Corpus data, for which only the family size effect of the right constituent reached significance. The CELEX data showed one of the predicted main effects and one main effect in the non-predicted direction.

We then investigated whether the observed family size effects persisted if other variables suspected of influencing compound stress assignment were factored in. In the T & W data, for which only family stress biases were available as additional predictors, it turned out that the stress biases, in particular that of the left family, were much stronger in their

effect. In addition, we found an interaction of size and bias, in that for left constituents with a bias towards rightward stess, an increase in left family size leads to an even more pronounced tendency towards rightward stress. In other words, increasing family size strengthens the pertinent stress bias. This means that for the left family, family size works in the predicted direction, but only primarily as a modifier of the much stronger family bias. There was no interaction between size and bias for the right family, but a decrease of the probability of rightward stress with increasing right family sizes across the board. This can be taken as evidence for the existence of an independent family size effect, but not a very strong one.

In the CELEX data and the Boston Corpus data the family sizes of both constituents interact with the much stronger stress biases. We have seen that increasing family size increases the chance of rightward stress for families that have a family bias towards rightward stress. For families that have a family bias towards leftward stress, increasing family size either has no effect whatsoever on the probability of rightward stress (particularly true for CELEX), or decreases the probability of rightward stress (particularly true for the Boston Corpus). The only instance where family size does not act as a mediator of family bias, but works as an independent predictor without interaction, is in the case of right constituents in the T & W data. Here, the probability of rightward stress decreases with increasing family size in general, irrespective of the family bias (at least, family bias is not significantly interacting with size).

What do these results mean for an information-based approach to compound stress, and for an account of compound stress in general? Overall, our analyses have found little evidence for a general effect of family size, if other factors are taken into account. One could, however, interpret our results as an indication that family bias becomes more important in the regression analysis when the family size is larger, both in left and right constituent families. This may have an interesting, yet unknown psycholinguistic reason having to do with the organization of the mental lexicon. It may, however, and quite disappointingly, be simply a reflection of a methodological issue. For small family sizes (which are prevalent in our data sets), the information encoded in the family bias is based on a very small number of observations. The information is bound to be much more unreliable than if the bias is calculated on the basis of a large number of observations. If that is the real explanation behind our results, it would independently strength en the idea that constituent family bias is a, perhaps *the* major force in stress assignment. Due to the limitations of our data sets, constituent families were necessarily small, and presumably much smaller than the constituent families in the minds of real speakers. If we now, based on the results of this paper, arrive at the conclusion that larger families allow better predictions, we can assume that real speakers necessarily can do a better stress assignment job than our models, which are based on rather small families. It is all the more striking, and supports the important role of family bias in stress assignment (as against other factors mentioned in the literature), that our models nevertheless reach acceptable classification results. Coming back to the initial question of whether informativeness plays a role in compound stress assignment, we have to say that we could not provide compelling evidence in favor of this idea. It seems, however, that larger data sets are needed to further substantiate this conclusion (or to prove it wrong).

Appendix 1[10]

Random sample of 100 compounds from T & W

fish finger, time capsule, ground rule, dust jacket, tea cloth, student nurse, saloon bar, Latin American, gas mask, love match, tree surgeon, student days, toll road, punch bag, ice hockey, rubbish dump, ring finger, ring road, thrift account, test match, bikini line, jelly roll, home owner, telephone number, swing shift, service station, soda cracker, watering can, county court, station wagon, air bag, doggy bag, pet food, street door, punch card, rear end, coffee house, window seat, fish farming, city hall, money belt, fish market, water sports, cottage industry, cocktail cabinet, fish stick, satellite town, power base, death trap, tape deck, train spotter, grade crossing, car pool, storage space, road test, road tax, shopping bag, driving school, video recording, sailing boat, service lift, beer garden, sea lion, wine gum, firing line, blood bath, flight bag, shower curtain, steel industry, county seat, sausage meat, truck stop, credit card, funeral service, family allowance, power plant, state education, town house, press agent, day release, cream soda, horse sense, publishing company, emery board, part timer, rock garden, dressing table, sleeping sickness, fuel oil, shoulder strap, set phrase, emergency room, field sports, staff meeting, litmus paper, field hockey, polling day, field day, wedding ring, polling place

Random sample of 100 compounds from CELEX

iron works, pot shot, employment exchange shirt sleeve, bed spread, coffee house, passion flower, eye ball, goal post, pan cake, lumber room, copy right, horse fly, pepper pot, motor boat, dessert spoon, gate keeper, safety belt, mountain lion, cab stand, hip bath, bridge work, clock dial, gas bracket, wall paper, tie break, sheep fold, ground staff, ear shot, tie pin, body stocking, country man, cottage loaf, junk shop, hair cut, crew cut, milk bar, ice cube, well water, pole axe, stock holder, night work, box number, traffic circle, concert master, brain pan, egg roll, car pool, telegraph pole, alms house, chocolate bar, sea weed, place name, neon lamp, church yard, sound barrier, sea god, class list, stone breaker, fire storm, school mistress, bed side, air cushion, tin foil, sailing boat, oyster bed, land slip, cod piece, blood bath, dust sheet, quarter staff, birth day, way side, car port, lime light, pudding head, sky hook, wind gauge, copper head, wedding band, pruning hook, eye sore, sugar candy, fire watcher, whipping boy, watch word, tape deck, stone cutter, police officer, sunday clothes, star light, choir master, corn flower, dress hanger, nose bag, jack tar, water finder, needle craft, sports jacket, hay cock

Random sample of 100 compounds from the Boston Corpus

Boston area, state senator, crew season, bar associations, pesticide chief concrete beams, rat traps, front runner, stomach pain, bulger breakfast, visitation rights, house members oil facilities immigration policy, nanny school, bookkeeper, house speaker, tax cut, hand guns, strategy session, budget process, soap opera, weekend, community activists, turbo tax, cabinet secretaries transplant surgeon,computer program, state aid repair

[10]For technical reasons, the spelling of the compounds listed in this appendix is not necessarily the same as in the original source.

costs, massachusetts cities, condo boom, baseball, government subsidy, budget cuts, tax package,boston mayor,world summit, shrewsbury institute's, n.h.l. play-offs, roadway, state employees, tax return, fenway park, turbo tax, treasury officials, temper tantrums, state treasurer, industry analysts, lemon survey, state representative, weekend, training facility, loan sharks, bathroom, dukakis administration, oil fires, newspapers, cigarette tax, solidarity shows, deputy superintendent, paper trail pension benefits, rescue effort, communications devices, arts funding, art copies, school children, toll plaza, taxpayers, job market, consumer office, lottery participants, tax revenue court system, campaign promise, households, auto fees, Boston harbor, health study, credit laws, seabrook, testing ground, aids care, kentucky derby, work week, student body, health clinics, interest rules, state prison, house negotiators, bar associations, business commentator, assault rifle, congressman, science reporter, kansas city, model tribe, ball game, police officers, gulf war, area residents

Appendix 2[11]

Table 12: Two constituent families and their stress biases from T & W.

lmember	rmember	stressPos	leftBias
family	allowance	left	0.20
family	name	left	0.20
family	planning	right	0.40
family	tree	right	0.40
family	unit	right	0.40
family	vault	right	0.40
box	office	left	0.60
head	office	right	0.80
home	office	right	0.80
press	office	left	0.60
ticket	office	left	0.60
tourist	office	left	0.60

[11]For technical reasons, the spelling of the compounds listed in this appendix is not necessarily the same as in the original source.

Table 13: Two constituent families and their stress biases from CELEX.

lmember	rmember	stressPos	leftBias
country	man	left	0.50
country	party	left	0.50
country	woman	left	0.50
country	seat	right	0.62
country	side	left	0.50
country	house	right	0.62
country	dance	right	0.62
country	music	right	0.62
country	club	left	0.50
cart	horse	left	0.78
clothes	horse	left	0.78
cock	horse	right	0.89
draught	horse	left	0.78
iron	horse	right	0.89
post	horse	left	0.78
sea	horse	left	0.78
side	horse	left	0.78
towel	horse	left	0.78
war	horse	left	0.78

Table 14: A sample of constituent families and their stress biases from the Boston Corpus.

lmember	rmember	stressPos	leftBias
business	man	left	0.67
business	official	left	0.67
business	owner	left	0.67
business	reporter	left	0.67
business	service	right	0.83
business	tax	right	0.83
business	men	left	0.67
computer	program	left	0.50
drug	program	left	0.50
emergency	program	right	0.62
government	program	right	0.62
housing	program	right	0.62
lead	program	left	0.50
metco	program	left	0.50
recycling	program	left	0.50
state	program	right	0.62

References

Allison, Paul D. & Stephen I. Allison. 1999. *Multiple Regression: A Primer*. Pine Forge Press.

Baayen, Harald. 2008. *Analyzing linguistic data. A practical introduction to statistics*. Cambridge: Cambridge University Press.

Baayen, Harald, Laurie B. Feldman & Robert Schreuder. 2006. Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language* 53:496–512.

Bauer, Laurie. 1983. *English word-formation*. Cambridge: Cambridge University Press.

Bauer, Laurie. 1998. When is a sequence of two nouns a compound in English? *English Language and Linguistics* 2(1):65–86.

Bell, Melanie. 2008. Noun noun constructions and the assignment of stress. Paper presented at the 1rst Conference of the International Society for the Linguistics of English (ISLE 1), Freiburg, 8-11 October, 2008.

Carvajal, Carol Styles & Jane Horwood. 1996. *The Oxford Spanish-English dictionary: New international edition*. Oxford: Oxford University Press.

Chomsky, Noam & Morris Halle. 1968. *The sound pattern of English*. New York: Harper and Row.

Fudge, Erik. 1984. *English word-stress*. London: George Allen & Unwin.

Giegerich, Heinz J. 2004. Compound or phrase? English noun-plus-noun constructions and the stress criterion. *English Language and Linguistics* 8:1–24.

Gussenhoven, Carlos & A. Broeders. 1981. *English pronunciation for student teachers*. Groningen: Wolters-Noordhoff-Longman.

Jaccard, James, Choi K. Wan & Robert Turrisi. 1990. The detection and interpretation of interaction effects between continuous variables in multiple regression. *Multivariate Behavioral Research* 25(4):476–478.

Jespersen, Otto. 1909. *A Modern English Grammar. On Historical Principles. Part I: Sounds and spelling*. London: Allen and Unwin. Reprinted 1961.

Kingdon, Roger. 1958. *The groundwork of English stress*. London: Longmans, Green and Co.

Kunter, Gero. 2009. *The phonetics and phonology of compound stress in English*. Ph.D. thesis, Universität Siegen.

Kunter, Gero & Ingo Plag. 2007. What is compound stress? In *Proceedings of the 16th International Congress of Phonetic Sciences, Saarbrücken*, 1005–1008.

Kuperman, Victor, Raymond Bertram & Harald Baayen. 2009. Processing trade-offs in the reading of Dutch derived words. *Ms* .

Kutner, Michael H., Christopher J. Nachtsheim, John Neter & William Li. 2005. *Applied linear statistical models*. Boston: McGraw-Hill.

Ladd, D. Robert. 1984. English compound stress. In Dafydd Gibbon & Helmut Richter (eds.) *Intonation, Accent and Rhythm*, 253–266. Berlin: de Gruyter.

Lappe, Sabine & Ingo Plag. 2007. The variability of compound stress in English: Towards an exemplar-based alternative of the compound stress rule. In *Proceedings of the ESSLLI workshop on exemplar-based models of language acquisition and use*. Dublin, Ireland.

Lappe, Sabine & Ingo Plag. 2008. The variability of compound stress in English: rules or exemplars? Paper presented at the 13th International Morphology Meeting, University of Vienna, 3–6 February 2008.

Liberman, Mark & Richard Sproat. 1992. The stress and structure of modified noun phrases in English. In Ivan A. Sag & Anna Szabolcsi (eds.) *Lexical matters*, 131–181. Stanford: Center for the Study of Language and Information.

Marchand, Hans. 1969. *The Categories and Types of Present-Day English Word Formation. A Synchronic-Diachronic Approach*. München: Beck'sche Verlagsbuchhandlung.

Milin, Petar, Victor Kuperman, Aleksandar Kostic & R. Harald Baayen. 2009. Paradigms bit by bit: an information-theoretic approach to the processing of paradigmatic structure in inflection and derivation. *submitted* .

O'Brien, Robert M. 2007. A caution regarding rules of thumb for variance inflation factors. *Quality and Quantity* 41(5):673–690.

Olsen, Susan. 2000. Compounding and stress in English: A closer look at the boundary between morphology and syntax. *Linguistische Berichte* 181:55–69.

Olsen, Susan. 2001. Copulative compounds: a closer look at the interface between syntax and morphology. In Geeert E. Booij & Jaap van van Marle (eds.) *Yearbook of Morphology 2000*. Dordrecht/Boston/London: Kluwer.

Plag, Ingo. 2003. *Word-formation in English*. Cambridge: Cambridge University Press.

Plag, Ingo. 2006. The variability of compound stress in English: structural, semantic, and analogical factors. *English Language and Linguistics* 10(1):143–172.

Plag, Ingo. 2009. Compound stress assignment by analogy: the consituent family bias. *submitted* 35 pp.

Plag, Ingo, Gero Kunter & Sabine Lappe. 2007. Testing hypotheses about compound stress assignment in English: a corpus-based investigation. *Corpus Linguistics and Linguistic Theory* 3(2):199–232.

Plag, Ingo, Gero Kunter, Sabine Lappe & Maria Braun. 2008. The role of semantics, argument structure, and lexicalization in compound stress assignment in English. *Language* 84.4.

Moscoso del Prado Martín, Fermín, Aleksandar Kostić & Harald Baayen. 2004. Putting the bits together: An information theoretical perspective on morphological processing. *Cognition* 94:1–18.

R Development Core Team. 2007. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

Sampson, Rodney. 1980. Stress in English N + N phrases: a further complicating factor. *English Studies* 61:264–270.

Schmerling, Susan F. 1971. A stress mess. *Studies in the Linguistic Sciences* 1:52–66.

Spencer, Andrew. 2003. Does English have productive compounding? In Geert E. Booij, Janet DeCesaris, Angela Ralli & Sergio Scalise (eds.) *Topics in Morphology. Selected papers from the 3rd mediterranean morphology meeting*, 329–341. Barcelona: Institut Universitari de Lingüística Applicada.

Sproat, Richard. 1994. English noun-phrase accent prediction for text-to-speech. *Computer Speech and Language* 8:79–94.

Stine, Robert A. 1995. Graphical interpretation of variance inflation factors. *The American Statistician* 49:53–56.

Teschner, Richard V. & Melvin Stanley Whitley. 2004. *Pronouncing English*. Washington, D.C.: Georgetown University Press.

Zwicky, Arnold M. 1986. Forestress and afterstress. In *Working Papers in Linguistics*, volume 32, 46–72. Columbus: Ohio State University.