Compound stress assignment by analogy: the consituent family bias

Ingo Plag

September 22, 2009

Revised version of September 22, 2009
to appear in *Zeitschrift für Sprachwissenschaft*

Ingo Plag
Universität Siegen
English Linguistics, Fachbereich 3
Adolf-Reichwein-Str. 2
D-57068 Siegen, Germany
plag@anglistik.uni-siegen.de
http://www.uni-siegen.de/ engspra

Compound stress assignment by analogy: the consituent family bias

Ingo Plag

September 22, 2009

Abstract

This paper tests the hypothesis that stress assignment to English compounds works on the basis of analogy. In particular, the role of the constituent family, i.e. the set of compounds that share the same right or left constituent with a given compound, is investigated. On the basis of large amounts of data from three different corpora it is shown that the tendency towards a certain kind of stress pattern within the constituent families of a given compound is a strong predictor for stress assignment. This challenges rule-based approaches to compound stress assignment and lends independent evidence to exemplar-based approaches to language structure.[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, 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 should be regarded as compounds by most 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,

[1]I would like to thank Sabine Arndt-Lappe, Kristina Kösling, Gero Kunter, Mareile Schramm, Linda Zirkel and the anonymous *ZS* reviewers for their feedback on an earlier version. Special thanks also to Heinz Giegerich and Harald Baayen for discussion and support. This work was made possible by two grants from the *Deutsche Forschungsgemeinschaft* (PL151/5-1, PL 151/5-3), which I gratefully acknowledge.

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 of noun-noun constructs. Until a few years ago systematic empirical work on the problem was lacking, but recent experimental and corpus studies have shown that determinstic approaches based on structural or semantic features are not very successful in predicting noun-noun stress correctly (Plag 2006, Plag et al. 2007, 2008, Lappe & Plag 2008).

The aim of the present paper is to test the adequacy of one particular alternative hypothesis that has been around for quite some time (see, for example, Schmerling 1971), but has not been thoroughly investigated so far. This hypothesis states that compound stress assignment is based on analogy to similar compounds in the lexicon. In particular, it has been claimed 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 what we call the 'constituent family bias'. A 'constituent family' is the set of compounds that share the first, or the second, constituent with a given compound. And the constituent family bias is the tendency of a given constituent family to favor a particular kind of stress, for example leftward stress.

In this paper we test this hypothesis using large amounts of data extracted from three different data sources: Teschner & Whitley (2004), the CELEX lexical database (Baayen et al. 1995) and the Boston University Radio Speech Corpus ('Boston Corpus' for short, Ostendorf et al. 1996). Our results show that in all three types of data the constituent family bias is indeed a strong predictor of noun-noun stress. In regression analyses including all potential factors, most of the other potential effects disappear as significant predictors of stress assignment, and the constituent family bias remains a robust, and often most important, factor.

Before we turn to a more detailed discussion of the existing hypotheses about compound stress assignment, a word is in order with regard to the notorious problem of whether noun-noun constructions should be analyzed as compounds or phrases. In general we remain agnostic with regard to this issue, because, first, the a priori exclusion of certain types of data might have biased our results in an undesired fashion. Thus, in the literature on the variability of compound stress, the notion of noun-noun compound is usually taken for granted, so that in a study that wants to test any claims in this domain a restrictive definition of noun-noun compound is inappropriate. Second, it has often been pointed out (e.g. more recently by Bauer (1998) or Spencer (2003)) that the stress criterion is inadequate to distinguish between the two types of construction (if one believes in this dichotomy in the first place). Other criteria, such as separability, spelling, or semantic transparency, do not yield consistent results either (cf. Bauer 1998). Hence we sometimes, and conservatively, speak of 'noun-noun constructs' in this paper, although the structures under investigation would probably be regarded as proper compounds by most analysts.

In what follows, we first review the hypotheses put forward in the literature and then describe our methodology (section 3). This is followed by the presentation of our results in sections 4 through 7, and a final discussion in section 8.

2 Hypotheses about compound stress assignment

Three 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-stress. 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 (Plag 2006, Plag et al. 2007), or for no effect at all (Plag et al. 2008, Lappe & Plag 2008) for argument structure, and a weak across-the-board lexicalization effect.

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 prominence 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-stress. 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*).

There are only a few systematic empirical studies available that investigate the role of semantics in variable compound stress assignment. The earliest one is Sproat (1994), who discusses a variety of methods for stress assignment in English compounds for the purpose of text-to-speech synthesis. The semantic information did not contribute much to successful compound stress classification in Sproat's study, neither in the form of semantic rules, nor in the form of cross-products of semantic categories instantiated in the two constituents.[2] Plag (2006) tested whether the semantic hypothesis makes the right predictions for compounds with a causative relation (as in *Kauffmann sonata*) against a relation that is not predicted by the literature to trigger right-hand stress (as in *Twilight Sonata*). It turned out that the data show either no effect, or show an effect in the opposite direction of what the semantic hypothesis would have predicted. Plag et al. (2007, 2008) tested many more semantic relations and found many effects, some of them new, some of them expected, but not all of the effects predicted by the literature. In general, large parts of the data were ill-behaved. A similar picture emerges from the study of Plag et al. (2008). Although they found a number of robust significant semantic effects, these effects were far from categorical and large parts of the data were unaccounted for.

[2]There are also some serious methodological problems with this study, see Plag et al. (2008) for discussion.

Finally, we turn to the third type of approach, the analogical one. Under this approach 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.

Liberman & Sproat (1992) are, however, the first to pave the way to an empirical method for testing constituent family effects by multiplying the probabilities of a certain type of stress for the two constituents of a compound. Unfortunately, they only give a "representative sample" (Liberman & Sproat 1992:176) of the two constituent families of the compound *safety board*, and do not test their hypothesis on a larger corpus. They simply state that "the method can work fairly well if properly trained. Its main drawback is that many words do not occur often enough in the needed constructions to generate useful statistics." We will show that this statement is too pessimistic. We will present robust statistical evidence in favor of an analogical effect of the left and right constituent families, even if these families are mostly rather small.

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 looked into in two corpus-based studies, Plag et al. (2007), based on data from CELEX, and Lappe & Plag (2007, 2008), based on data from CELEX and from the Boston University Radio Speech Corpus. These studies made use of exemplar-based computational algorithms that tested not only the effect of the left and right constituent, but also of various other properties of compounds, such as semantic and structural ones.

Such exemplar-based models roughly work along the following lines. When a new compound is input to the system in order to be assigned stress, the new compound is compared in all its properties with all the exemplars that are already stored in the lexicon. The algorithm selects the set of compounds that are most similar in its properties to the input. The algorithm then assigns the kind of stress to the input form that is most frequent among this set of most similar compounds. In all three studies mentioned it was found that constituent family is the most successful predictor. However, it also turned out that

the predictions were not always very accurate, and that the prediction of rightward stress was especially problematic, with its accuracy scores reaching no more than 20 percent for the CELEX data and 61 percent for the Boston Corpus data (Lappe & Plag 2008)

The exemplar-based approach raises the interesting question whether the semantic effects on stress assignment found in Plag et al. (2007, 2008) could be explained as an epiphenomenon of the constituent family bias on stress. Gagné & Shoben (1997) and Gagné (2001) provided evidence that the constituent family has an effect on compound interpretation, in that compounds with the same left or right constituent tend to show the same kinds of semantic relationships. For example, compounds with the right constituent *magazine* tend to show the relation 'N2 ABOUT N1', as in *mountain magazine*. The present study addresses the crucial question whether it is the constituent family or the semantics or both that is responsible for stress assignment. If semantics has an independent effect on stress assignment, semantic factors should emerge as independent predictors even in those regression models that include both semantic factors and constituent family stress biases as predictors. We will therefore provide three different kinds of analyses: one based of the effects of constituent family bias alone, one of the effects of all predictors but constituent family bias, and one that includes all factors. As we will see, both constituent family and semantics are independent and significant predictors of stress assignment.

Before turning to the details of our methodology, let us look at some sources of variability not yet discussed. Most previous studies of compound stress assume that, apart from cases of contrastive stress, any given compound, i.e. type, has always the same kind of stress pattern. This assumption is problematic on two grounds. First, there are dialectal differences so that some compounds may be left-stressed in one variety of English, but right-stressed in another variety. It should be noted, however, that, in spite of potential regional differences in the stressing of individual compounds, recent studies using corpora from British English (Plag et al. 2007) and American English (Plag et al. 2008) yielded very similar results with regard to the mechanisms at work in compound stress assignment. The present study also uses corpora from these two varieties.

Second, as discussed, for example, in Bauer (1983:103), Plag et al. (2008), Kunter (2009), there can be within-speaker and across-speaker variation in the stressing of a single type, even within one variety of English. In his systematic study of this kind of variation in the Boston Corpus, Kunter (2009: chapter 8) finds that both within- and across-speaker variation are frequent phenomena. For example, for speaker F3, *morning edition* has 12 left- and 5 right-prominent tokens in the corpus. The types *budget deficit* and *state trooper* are compounds showing across-speaker variation, with about half of their Boston Corpus tokens being left-stressed, the other half right-stressed. Importantly, this variation is not random, and is therefore not attributable to mere 'performance' noise. Future studies will have to substantiate and clarify which factors contribute to the variability or non-variability of certain compounds (see Plag et al. 2008:787 for some hypotheses). The observed within-type variation presents, in any case, a severe problem for rule-based structural or semantic approaches, which mostly rely on the non-variability of stress assignment across tokens and speaker. In the present study, one of our sources, the Boston Corpus, allows us to take into account token-variability (see more on this in the following section).[3]

[3]The variability just discussed has the unfortunate consequence that not all readers will find the stress

3 Methodology

3.1 The corpora

We took our 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). 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 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 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. In addition to testing the individual effect of constituent family for those two corpora, this allows us also 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, 'l' for left, 'r' for right, and 'b' for 'both'. There is some confusion in the literature about how many stress levels should be assumed, and whether, when more than two levels 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), it was shown that rightward stress manifests itself 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', 'even', or, as in this case, 'both'. We have therefore collapsed Teschner & Whitley's 396 'b'-marked items and the 36 'r'-marked 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

patterns of our examples conforming to their own pronunciations or intuitions. The same problem occurs when one compares different dictionaries, which sometimes provide different for individual words. We document here the stresses as gleaned from our sources.

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 24 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 files were digitized at a 16k Hz sample rate using a 16-bit A/D conversion. The orthographic transcripts were generated by hand.

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 we 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 we eliminated proper names such as *Barney Frank* and those with an appositive modifier, such as *Governor Dukakis*. We finally arrived at an overall number of 4353 tokens of noun–noun constructs, representing 2450 word types. As mentioned already above, the data from the Boston Corpus thus present us with two different options. One could analyze tokens, or one could generalize over tokens and provide a type-based analysis. Given that there is also variability within types, a token-based approach seems conceptually superior and more in line with the idea of exemplar-based approaches, since each token contributes to the set of exemplars over which analogies may be computed. In any case, we explored both options and present the results of both type-based and token-based analyses.

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 that 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.

The 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 has 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).

3.2 Computing constituent family biases

In order to test the effect of analogy in compound stress assignment, we used what we call the 'constituent family bias' as manifested in each data source. This bias derives from the proportion of rightward and leftward stresses within a constituent family, and hence can be taken as a measure of the probability of the members of the family to take either left or rightward stress. The biases are computed 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 lead to a considerable reduction in the size of the data, but the remaining data are still large enough to allow serious testing (T&W: $N = 782$ (types), CELEX: $N = 2638$ (types), Boston Corpus: $N = 536$ (types), $N = 1154$ (tokens).

We then computed for the left constituent of each compound in each corpus the stress bias in its constituent family, i.e. the bias with which all other compounds in that corpus that have the same left constituent as our given compound take leftward or rightward stress. We then did the same for each right constituent of each compound. 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*), 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. 0.83333, 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, we do not take into account the stress of the compound in question when computing the family bias for this compound. We do so 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.

We then turned the gradient constituent family biases into three discrete categories. We assigned the value `left bias` for probabilities of leftward stress larger than 70 percent, the value `right bias` for probabilities of leftward stress smaller than 30 percent, and `neutral` for all probabilities between and including 70 and 30 percent.[4] We then

[4]This kind of procedure was also used in the studies by Krott et al. (2001, 2002b, 2007), in which the effect of constituent family on the choice of linking elements in Dutch and German compounds was tested. We also ran analyses based on the gradient biases, i.e. using the proportions directly, but we found

used logistic regression models to estimate the effect of these two variables. To return to our hypothesis, if analogy plays a role, we should find a significant effect of constituent family bias in our regression models. In addition to models that use only family bias as predictors, we also present models based on other, i.e. structural and semantic predictors, and models that are based on all available predictors.

For the statistical analysis we used the statistical package R (R Development Core Team, 2007). The final models we present 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 whether semantic factors and familiy bias are 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). The models presented below show no danger of collinearity, with all VIFs having values below 2. Predictors with VIF values exceeding 2 were removed during model simplification. These cases were rare and are explicitly documented. 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. In all simulations all predictors remained in the models, and only very small corrections of R^2 occured.

4 Results 1: Exploring the data bases

Let us first have a look at the distribution of stresses in the four sources. Table 1 gives these distributions 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 amount as the one we find in our sample from the Boston Corpus news texts.

Table 1: Distribution of stresses across corpora

	T&W	CELEX	Boston Corpus (types)	Boston Corpus (tokens)
leftward stress	700	2483	359	821
rightward stress	82	155	176	333
percent leftward stresses	89.5	94.1	67.1	71.1

Another interesting question is the distribution of family sizes. How large are these families in our corpora? Table 2 illustrates for the Teschner & Whitley corpus that the families are generally quite small, with 60.2 percent of the 782 compounds having left constituent families with only one or two other members, and 63.6 percent having right

basically the same statistically significant effects. We therefore decided to present here the results of the analysis using the categorically transformed biases since these are conceptually more easy to handle.

constituent families with only one or two other members. We will see below that such a small basis for generalizations is still large enough to make fairly good predictions concerning stress assignment.

Table 2: Distribution of constituent family sizes, T&W data

Left constituent										
Family size	2	3	4	5	6	7	8	9	10	11
Frequency	306	165	96	90	30	35	40	9	0	11
Right constituent										
Family size	2	3	4	5	6	7	8	9	10	11
Frequency	326	171	68	85	42	35	16	18	10	11

For CELEX and the two other corpora we find a similar picture. As illustrated in table 3, the CELEX families are again quite small, with the majority of compounds having families with only one or two other members, i.e. two or three members overall.

Table 3: Distribution of constituent family sizes, CELEX data.

Left constituent															
Family size	2	3	4	5	6	7	8	9	10	11	12	13	...	34	37
Frequency	267	157	67	53	23	15	14	11	10	5	5	3	...	1	1
Right constituent															
Family size	2	3	4	5	6	7	8	9	10	11	12	13	...	38	76
Frequency	239	121	59	32	28	16	16	10	10	7	6	5	...	1	1

In the type-based Boston Corpus (see table 4), a similar preponderance of small families can be observed.[5]

[5]The left constituent with the highest number of family members, i.e. 31, is *state*. This family constists of the following items: *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*. This family has a strong bias towards leftward stress, with only 3 out of the 31 compounds having rightward stress.

Table 4: Distribution of constituent family sizes, Boston Corpus, type data

Left constituent													
Family size	2	3	4	5	6	7	8	9	11	12	15	17	31
Frequency	83	33	15	6	4	4	3	2	1	1	1	1	1
Right constituent													
Family size	2	3	4	5	6	7	8	9	11	13			
Frequency	88	44	15	10	9	2	2	1	1	1			

Another interesting question is whether the two families basically agree on their stress biases, or whether there are large numbers of compounds where the bias of the left constituent family and the bias of the right constituent family would work against each other. Tables 5 through 8 crosstabulate the stress biases of the left and right constituent families for the four data sets.

Table 5: T & W

		right constituent family		
		left bias	neutral	right bias
left constituent family	left bias	626	34	28
	neutral	20	2	5
	right bias	44	7	16

Table 6: CELEX

		right constituent family		
		left bias	neutral	right bias
left constituent family	left bias	2306	85	64
	neutral	113	11	3
	right bias	45	7	4

Table 7: Boston Corpus, type data

		right constituent family		
		left bias	neutral	right bias
left constituent family	left bias	197	64	56
	neutral	42	22	16
	right bias	61	37	40

Table 8: Boston Corpus, token data

		right constituent family		
		left bias	neutral	right bias
left constituent family	left bias	548	95	55
	neutral	157	87	75
	right bias	54	32	51

In all four data sets we can see that the biases of the left and right constituent families for leftward stress assignment largely coincide, but that the biases for rightward stress

largely contradict each other. For illustration, let us look at the T & W table. Of all right constituent families, 690 (i.e. 626 + 20 + 44) have a leftward stress bias. In the vast majority of these 690 cases, namely in 626 cases, the compounds with a leftward stress bias in the right constituent family also have a leftward stress bias in their left constituent family. In contrast, of the 49 (28 + 5 + 16) compounds that have right constituent families with a rightward stress bias, 28 compounds have a leftward stress bias in their left constituent family and 16 have a rightward stress bias in their left constituent family.

In other words, if one of the two constituents has a family bias for leftward stress, chances are high that the other constituent family has the same kind of bias. But if one of the constituents has a family with a bias for rightward stress, chances are high that the other constituent shows the opposite tendency in its family. Neutral biases also do not coincide across the two constituents. These tendencies hold for all corpora. Overall, this means that left and right families do not generally provide the same kind of information.[6]

To summarize, our data show enough variation in stress assignment and provide the necessary information on constituent family to make the data an appropriate testing ground for the effect of constituent family on compound stress.

Our investigation addresses three different research questions. First, how well can compound stress be predicted solely on the basis of constituent family information? Second, how does the performance of models based on constituent family only compare with the performance of models using other kinds of predictors, i.e. semantic and structural? Third, how do models perform that have all types of information at their disposal? In particular, which factors survive in such an overall model?

The following three sub-sections address each of the three research questions in turn.

5 Results 2: Stress assignment on the basis of constituent family bias alone

5.1 Teschner & Whitley (2004): family bias alone

According to the hypothesis that left and right family biases determine stress asignment, we should expect a majority of compounds with a bias towards leftward stress to have leftward stress and a majority of compounds with a bias towards rightward stress to have rightward stress. The mosaic plot in figure 1 shows the distribution of left and rightward stresses according to the stress bias of the left and right constituent family. Mosaic plots represent the number of observations in each subset of the data as an area.

[6]This is also evidenced by the generally low variance inflation factors for these two predictors.

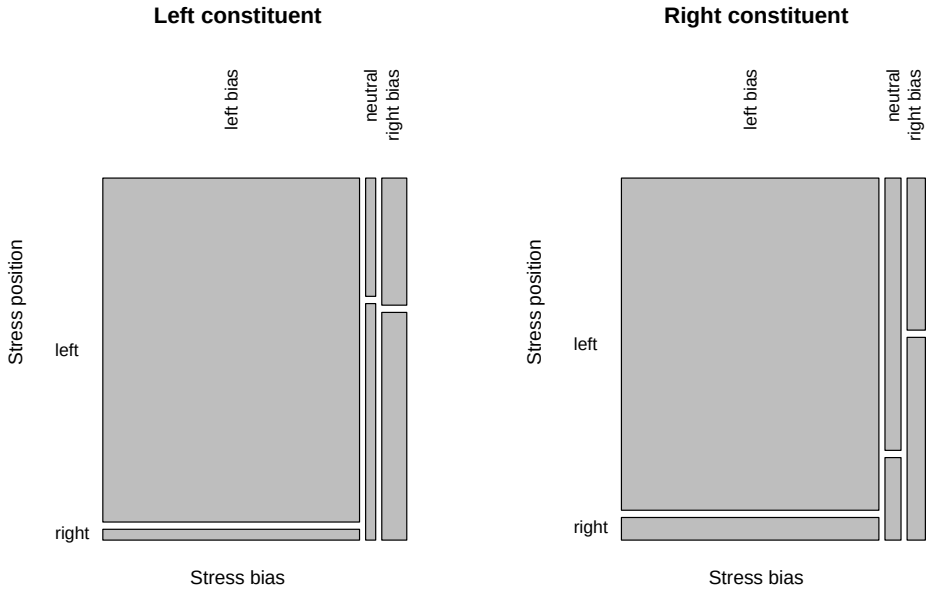

Figure 1: Stress patterns by left and right constituent family bias, T&W data.

Let us first look at the left panel, which shows the effect of the left constituent family. We can see that the vast majority of the compounds with a left constituent family bias take leftward stress (667 out of 688, i.e. 96.9 percent). Compounds with a neutral bias have a two third preponderance of rightward stresses (18 out of 27, i.e. 66.7 percent), and compounds with a bias for rightward stress show almost the same behavior as those with a neutral left bias (43 out of 67, i.e. 64.2 percent). A similar story can be told for the effect of the right constituent family, as shown in the right panel of figure 1. A bias in the right family for leftward stress goes together with a vast majority of leftward-stress compounds, compounds with a neutral bias in their right constituent family still favor leftward stress, but, crucially, the majority of compounds with a family bias for rightward stress have rightward stresses (57.1 percent).

In a logistic regression analysis with STRESS POSITION as the dependent variable and LEFT CONSTITUENT BIAS and RIGHT CONSTITUENT BIAS as the two predictor variables, both biases turn out to be highly significant. [7]. There were only the two main effects and no interaction between the two predictors. The full model is documented in table 9. Here and in the models to follow, positive coefficients indicate changes in the logits in the direction of rightward stress. Note that the overall fit of the model is very good (cf., for example, $C = 0.906$).[8]. Interestingly, the effect of the left constituent family bias is

[7] A full documentation of the ANOVA can be found in the appendix, in table 26

[8] C is a measure of the discriminative power of the logistic regression model and is the percent of all possible pairs of cases in which the model assigns a higher probability to a correct case than to an incorrect case. C ranges from 0.5 to 1.0, with 1.0 showing perfect matches of high probabilities and correct classification. Standardly, values of 0.9 indicate excellent fit, values between 0.8 and 0.9 indicate a good fit of the model. Technically, C is the area under a Receiver Operating Characteristics (ROC) curve (see, e.g., Fawcett 2003)

stronger than that of the right constituent family bias. While this seems to run counter to intuition based on the above-mentioned textbook examples (cf. again the effect of *street* vs. *avenue* as right constituents), the existence of both left and right constituent effects was claimed to exist by Liberman & Sproat (1992). Our analysis provides the first empirical validation for this claim.

Table 9: Logistic regression model with left and right constituent bias as predictors, T&W data

	Estimate	Std. Error	z value	Pr(>|z|)	Odds Ratio
(Intercept)	-4.0585	0.2906	-13.96	0.0000	0.01727517
leftConstituentBias=neutral	4.3482	0.5146	8.45	0.0000	77.33752975
leftConstituentBias=right bias	4.0522	0.3866	10.48	0.0000	57.52442419
rightConstituentBias=neutral	1.5353	0.5695	2.70	0.0070	4.64293857
rightConstituentBias=right bias	3.1149	0.4713	6.61	0.0000	22.53068923

Obs	Max Deriv	Model L.R.	d.f.	P	C	Dxy	Gamma	Tau-a	R^2	Brier
782	5e-14	262.44	4	0	0.906	0.812	0.936	0.153	0.583	0.047

5.2 CELEX: family bias alone

The mosaic plot in figure 2 shows the distribution of left and rightward stresses according to the stress bias of the left and right constituent family for the CELEX data.

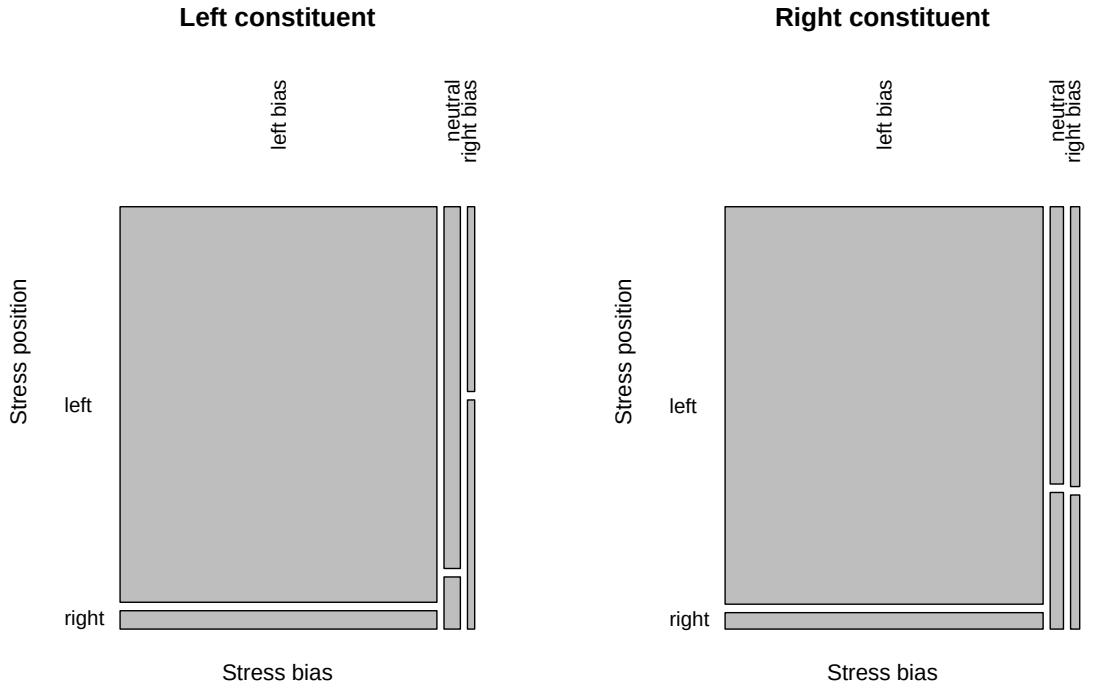

Figure 2: Stress patterns by left and right constituent family bias, CELEX data.

In the left panel we see that the vast majority of the compounds with a left constituent family bias for leftward stress actually take leftward stress (2347 out of 2455, i.e. 95.5 percent), and compounds with a neutral bias also tend heavily towards leftward stresses (111 leftward stresses out of 127, i.e. 87.4 percent). Compounds with a left constituent bias for rightward stress show only a slight tendency towards rightward stress (31 out of 56, i.e. 55.4 percent). The effect of the right constituent family is similar apart from those compounds that have a bias for rightward stress. They fare no better in taking rightward stress than those with a neutral bias, and still tend towards leftward stress (67.6 percent leftward stresses, 48 out of 71), as shown in the right panel of figure 2.

In a logistic regression analysis with STRESS POSITION as the dependent variable and LEFT CONSTITUENT BIAS and RIGHT CONSTITUENT BIAS as the two predictor variables, both biases turn out to be highly significant[9]. There were only the two main effects and no interaction between the two predictors. The model is documented in table 10. The model predicts the probability of rightward stress. Although the model is highly signifcant, its overall fit is not very satisfactory ($C = 0.753$). Again, the effect of the left constituent family bias is stronger than that of the right constituent family bias.

[9]A full documentation of the ANOVA can be found in the appendix, in table 27

Table 10: Logistic regression model with left and right constituent bias as predictors, CELEX data

	Estimate	Std. Error	z value	Pr(>|z|)	Odds Ratio
(Intercept)	-3.4925	0.1204	-29.01	0.0000	0.03042444
leftConstituentBias=neutral	1.0392	0.3102	3.35	0.0008	2.82694402
leftConstituentBias=right bias	3.3557	0.3093	10.85	0.0000	28.66630490
rightConstituentBias=neutral	2.4491	0.2521	9.72	0.0000	11.57819274
rightConstituentBias=right bias	2.5364	0.2917	8.69	0.0000	12.63449219

Obs	Max Deriv	Model L.R.	d.f.	P	C	Dxy	Gamma	Tau-a	R^2	Brier
2638	2e-09	241.76	4	0	0.753	0.506	0.828	0.056	0.243	0.046

5.3 Boston Corpus, type data: family bias alone

The mosaic plot in figure 3 shows the distribution of left and rightward stresses according to the stress bias of the left and right constituent family in the type-based Boston Corpus.

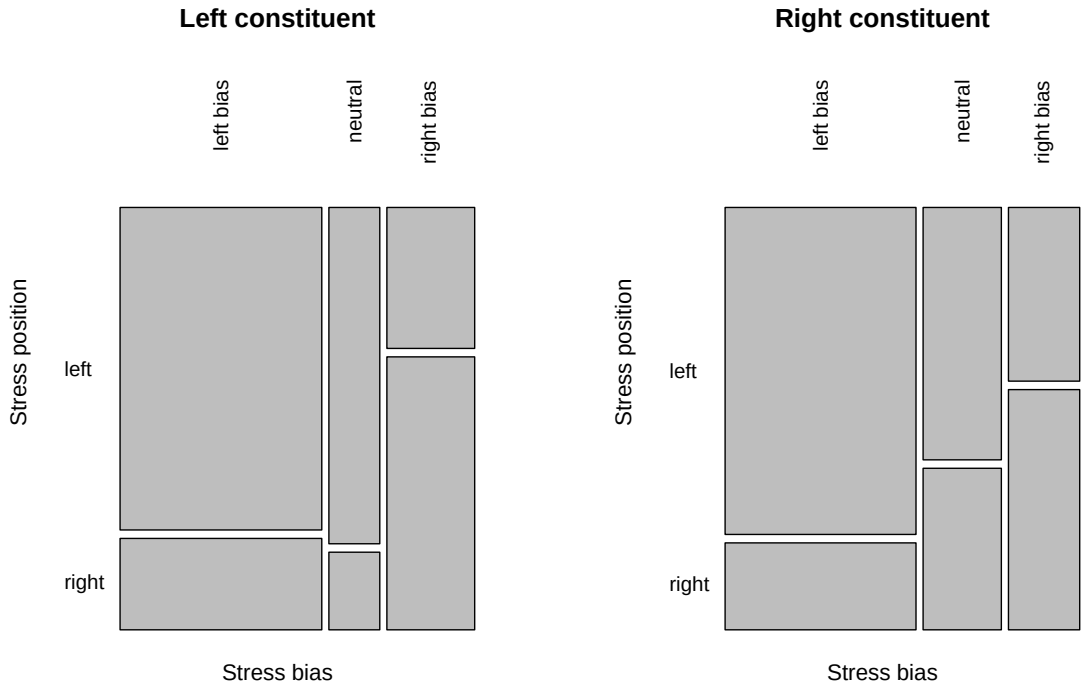

Figure 3: Stress patterns by left and right constituent family bias, Boston corpus, type data.

Let us first look at the left panel, which shows a very clear effect of left constituent

family bias. The vast majority of the compounds with a left constituent family bias for leftward stress actually take leftward stress, and compounds with a right bias in the left constituent family have a strong tendency for rightward stress. Compounds with a neutral bias tend toward leftward stress. Simlarly, as shown in the right panel of figure 3, a left bias in the right contituent family leads to mostly leftward stress, a right bias to a majority of rightward stresses. A neutral bias leads to a more even distribution, with a slight preponderance of leftward stresses.

In a logistic regression analysis with STRESS POSITION as the dependent variable and only LEFT CONSTITUENT BIAS and RIGHT CONSTITUENT BIAS as the two predictor variables, both biases emerge as highly significant[10]. There were only the two main effects and only one significant interaction (namely between neutral bias and right constituent family, as already hinted at above in the discussion of the mosaic plot). Since none of the interactions was overall significant in an ANOVA, the interaction between neutral bias and right constituent family was dropped from the model. The resulting model is documented in table 11. The model predicts the probability of rightward stress. Although the model is highly signifcant, its overall fit is not too impressive, though slightly better than that of the CELEX data (cf. the C-values of 0.778 vs. 0.753). Again, the effect of the left constituent family bias is stronger than that of the right constituent family bias.

Table 11: Logistic regression model with left and right constituent bias as predictors, Boston Corpus type data

	Estimate	Std. Error	z value	Pr(>|z|)	Odds Ratio
(Intercept)	-1.8225	0.1808	-10.08	0.0000	0.1616135
leftConstituentBias=neutral	-0.3255	0.3300	-0.99	0.3240	0.7221937
leftConstituentBias=right	1.8686	0.2363	7.91	0.0000	6.4791249
rightConstituentBias=neutral	0.8269	0.2560	3.23	0.0012	2.2862952
rightConstituentBias=right	1.6308	0.2609	6.25	0.0000	5.1081315

Obs	Max Deriv	Model L.R.	d.f.	P	C	Dxy	Gamma	Tau-a	R^2	Brier
535	1e-07	130.94	4	0	0.778	0.557	0.631	0.246	0.302	0.167

5.4 Boston Corpus, token-based: family bias alone

The mosaic plots in figure 4 show that the vast majority of the compounds with a left constituent family bias for leftward stress actually take leftward stress (593 out of 698, i.e. 85 percent). Compounds with a neutral bias have a much more even distribution (188 leftward stresses out of 319, i.e. 59 percent), and compounds with a bias for rightward stress show a clear majority of rightward stresses (97 out of 137, i.e. 71 percent). Basically the same story can be told for the effect of the right constituent family, as shown in the right panel. A bias in the right family for leftward stress goes together with a vast majority of leftward-stress compounds, compounds with a neutral bias in their right constituent

[10]A full documentation of the ANOVA can be found in the appendix, in table 28

family show a more variable behavior, and compounds with a family bias for rightward stress have a clear tendency towards rightward stress.

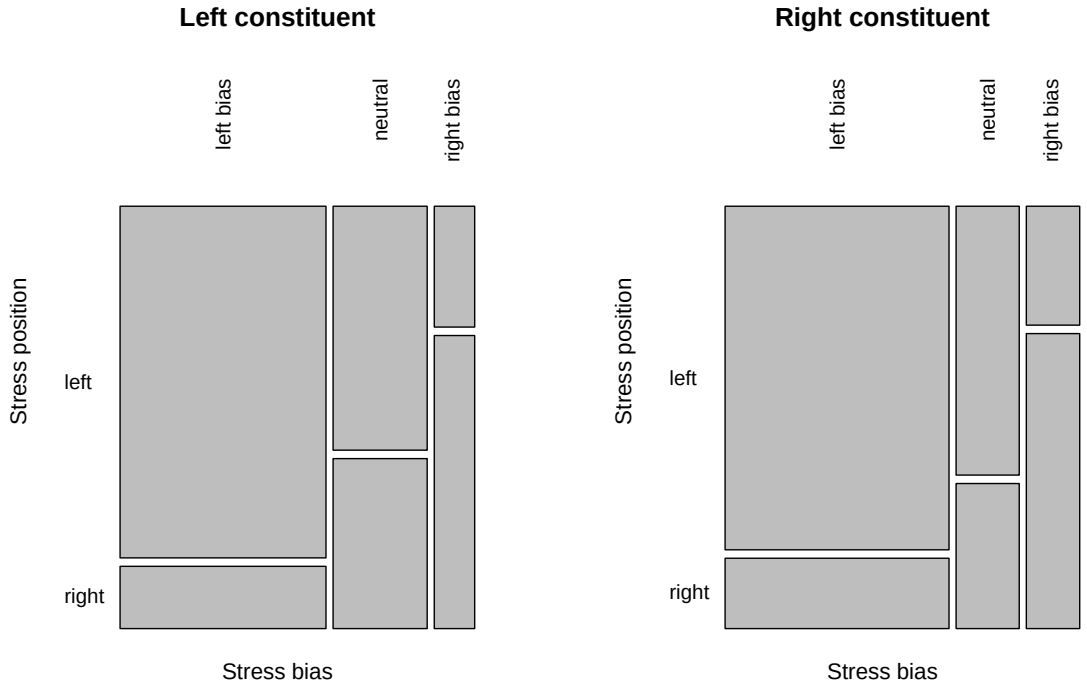

Figure 4: Relation between consitutent family bias and stress assignment, Boston Corpus, token data

In order to examine the effect of constituent family more closely we again performed a logistic regression analysis with STRESS POSITION as dependent variable and LEFT CONSTITUENT BIAS and RIGHT CONSTITUENT BIAS as categorical predictors. Both factors turned out to be highly significant[11]. As was the case in the earlier analyses, the left constituent bias shows the stronger effect, and there was no significant interaction. The model is summarized in table 12.

[11] A full documentation of the ANOVA can be found in the appendix, in table 29

Table 12: Logistic regression analysis using left and right constituent family biases as predictors, Boston Corpus, token data

	Estimate	Std. Error	z value	Pr(>|z|)	Odds Ratio
(Intercept)	-2.1047	0.1235	-17.04	0.0000	0.1218823
rightConstituentBias=neutral	0.6541	0.1871	3.50	0.0005	1.9235057
rightConstituentBias=right bias	2.0991	0.2037	10.31	0.0000	8.1586477
leftConstituentBias=neutral	1.0510	0.1689	6.22	0.0000	2.8604248
leftConstituentBias=right bias	2.2441	0.2304	9.74	0.0000	9.4318879

Obs	Max Deriv	Model L.R.	d.f.	P	C	Dxy	Gamma	Tau-a	R^2	Brier
1154	6e-14	312.27	4	0	0.799	0.599	0.696	0.246	0.339	0.148

We can see that the differences between the different levels of both factors are all highly significant, which means that there is a strong relation between constituent family bias and the position of stress. The overall effect of the bias is satisfactory ($C = 0.799$).

To summarize, constituent family turned out to be a significant predictor in all four corpora, with the bias of the left constituent having generally a greater effect size than the bias of the right constituent. We now take a look at how well the models actually predict the stress.

5.5 Prediction accuracies across corpora: family bias alone

To make the model estimates from above more tangible, and to more easily compare the accuracies of the models in assigning categorical stress we transformed the estimated probabilities of left and rightward stresses into categorical decisions. If the probability of rightward stress for a given compound as estimated by the model was less than 0.5 for a given item, we interpreted this item as left-stressed, and as rightward-stress if otherwise. These predictions were then compared to the stress positions as found in the corpus – a match was counted as a correct prediction. In table 13 we have listed the accuracy scores of all models that only contained the two constituent biases as predictors.[12] We also give the C values for a comparison of the overall model fits.

[12]There is a rich literature on the different measures of performance for classifying algorithms in the field of machine learning (see, for example, Demšar (2006) for a survey). Such measures are usually built from a confusion matrix in which correctly and incorrectly classified examples are recorded. Frequently used measures are accuracy, precision, recall, F-score and ROC analysis (but see Sokolova et al. (2006) for alternative measures). Since the primary interest in this paper does not lie in the intricacies of the classificatory performance of the different regression models, but rather on the role or non-role of certain variables as significant predictors, we do not document and discuss the different measures, but restrict ourselves to the probably most intuitive of these measures, accuracy. We use the term accuracy as the percentage of correctly classified instances. In technical terms, accuracy refers to either the percentage of true positives among all positives, the percentage of true negatives among all negatives, or to the percentage of the sum of the true positives and true negatives among all instances.

Table 13: Comparison of accuracy scores across corpora and approaches, based on constituent family information only.

	T&W	CELEX	Boston types	Boston tokens
C	0.906	0.753	0.778	0.799
Overall accuracy in percent	93.1	94.4	75.3	80.3
Leftward stress accuracy	98.4	99.9	86.9	93.7
Rightward stress accuracy	47.6	6.1	51.7	47.7

The figures for the overall accuracy show that, across corpora, knowledge of how the constituent families of a given compound are stressed suffices to rather successfully predict the assignment of stress. The overall fit of the models is not bad, and is comparable across the CELEX and Boston corpora, with the T&W model clearly outperforming the other three. The categorically transformed model estimates, however, are much better for the dictionary data with accuracies of 93.1 and 94.4 percent correct predictions as against only 75.3 and 80.3 for the two Boston Corpora. The figures in the third and fourth row indicate that the prediction of rightward stress is not nearly as successful as the prediction of leftward stress. This is especially true for CELEX, which is the corpus with the smallest proportion of rightward stresses, and the smallest proportion of correct predictions for rightward stress.

6 Results 3: Stress assignment on the basis of other predictors

We now turn to the effects of predictors other than constituent family. For the CELEX and Boston Corpus compounds Plag et al. (2007, 2008) coded each compound according to the categories held to be responsible for stress assignment in the literature (and some more), and we will use these codings in this subsection to check their predictive power for stress assignment for the same data sets from these two corpora that we used in the previous section. Which properties were coded? 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[13] Furthermore, the factor SPELLING is coded as a proxy of lexicalization (with the values `1` for one-word, `h` for hyphenated, and `2` for two-word spellings).[14] With regard to semantic properties,

[13]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.

[14]As discussed in detail in Plag et al. 2007, 2008, one-word spellings should be most prevalent with lexicalized compounds, while less lexicalized compounds should prefer two-word spellings. We are aware that a connection between spelling and lexicalization does not mean that stress would be dependent on orthography (to the effect that only literate speakers would know how to stress correctly). Rather, given the options of English spelling, speakers would express their intuition that a given compound is felt to be more or less integrated by choosing a more or less integrated spelling. 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 written as two words are much more variable in their stress pattern.

each compound is coded according to the following categories, all of which are mentioned in the literature (e.g. Fudge 1984:144ff, Gussenhoven & Broeders 1981, Liberman and Sproat 1992, Zwicky 1986) to trigger rigthward stress:

(2) 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. 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. In table 14 we present the final list of the semantic relations coded. The relations are expressed by supposedly language-independent predicates that link the concepts denoted by the two constituents (see Levi 1978 for discussion). Table 14 gives the 18 semantic relations. A subset of these, as given in table 15 have been claimed to trigger rightward stress (e.g. Fudge 1984:144ff, Zwicky 1986, Liberman and Sproat 1992).

Table 14: List of semantic relations coded, illustrated with one example 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	*firelight*
7.	N2 IS MADE OF N1	*potato crisp*
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 15: 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*

In the following subsections we discuss for each corpus how well the overall 26 different structural and semantic predictors can predict compound stress assignment. At the end of this section we will compare the accuracies of the models with these predictors to the accuracies of the models that use only constituent family as predictors.

6.1 CELEX: other predictors

We fitted a logistic regression analysis with the structural and semantic predictors to the CELEX data. The final regression models is given in table 16 (the ANOVA is again documented in the appendix, in table 30).[15]

Table 16: Final logistic regression model, based on all predictors but family bias, CELEX data

	Estimate	Std. Error	z value	Pr(>|z|)	Odds Ratio
(Intercept)	-4.4402	0.2545	-17.45	0.0000	
spellNew=separate	2.4118	0.2542	9.49	0.0000	
semRel4=yes	1.1781	0.2429	4.85	0.0000	
semRel7=yes	1.4011	0.2348	5.97	0.0000	
semRel12=yes	-1.6242	0.2519	-6.45	0.0000	
semRel16=yes	1.2642	0.3732	3.39	0.0007	

Obs	Max Deriv	Model L.R.	d.f.	P	C	Dxy	Gamma	Tau-a	R^2	Brier
2638	3e-10	246.12	5	0	0.832	0.664	0.745	0.073	0.247	0.049

Of the 26 predictors, only five survive as significant in the logistic regression analysis. The five are spelling, 'N1 HAS N2', 'N2 IS MADE OF N1', 'N2 FOR N1' (the only semantic predictor with a tendency towards leftward stress), and 'N2 DURING N1'. The fit of the model is good ($C = 0.868$).

[15]The variance inflation factors for SPELLING were 2.11 for two word spellings and 2.05 for hyphenated compounds. Since these values exceeded the threshold value of 2 this could be taken as an indication that the two values tap essentially the same phenomenon. It was therefore devised to collaps the two levels into a single one comprising both separate spellings. This recoded variable is named 'spellNew' in the ANOVA table. The variance inflation factor for the recoded binary variable was 1.02

6.2 Boston Corpus, types: other predictors

This analysis presents the problem that for many of the semantic predictors we have very few observations. This may be one reason why in the final model only three factors survive, 'N1 is a proper noun', the righthand morpheme, and 'N1 HAS N2'. The final model is documented in table 17, the ANOVA again in the appendix, in table 31.

Table 17: Final logistic regression model based on all predictors but family bias, Boston Corpus data, types

	Estimate	Std. Error	z value	Pr(>|z|)	Odds Ratio
(Intercept)	-1.4596	0.2556	-5.71	0.0000	0.2323215
isN1PNyes_N1pn	0.6176	0.3115	1.98	0.0474	1.8544703
morphRight=er	0.3602	0.3974	0.91	0.3647	1.4335783
morphRight=ing	1.8861	0.6258	3.01	0.0026	6.5936261
morphRight=ion	1.1202	0.5505	2.03	0.0419	3.0654052
morphRight=none	0.4019	0.2749	1.46	0.1438	1.4946343
semRel4=yes	0.9258	0.2059	4.50	0.0000	2.5239383

Obs	Max Deriv	Model L.R.	d.f.	P	C	Dxy	Gamma	Tau-a	R^2	Brier
535	4e-12	37.62	6	0	0.657	0.313	0.394	0.139	0.095	0.205

Compounds which belong to the two semantic categories are significantly more rightward-stressed than all other compounds, and compounds that end in *-ion* or *-ing* are significantly more rightward-stressed than those that have converted right constituents. The overall fit of the model is not impressive at all ($C = 0.657$).

6.3 Boston Corpus, tokens: other predictors

Testing the power of the traditional semantic and structural predictors on this data set yielded the following results. After the usual model simplification, we ended up with seven significant predictors (see table 32 in the appendix) and no great fit ($C = 0.723$). Proper noun status, spelling, and five semantic relations have a significant effect. The final model is documented in table 18.

Table 18: Final logistic regression model based on all predictors but family bias, Boston Corpus token data

	Estimate	Std. Error	z value	Pr(>|z|)	Odds Ratio
(Intercept)	-4.8229	0.7223	-6.68	0.0000	0.008043699
isPropN=True	2.0830	0.6431	3.24	0.0012	8.028719798
spell=2	3.3855	0.7192	4.71	0.0000	29.533114896
semRel2=yes	0.6095	0.2675	2.28	0.0227	1.839463671
semRel4=yes	1.0454	0.1466	7.13	0.0000	2.844545985
semRel6=yes	0.7036	0.3081	2.28	0.0224	2.020981541
semRel12=yes	0.2961	0.1461	2.03	0.0427	1.344654008
semRel14=yes	0.7564	0.2089	3.62	0.0003	2.130549027

Obs	Max Deriv	Model L.R.	d.f.	P	C	Dxy	Gamma	Tau-a	R^2	Brier
1147	8e-06	183.68	7	0	0.723	0.446	0.507	0.184	0.211	0.176

6.4 Prediction accuracies across corpora: other predictors

To finish our discussion of other predictors of compound stress assignment, we compare the results for the different corpora with each other, and with the results of the models that had only constituent family as predictors. For ease of exposition, we only compare the model fits. Consider table 19.[16]

Table 19: Comparison of model fits across corpora and approaches: based on constituent family bias only vs. all predictors but constituent family bias.

	CELEX	Boston types	Boston tokens
C based on constituent family only	0.753	0.778	0.799
C based on other predictors only	0.832	0.657	0.723

We find a mixed picture. For the CELEX data the model fit is much better if we use the other predictors, while for the two Boston Corpus data sets the fit on the basis of constituent family is better. The interesting question is of course what happens if we take both kinds of information into account. This will be done in the following subsection.

7 Results 4: Stress assignment using all predictors

In section 5 it was shown that taken in isolation, constituent family bias is a significant predictor for compound stress assignment across corpora and kinds of data. Similarly, we have found some effects for other predictors, thus partially replicating results from earlier studies that used the same two corpora but with the full set of forms. Recall that our sets are subsets from these corpora because we used only those compounds that had

[16]The differences between the two competing models for each data set are all significant (ANOVA results: $p = 0.04$ (CELEX), $p = 0.00$ (Boston, type data), $p = 0.00$ (Boston, token data))

families for both of their left and right constituents. In view of the multiplicity of factors that have been shown to have an effect on compound stress assignment it is crucial to assess the significance and predictive power of the many different factors in a single model that is based on all possible predictors. In particular, such an anlysis can show whether semantic or structural effects are just epiphenomenal of constituent family effects, or the other way round. If both types of factor survive as significant in a regression model, there is good reason to believe that they are independently doing their work. The following subsections will explore this.

7.1 CELEX: all predictors

We fitted a logistic regression model with all orthographic, semantic and structural criteria to the data. After the removal of insignificant predictors, a final model with seven predictors emerged. The model is documented in tables 33 (see appendix) and 20.

Table 20: Logistic regression model using all predictors, CELEX data

	Estimate	Std. Error	z value	Pr(>|z|)	Odds Ratio
Intercept	-4.7727	0.2723	-17.53	0.0000	0.008457266
spell=2	2.6764	0.2821	9.49	0.0000	14.532424165
spell=h	1.3806	0.3024	4.57	0.0000	3.97718209
semRel4=yes	0.9432	0.2860	3.30	0.0010	2.568179361
semRel7=yes	1.3167	0.2694	4.89	0.0000	3.731263134
semRel12=yes	-1.4753	0.2769	-5.33	0.0000	0.228699027
semRel16=yes	1.5389	0.3972	3.87	0.0001	4.659278213
leftConstBias=neutral	0.6237	0.3318	1.88	0.0601	1.865795364
leftConstBias=right bias	2.6301	0.3648	7.21	0.0000	13.875756173
rightConstBias=neutral	2.3751	0.2932	8.10	0.0000	10.752299936
rightConstBias=right bias	1.4810	0.3282	4.51	0.0000	4.397184222

Obs	Max Deriv	Model L.R.	d.f.	P	C	Dxy	Gamma	Tau-a	R^2	Brier
2638	7e-08	431.17	10	0	0.899	0.798	0.82	0.088	0.418	0.039

We can see from the coefficients and the odds ratios in the table that the effect of the left constituent bias is stronger than that of the right constituent, and that, apart from spelling (which is indeed the strongest factor), all other factors are much weaker than the constituent family biases. Note that four of the five other significant predictors were also found to show a significant effect in Plag et al. (2007), where the full set of NN compounds from CELEX was used. These predictors are SPELLING and the following three semantic relations: '4. N1 HAS N2', '7. N2 IS MADE OF N1', and '16. N2 DURING N1'.

7.2 Boston Corpus, type data: all predictors

We again performed a logistic regression analysis with all predictor variables included. In the final model, only the two biases and only two of the other predictors (the semantic

relations 'N1 HAS N2', 'N2 ABOUT N1') survived (see 34 in the appendix for full documentation of the ANOVA). Notably, the factor SPELLING was again insignificant for this data set. The model, which is documented in table 21, has a somewhat better fit than the model with only family bias ($C = 0.794$ as against $C = 0.778$).

Table 21: Logistic regression model using all predictors, Boston Corpus, type data

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.0460	0.1998	-10.24	0.0000
semRel4=yes_4	0.6090	0.2358	2.58	0.0098
semRel13=yes_13	0.6719	0.3174	2.12	0.0343
leftConstituentBias=neutral	-0.3364	0.3345	-1.01	0.3144
leftConstituentBias=right	1.8175	0.2445	7.43	0.0000
rightConstituentBias=neutral	0.7298	0.2603	2.80	0.0051
rightConstituentBias=right	1.6211	0.2639	6.14	0.0000

Obs	Max Deriv	Model L.R.	d.f.	P	C	Dxy	Gamma	Tau-a	R^2	Brier
535	5e-07	140.89	6	0	0.794	0.587	0.619	0.26	0.322	0.163

7.3 Boston Corpus, token data: all predictors

In a logistic regression that includes all predictors, a final model emerges that has the two biases and five additional factors as significant predictors (see table 35 in the appendix for full doumentation of the ANOVA). However, its fit is only slightly improved as against the one we get if we take only family bias as predictor ($C = 0.828$ as against $C = 0.799$, $p = 0.00$, *anova*). In other words, the additional five predictors add very little to the success of the model, and we see that their effect sizes (apart from that of spelling) are very small. The model is documented in table 22.[17]

[17]Possibly due to the very low number of hyphenated observations in the data set, the value `hyphenated` for the factor SPELLING did not reach significance. It was therefore decided to remove the seven hyphenated items from the data set (instead of to recode them, as we would have done with if their number had been much higher, as was the case with the CELEX data above). After the removal of the seven items, we are still left with 1147 observations.

Table 22: Logistic regression model using all predictors, Boston Corpus, token data

	Estimate	Std. Error	z value	Pr(>|z|)	Odds Ratio
(Intercept)	-4.6471	0.7188	-6.47	0.0000	0.009589455
leftConstituentBias=neutral	0.6293	0.1814	3.47	0.0005	1.876242915
leftConstituentBias=right bias	2.0089	0.2370	8.48	0.0000	7.454923133
rightConstituentBias=neutral	0.3969	0.1938	2.05	0.0406	1.487214955
rightConstituentBias=right bias	1.8448	0.2110	8.74	0.0000	6.326837490
isPropN=True	1.5791	0.7254	2.18	0.0295	4.850793422
spell=2	2.5582	0.7256	3.53	0.0004	12.911894285
semRel2=yes	0.6139	0.2924	2.10	0.0358	1.847557624
semRel4=yes	0.7688	0.1717	4.48	0.0000	2.157197142
semRel6=yes	0.7520	0.3428	2.19	0.0282	2.121329347

Obs	Max Deriv	Model L.R.	d.f.	P	C	Dxy	Gamma	Tau-a	R^2	Brier
1147	6e-06	370.28	9	0	0.828	0.656	0.689	0.271	0.394	0.141

To summarize, when all predictors are taken into account, the constituent family bias emerges as a robust and significant effect across all corpora. Of the other predictors, some semantic predictors also come out as significant, but their effect sizes are much weaker than that of family bias. Effects of argument structure or morphology did no longer emerge. The effect of spelling was, however, very strong in those data sets where is significant, i.e. in the two larger data sets.

8 Summary and conclusion

Table 23 gives an overview of the fit of all models presented above.

Table 23: Comparison of model fits across corpora and approaches

	CELEX	Boston types	Boston tokens
C based on constituent family only	0.753	0.778	0.799
C based on other predictors only	0.832	0.657	0.723
C based on all predictors	0.899	0.794	0.828

Across all corpora, the models that include all predictors are significantly more successful than those that use only one set of predictors. As we saw in the previous section, family bias, lexicalization and semantics are independent significant predictors of noun-noun stress in English. Given all available information sources, family bias and spelling are the most important of these. Argument structure does not play an independent role. Together, these results provide very robust evidence for an important and independent role of analogy in stress assignment to compounds. At the same time, our findings suggest that semantic effects are not epiphenomenal to constituent family effects, but exist alongside of them. For a model of grammar and lexicon this could be interpreted as evidence

for the idea that generalizations across lexical items emerge at all levels of representation, and that language users, or rather their minds, make use of all kinds of information. While this may make life harder for theorists looking for lean and parsimonious models of processing, it is in line with the bulk of more recent psycholinguistic research on lexical processing, as captured in Libben's (2006) term 'maximization of opportunity'.

How do these results, obtained through regression analysis, compare to analyses of compound stress that use deterministic rules along the lines of the structural or semantic hypothesis, or to exemplar-based computational algorithms?

The application of the structural and semantic rules as proposed in the literature and summarized in section 2 is rather straightforward. For the structural rule we assign leftward stress if N1 is an argument of N2, as in *opera singer*, and assign rightward stress elsewhere (e.g. *steel bridge*). For the different semantic rules we assign rightward stress in cases in which the semantics favours rightward stress (e.g. 'N2 is located at N1' *town house*, or 'N2 is made of N1' *steel bridge*), and leftward stress elsewhere.

With regard to exemplar-based algorithms, the methodology is more complex, but need not be discussed here. We report instead the results from the investigation by Lappe & Plag (2008), who used the same subset of the CELEX data base and the type-based Boston Corpus to test analogy with TiMBL (Daelemans et al. 2007) and A::M (Skousen et al. 2002). In Lappe & Plag's study, the most successful models were those that were exclusively based on constituent family information. Models that contained all predictors as well as models that contained only structural and semantic information were significantly less successful in predicting stress correctly. Let us therefore compare Lappe & Plag's accuracy scores with those from our study. Ours are parallel to theirs in either being based on the same predictors (i.e. only constituent family), or being derived from the most successful models (i.e. based on all predictors). Table 24 presents the relevant figures.[18] In the last section of the table we report the scores that arise from the application of the structural and semantic rules, if applied in the determinstic fashion described above.

Table 24: Comparison of scores of overall accuracies across corpora and approaches. 'L &P' stands for Lappe & Plag (2008)

	CELEX	Boston Corpus type data
Exemplar-based modeling		
L & P (AM model)	94.9	80.4
L & P (TiMBL model)	94.3	77.2
Regression		
Only family bias	94.4	75.3
Family bias, lexicalization and semantics	94.9	77.8
Determinstic rules		
Rule-based overall accuracy (argument structure)	19.1	41.4
Rule-based overall accuracy (semantics)	70.1	59.3

[18]The accuracy scores for our most successful models (see section 7) were computed in the same way as the accuracy scores for the models with only constituent bias from section 5.

The table nicely shows that, across corpora, knowledge of how the constituent families of a given compound are stressed suffices to rather successfully predict its stress pattern. Family bias emerges as the most important predictor of compound stress in Lappe & Plag (2008) and, depending on the corpus, as a very important (CELEX) or most important (Boston Corpus) predictor, independent of the kind of model that is being employed. Both regression analysis as used in this paper, and exemplar-based modeling as employed by Lappe & Plag reach almost the same levels of accurracy when stress assignment is based solely on constituent family information. The present study therefore provides strong empirical evidence that constituent family effects are not methodological artefacts. The table also shows that deterministic, rule-based approaches are hopelessly inadequate for the task of assigning stress correctly. This finding is line with the most recent, empirical studies of compound stress (e.g. Bell 2008, Kunter 2009, Plag et al. 2007, 2008).

However, we saw earlier that our models had considerable difficulties assigning rightward stress correctly. Table 25 allows a closer look at that problem.

Table 25: Comparison of accuracy scores for rightward stress across corpora and approaches. 'L &P' stands for Lappe & Plag (2008)

	CELEX	Boston Corpus type data
Exemplar-based modeling		
L & P (AM model)	19.9	60.8
L & P (TiMBL model)	19.2	50.0
Regression		
Only family bias	6.1	51.7
Family bias, lexicalization and semantics	31.6	50.0

For the Boston Corpus, the accuracies across all models is almost the same at around 50 percent, but for CELEX we find some differences. In regression, family bias alone makes overwhelmingly wrong predictions, the regression model with all predictors predicts at least about a third of the rightward stresses correctly, but flipping a coin would have been much better still. One can only speculate about the reasons why rightward stresses are so hard to predict. The low proportion of rightward stresses in CELEX is of course a problem. The T & W corpus has a similarly low proportion of rightward stresses, but the proportion of correctly predicted rightward stresses is in the same range as the ones for the Boston Corpus (see again table 13 above). Thus, the extremely low predictability of rightward stresses seems to be a peculiarity of the CELEX corpus. This leaves us still with the unsatisfactory accuracy in the other corpora, which does not exceed chance level. At present we have no good explanation to offer for this fact.

Having shown the robust effect of constituent family across corpora and methodologies, the question may arise how an analogical approach can account for completely new formations, for which we might not have two constituent families available that may help us to assign stress to the new compound. After all, the treatment of novel expressions is what the human language faculty is all about, and under a traditional rules-based approach the treatment of new expressions is no problem , but is rather what this mecha-

nism is designed for. So how would stress assignment for completely new formations work under the absence of a rule?

One can distinguish two cases, and we will discuss each in turn. In the first case, at least one of the two nouns has occurred in other compounds before. In this case, we have constituent family information for one constituent. Would that be enough for stress assignment? Of course it would have to be enough in terms of constituent family information, but we have also seen that other kinds of factor are also at work, and these factors would then perhaps gain more weight. But it is an interesting question (and an empirical one) whether this reduced type of constituent family information is still able to make correct predictions. In order to test this in an at least exploratory fashion, we took the whole set of compounds from Teschner & Whitley and computed the constituent family in such a way that we included all compounds that had a constituent family for at least one of the two constituents.[19] This enlarged the data set from 782 to 1138, with 87.9 percent left stresses (as against 89.5 percent in the more restricted data set). A logistic regression model was fitted to this data set in the same way as for the enlarged data set. The new model has an even slightly better fit than the model for the restricted data set ($C = 0.916$ vs. $C = 0.906$), and its accuracy scores based on the categorically transformed estimated probabilities are very close to that of the restricted set (92.6 percent vs. 93.1 overall accuracy), or even better (71.1 vs. 47.6 for rightward stress prediction). The latter finding may also give a hint concerning the explanation of the unsatisfactory prediction of rightward stresses in our models. The reason may simply be the limited amount of information on right-stressed items in the more restricted data sets. Needless to say, this point would have to be investigated more closely in future studies that make use of less restricted data sets.

The second case of missing information would be the, rather unlikely, case[20] that neither the left nor the right constituent had a constituent family from which stress information could be gleaned. Would this make analogical stress assignment impossible? It would not, since analogies could be computed over other types of information. And we even know which other types of information may be involved: lexicalization and semantics (and perhaps many other properties, such as phonological similarities, morphological similarities etc.). In sum, the treatment of novel forms is not a principled problem for analogical formations, even if the amount of available information on which to base analogies is smaller than for existing words.

One other interesting result across corpora is that the left constituent family consistently has a greater effect size than the right constituent family. This may be surprising for two reasons. First, the textbook examples of analogy exclusively illustrate the effect of the right constituent (e.g. *street*, *avenue*, *pie*). Second, the second constituent is normally considered the more important constituent for other compound properties, e.g. semantics and grammar, as reflected in the right-hand head rule. However, from a psycholinguistic point of view, the left constituent is very prominent, and in some sense more important for lexical processing than the right constituent, especially for word recognition.

[19]Recall that for all our above models we used data sets that had constituent families for both constituents.

[20]One reason why this is highly unlikely is the fact that noun-noun compounding is generally held to be the most productive word-formation process in English, which means that the chances for a given noun to be used in such a construct are high.

Furthermore, other studies (e.g. by Krott and colleagues) have also found that it is the left constituent family that has a decisive (and more important) influence on compound behavior. Krott et al. (2001, 2002b, 2007) studied the morphological properties of compounds in German and Dutch and demonstrated that the constituent family, and the left constituent family in particular, has a significant influence on the choice of the linking element. For German Krott et al. (2007) it was even shown that the right constituent does not contribute at all to the decision which linking element will be used. Krott et al. (2002a) investigate semantic effects and find that there is a relation between the semantic class of the left and right constituent (in terms of animacy and concreteness) and the choice of the linking morpheme in Dutch, but the semantic effects are generally stronger for the left constituent, and are sometimes even totally absent for the right constituent. Thus we can say that the greater impact of the left constituent in analogical decisions is not only very plausible from a psycholinguistic point of view but is also independently, and cross-linguistically, confirmed for other compound phenomena.

That constituent family is not the only factor involved in analogical computation should not surprise us either. In the studies by Krott and her colleagues it was also the case that apart from the strong constituent family effects, other similarities played an important role, namely semantics (see the above paragraph), phonological and morphological structure. In fact, in an analogical framework, one would expect a multitude of factors having an effect since, in principle, all kinds of factor may be chosen for the computation of similarity between linguistic entities, and compounds in particular (see again Libben 2006). Why the language user picks certain properties but not others for building analogies is a more general problem of analogical approaches, but interestingly enough, this kind of problem also extends to rule-based frameworks. Rules also make crucial reference to certain properties, and not to others. An illumination of this problem is therefore an important goal of future research, irrespective of the theoretical framework being employed in the analysis.

References

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

Baayen, Harald, Richard Piepenbrock & Leon Gulikers. 1995. *The CELEX lexical database (CD-ROM)*. University of Pennsylvania, Philadelphia, PA: Linguistic Data Consortium.

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.

Daelemans, Walter, Jakub Zavrel, Ko van der Sloot & Antal van den Bosch. 2007. TiMBL: Tilburg Memory Based Learner, version 6.0, reference guide. ilk technical report 07-03. Technical report, Computational Linguistics, Tilburg University.

Demšar, J. 2006. Statistical Comparisons of Classifiers over Multiple Data Sets. *The Journal of Machine Learning Research* 7:1–30.

Fawcett, Tom. 2003. ROC graphs: Notes and practical considerations for data mining representation. Tech report. HPL-2003-4, HP Laboratories, Palo Alto.

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.

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.

Krott, Andrea, Harald Baayen & Robert Schreuder. 2001. Analogy in morphology: modeling the choice of linking morphemes in Dutch. *Linguistics* 39(1):51–93.

Krott, Andrea, Loes Krebbers, Robert Schreuder & Harald Baayen. 2002a. Semantic influence on linkers in Dutch noun-noun compounds. *Folia Linguistica* 36:7–22.

Krott, Andrea, Robert Schreuder & Harald Baayen. 2002b. Linking elements in dutch noun-noun compounds: constituent families as predictors for response latencies. *Brain and Language* 81:708–722.

Krott, Andrea, Rob Schreuder, Harald R. Baayen & Wolfgang U. Dressler. 2007. Analogical effects on linking elements in German compound words. *Language and Cognitive Processes* 22(1):25–57.

Kunter, Gero. 2009. *Compound stress in English: The phonetics and phonology of prosodic prominence*. 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.

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.

Libben, Gary. 2006. Why study compound processing? In Gary Libben & Gonia Jarema (eds.) *The representation and processing of compound words*, 1–22. Oxford: Oxford University Press.

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.

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.

Ostendorf, Mari, Patti Price & Stefanie Shattuck-Hufnagel. 1996. *Boston University Radio Speech Corpus*. Philadelphia: Linguistic Data Consortium.

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, 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.

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.

Skousen, Royal, Deryle Lonsdale & Dilworth B. Parkinson (eds.) . 2002. *Analogical modeling: An exemplar-based approach to language*. Amsterdam: Benjamins.

Sokolova, M., N. Japkowicz & S. Szpakowicz. 2006. Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation. *Lecture notes in computer science* 4304:1015.

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.

Appendix

Table 26: Analysis of variance of logistic regression model with left and right constituent bias as predictors, T&W data

	Chi-Square	d.f.	P
leftConstituentBias	134.38	2.00	0.00
rightConstituentBias	45.64	2.00	0.00
TOTAL	141.22	4.00	0.00

Table 27: Analysis of variance of logistic regression model with left and right constituent bias as predictors, CELEX data

	Chi-Square	d.f.	P
leftConstituentBias	122.38	2.00	0.00
rightConstituentBias	145.36	2.00	0.00
TOTAL	233.01	4.00	0.00

Table 28: Analysis of variance of logistic regression model with left and right constituent bias as predictors, Boston Corpus type data

	Chi-Square	d.f.	P
leftConstituentBias	71.95	2.00	0.00
rightConstituentBias	40.20	2.00	0.00
TOTAL	101.34	4.00	0.00

Table 29: Analysis of variance of logistic regression model with left and right constituent bias as predictors, Boston Corpus, token data

	Chi-Square	d.f.	P
leftConstituentBias	105.03	2.00	0.00
rightConstituentBias	106.49	2.00	0.00
TOTAL	229.85	4.00	0.00

Table 30: Analysis of variance of final logistic regression model, based on all predictors but family bias, CELEX data

	Chi-Square	d.f.	P
spellNew	90.00	1.00	0.00
semRel4	23.53	1.00	0.00
semRel7	35.62	1.00	0.00
semRel12	41.57	1.00	0.00
semRel16	11.48	1.00	0.00
TOTAL	178.59	5.00	0.00

Table 31: Analysis of variance, final logistic regression model based on all predictors but family bias, Boston Corpus data, types

	Chi-Square	d.f.	P
isN1PN	3.93	1.00	0.05
morphRight	11.43	4.00	0.02
semRel4	20.21	1.00	0.00
TOTAL	35.50	6.00	0.00

Table 32: Analysis of variance of final logistic regression model based on all predictors but family bias, Boston Corpus token data

	Chi-Square	d.f.	P
isPropN	10.49	1.00	0.00
spell	22.21	2.00	0.00
semRel2	5.19	1.00	0.02
semRel4	50.87	1.00	0.00
semRel6	5.21	1.00	0.02
semRel12	4.11	1.00	0.04
semRel14	13.10	1.00	0.00
TOTAL	104.23	8.00	0.00

Table 33: Analysis of variance of logistic regression model using all predictors, CELEX data

	Chi-Square	d.f.	P
spell	97.27	2.00	0.00
semRel4	10.88	1.00	0.00
semRel7	23.88	1.00	0.00
semRel12	28.40	1.00	0.00
semRel16	15.01	1.00	0.00
leftConstituentBias	53.31	2.00	0.00
rightConstituentBias	77.25	2.00	0.00
TOTAL	272.71	10.00	0.00

Table 34: Analysis of variance of logistic regression model using all predictors, Boston Corpus, type data

	Chi-Square	d.f.	P
semRel4	6.67	1.00	0.01
semRel13	4.48	1.00	0.03
leftConstituentBias	63.83	2.00	0.00
rightConstituentBias	38.23	2.00	0.00
TOTAL	104.89	6.00	0.00

Table 35: Analysis of variance of logistic regression model with only semantic and structural predictors, Boston Corpus, token data

	Chi-Square	d.f.	P
leftConstituentBias	72.19	2.00	0.00
rightConstituentBias	76.63	2.00	0.00
isPropN	4.74	1.00	0.03
spell	12.41	1.00	0.00
semRel2	4.41	1.00	0.04
semRel4	20.04	1.00	0.00
semRel6	4.81	1.00	0.03
TOTAL	222.13	9.00	0.00