Informativity and analogy in English compound stress

Melanie J. Bell & Ingo Plag

Revised version

March 8, 2013

Abstract

It has long been claimed (e.g. Sweet 1892, Bolinger 1972, Ladd 1984) that informativity has an influence on the leftward or rightward stress assigned to noun-noun combinations in English. However, the few available empirical studies of this hypothesis have provided contradictory findings (Plag & Kunter 2010, Bell 2013, Bell & Plag 2012). The present paper replicates the effect of informativity using the same measures as Bell & Plag (2012) but with a different set of data. More informative constituents in N2 position tend to be stressed. This result fits with the general propensity of speakers to accentuate important information (e.g. Bolinger 1972). The results also raise the question of the relationship between informativity and constituent identity, which is the strongest known predictor of compound stress pattern (e.g. Plag 2010, Arndt-Lappe 2011). An exploration of this problem shows that the two factors are interrelated. We argue that informativity is best conceptualized as underlying other predictors of prominence, including constituent identity.

1 Introduction[1]

It is well known that, in Present-day English, some noun-noun (NN) constructions are produced with main stress on the first element (N1), e.g. *táble lamp*, while others have stress on the second element (N2), e.g. *silk shírt*. In this paper, we use the term COMPOUND for all such constructions, in which two nouns combine to name a single entity or class of entities. In other words, following e.g. Bauer (1998), Olsen (2000) and Bell (2011), we do not distinguish between compound nouns and constructions which in some other analyses (e.g. Payne & Huddleston 2002) would be regarded as noun phrases with nominal modifiers. For describing the prosody of these constructions, we use the terms STRESS and PROMINENCE interchangeably.

In running speech, about one third of NN compounds (as defined above) are stressed on N2, while two thirds are stressed on N1 (e.g. Sproat 1994, Plag *et al.* 2008, Bell & Plag 2012). An adequate account of compound stress in English should therefore be able to predict which of these patterns will apply in any given case. Many scholars of English have addressed this problem. For example, Chomsky & Halle (1968) suggest that, in cases of left prominence, main stress is assigned to N1 by a COMPOUND RULE. However, they make no attempt to define the strings to which this rule applies, and simply state that an investigation is needed 'of the conditions, syntactic and other, under which the Compound Rule is applicable' (Chomsky & Halle 1968: 156).

Recent empirical studies by Plag and colleagues (e.g. Plag 2006, Plag *et al.* 2007,

[1] The authors wish to thank Gero Kunter for his help with COCA data. We are also grateful for the very helpful comments we received from the two reviewers and the editor Heinz Giegerich. This work was made possible by an AHRC postgraduate award (114 200) and a major studentship from Newnham College, Cambridge, to the first author as well as two grants from the Deutsche Forschungsgemeinschaft (PL151/5-1, PL 151/5-3) to the second author, all of which are gratefully acknowledged.

2008, Plag 2010, Arndt-Lappe 2011) constitute such an investigation. These studies have shown that a variety of different factors can be used to predict the stress pattern of a compound, including lexicalization, semantics, and the identities of the constituents (N1 and N2): more lexicalized compounds tend to be more prone to left stress, compounds exhibiting the same semantic relation between constituents tend to have the same kind of stress pattern, and compounds that share the same N1 or N2 also tend to be stressed in the same way. The latter effect has been interpreted as analogical. For ease of reference and to be specific about the kind of analogy involved, we will label this effect the CONSTITUENT IDENTITY effect.

In addition to these well-established effects, there are two further factors that have been claimed to influence compound stress, but which have been less extensively tested. These are the length of the constituents and their relative INFORMATIVENESS (also called INFORMATIVITY or INFORMATION CONTENT). This paper investigates both of these factors, with the focus on informativeness. The claim is that less informative constituents tend to be unstressed (e.g. Sweet (1892), Bolinger (1972), Ladd (1984)). Until now, however, large-scale empirical evidence for this idea has been scarce. Bell (2013) and Bell & Plag (2012) find strong informativity effects in a large database of compounds from the British National Corpus (BNC, Davies 2004-), but these findings are in need of replication. Plag & Kunter (2010) found no independent effect of informativity on compound prominence in a large database of compounds taken from the Boston University Radio Speech Corpus (Ostendorf *et al.* 1996), which we will refer to as the BU CORPUS. However, that study had two crucial limitations. Firstly, only a single measure of informativity was used and, secondly, this measure was problematic, since it was calculated on the basis of a very small corpus.

In the present study we will use the BU corpus data again, addressing the crucial

methodological problems of Plag & Kunter's (2010) study: firstly, we will include a greater number and variety of informativity measures, and secondly, these informativity measures will be derived from much larger databases, namely the Corpus of Contemporary American English (Davies 2008-) and the WordNet lexical database (Fellbaum 1998). The informativity measures are the same as used by Bell & Plag (2012), so that a direct comparison of the results is possible. Furthermore, in contrast to the data used by Bell & Plag (2012), many of the compounds in the BU corpus data share each of their constituents with at least one other compound in the dataset. This means that it is possible to calculate the tendency of each constituent to be associated with a particular stress pattern, and hence to include the constituent identity effect in the analysis. This paper therefore has two objectives: firstly, to replicate the effect of informativity on compound stress in a different set of data, and secondly, to investigate the relationship between this effect and the strongest known predictor of compound stress, namely constituent identity. For the statistical analysis we will use multiple mixed effects regression modeling.

The results show that measures of informativity are indeed highly predictive of prominence placement. The finding that more informative constituents tend to be stressed fits with the general propensity of speakers to accentuate important information (e.g. Bolinger 1972). An exploration of the relationship between informativity effects and constituent identity effects shows that constituent identity subsumes most other known effects on compound stress, including informativity: when constituent identity is included as a predictor in our models, other predictors become less influential. However, it is also shown that a constituent's informativity strongly predicts its bias for a particular stress pattern (i.e. the constituent identity effect), and it is argued that informativity is in fact the most important determinant of

compound stress, underlying most other known predictors.

The paper has the following structure: section 2 gives an overview of the factors known to determine the variation in NN prominence, section 3 describes the methodology used in the present study, section 4 describes the results of our analyses and section 5 summarizes the findings and discusses their implications for theories of compound stress.

2 Compound stress: What we know and what we don't know

In recent years the problem of compound stress variation has been addressed in a number of empirical studies (Plag 2006, Plag *et al.* 2007, 2008, Plag 2010, Plag & Kunter 2010, Bell 2013, Bell & Plag 2012, Kunter 2011, Arndt-Lappe 2011). Using different methodological tools and different data sets from different varieties of English, these studies have tested various factors that have been claimed in the literature to be influential in stress assignment. Since most of these factors will also be included in the analyses to follow, we will take a short look at them in this section.

Before doing so, a note on the nature of the phenomenon we are looking at and the pertinent terminology seems appropriate. Most authors speak of (COMPOUND) STRESS, while others speak of (PROSODIC) PROMINENCE. Phonetically, compound prominence manifests itself in most cases through pitch accents (see Kunter & Plag 2007, Kunter 2011 for detailed acoustic analyses[2]). Over and above the lexical stress(es) of each constituent in any compound, left-stressed compounds usually have one pitch accent, on N1, whereas right-stressed compounds usually have two pitch accents, one on each constituent. Given that N1 always receives an accent, the difference between left-stressed and right-stressed compounds can therefore be straightforwardly

[2] See also Plag *et al.* (2011) for a parallel analysis of primary and secondary stresses in derived words.

conceptualized as one of N2 accentuation. Left-stressed compounds have no accent on N2, whereas right-stressed compounds do have an accent on N2. This conceptualization is helpful for understanding the relevance of informativity: when N2 is informative relative to N1, it receives an accent. However, not much hinges on the terminology itself and we will use the terms PROMINENCE and STRESS more or less interchangeably.

In the pre-2006 literature on compound stress assignment, possible explanatory factors are usually formulated in a deterministic fashion: compounds of a given type are claimed to exhibit either left or right stress categorically. However, practically all of the empirical studies carried out since 2006 have shown that such deterministic, rule-based approaches are inadequate. In contrast, probabilistic and analogical models have been shown to be quite successful in predicting the prominence type of a given compound. These empirical studies have consistently shown that a compound's semantics and degree of lexicalization, as well as the identities of its constituents, are all predictive of its stress pattern. We will therefore discuss these three factors in more detail, as well as the two additional factors investigated in this study, namely informativity and length.

2.1 Semantics

There are many claims in the literature that right prominence in compounds goes together with certain semantic properties. These may be properties of the individual constituents, or of the relation between the two constituents (see, for example, Plag *et al.* 2008 for a review of the literature). Large-scale empirical studies have indeed found probabilistic effects of certain semantic categories. Plag *et al.* (2007, 2008), for example, found the effects shown in table 1.

Insert table 1 here

2.2 Constituent identity

The idea that the identity of the constituent nouns plays a role in compound stress assignment has been around for some time, too (e.g. Schmerling 1971). An illustration of this is the behavior of compounds that have *street* as their N2, compared with compounds that have *avenue* as their N2. When they refer to the names of thoroughfares, the former are categorically left-stressed (e.g. *Óxford Street*, *Báuer Street*, *Wáll Street*), the latter right-stressed (e.g. *Madison Ávenue*, *Sproat Ávenue*, *Victory Ávenue*). This can be interpreted as an analogical effect based on the positional constituent family. The positional family is defined as the set of compounds that share the first, or the second, constituent with a given compound. For example, the left constituent positional family of *country house* would include compounds such as *country club*, *country music*, *countryside*, while the right constituent positional family would feature compounds like *town house*, *jailhouse*, *summer house*. For each positional family, one can compute the tendency towards a particular kind of stress, and it has been shown that this so-called CONSTITUENT FAMILY BIAS, especially the left constituent bias, is highly predictive for the stress of new compounds in that family, i.e. compounds with a particular constituent in a given position (Plag 2010, Arndt-Lappe 2011).

2.3 Lexicalization

Lexicalization has long been claimed (e.g. by Sweet 1892: 289) to be a contributory factor in compound stress assignment, with more lexicalized compounds being more prone to left stress. Empirical studies have used spelling, frequency and listedness in

dictionaries as measures of lexicalization, with considerable success. The assumptions are that more lexicalized compounds are more prone to hyphenated and one-word spellings, have a higher frequency and are more likely to be listed in dictionaries (e.g. due to their semantic opacity), than less lexicalized compounds. It is now well established that degree of lexicalization, as indicated by one or more of these measures, is an important determinant of a compound's stress pattern.

2.4 Informativity

Sweet (1892: 288) was probably also the first author to put forward the idea that compound stress assignment might depend on the amount of information carried by a given constituent vis-à-vis the other. Sweet (*ibid.*) uses the term LOGICAL PROMINENCE for this effect and writes that stress on N1 'seems to be the result of the second element being less logically prominent than the first, through being a word of general meaning and frequent occurrence in compounds'. This reasoning is based on the assumption that, in general in language, uninformative elements tend to be unaccented, while more informative and unexpected information is accented. A further assumption is that the information content of a word depends both on its semantics and its frequency, with more semantically general and more frequent words being less informative. These assumptions have received empirical support from studies by Pan & McKeown (1999) and Pan & Hirschberg (2000), which demonstrate that pitch accent placement in texts can quite successfully be predicted on the basis of semantic and frequency-based measures of informativity.

The relationship between semantic specificity, informativeness and compound stress has been discussed by various authors, using a variety of terms. For example, Jones (1922: 126) suggests that when N2 is felt to have 'special importance' it attracts

stress and is therefore associated with right prominence. A similar idea is expressed by Bolinger (1972), who explicitly links the ideas of accent, semantics and informativity. In his view, '(a)ccented words are points of information focus', and this in turn is a matter of semantics. But rather than highly informative constituents attracting stress, Bolinger (*ibid.*) suggests that more semantically predictable elements are DEACCENTED. He gives the example *It's a geránium plant*, in which 'we understand that if it is a geranium, it is a plant': *plant* therefore adds very little to the meaning of the sentence, and is deaccented. Ladd (1984) also argues that left prominence in compounds is the result of deaccenting the second constituent, and suggests that this is most likely when N2 is least semantically specific. For example, he argues that the reason why thoroughfare names ending in street are stressed on N1, while those ending in *road*, *avenue*, *place* etc. are stressed on N2, is that *street* is the least specific and hence least informative of the group, and is therefore deaccented.

The idea that informativity can be gauged in terms of frequency comes from information theory, where a standard measure of INFORMATION CONTENT is the negative log likelihood of a word in a corpus (Shannon, 1948). The less frequent a word, the less likely it is to occur, hence the less expected and more informative it is taken to be when it does occur. This idea is foreshadowed by Sweet's reference (*ibid.*) to the 'frequent occurrence' of N2 in compounds being associated with stress on N1. In the same vein, Marchand (1969: 23) states that '[t]he frequent occurrence of a word as second constituent is apt to give compound character [i.e. left prominence] to combinations with such words'.

Despite the long history of these ideas, there were until recently no empirical studies that tested whether one can predict compound stress assignment on the basis of informativity: Bell (2013) and Bell & Plag (2012) were the first to show that

measures of informativity are indeed predictive of prominence placement, at least in one set of data. Both these studies used the same set of compounds, whose stress patterns were elicited in a controlled experimental procedure. The compounds were sampled from the demographic section of the BNC, which consists of conversational British English. Given that we assume N1 will receive an accent in all compounds, be they left- or right-stressed, what is of interest is whether or not N2 receives an accent. If this is related to informativity, then the pertinent variables are the inherent informativity of N2, and its informativity relevant to N1. Bell (2013) and Bell & Plag (2012) used a variety of measures for these two types of informativity.

As a semantic measure of informativity, the studies mentioned in the previous paragraph used SYNSET COUNTS. The term SYNSET comes from the WordNet lexical database (Fellbaum 1998), where a synset is a set of words with similar meanings. The synset count of a word is the number of synsets to which it belongs, each of which represents a different sense of the word in question. For example, the noun *house* has one synset, consisting of *house*, *firm* and *business firm*, that embodies the meaning 'business organization', and another synset, consisting of *house*, *family*, *household*, *home* and *menage*, that means 'social unit living together'. Overall, the noun *house* belongs to 12 such synsets, i.e. has 12 different senses, in WordNet. Our assumption is that the greater the number of senses a word has, i.e. the higher its synset count, the less semantically specific it is, and hence the less informative. The noun *house*, with its twelve synsets, would be a semantically rather nonspecific constituent, and hence less informative than a semantically much more specific noun like *desk*, which belongs to only one synset (which, in fact, has only itself as its member). Bell (2013) and Bell & Plag (2012) show that the greater the synset count of N2, the less likely it is to receive an accent. On the other hand, the greater the synset count of N1, the more likely is N2

to receive an accent, as it becomes more informative relative to the nonspecific first constituent.

For frequency-based measures of informativity, the above-mentioned studies used the constituent family sizes of N1 and N2. The positional family size of N2 is the number of different compounds in which that particular word appears in the right-hand position: it can therefore be used to estimate the probability of the word occurring as the second element of a compound. The greater the positional family size of N2, the greater the number of compound types in which it occurs as the right-hand constituent, hence the more expected it is following a noun, and the less informative it is in that position. The conditional probability of N2 is the probability of N2 occurring after a given N1, and can be estimated using the family size of N1. The greater the positional family size of N1, the greater the number of nouns that might potentially follow it, hence the less expected and more informative is N2 relative to N1. Mathematically, this family-size-based conditional probability of N2 can be expressed as 1 divided by the family size of N1. Bell (2013) and Bell & Plag (2012) show that, as predicted, the greater the family size of N2, the less likely it is to be accented, while the greater the family size of N1, i.e. the lower the conditional probability of N2, the more likely is N2 to be accented.

In contrast to the studies mentioned in the preceding paragraphs, Plag & Kunter (2010) did not find straightforward informativity effects in their data. Rather, their measure of informativity only acted as a rather weak modulator for the much more significant effect of constituent identity. Given these contradictory results, a study is called for that tries to replicate the effect of informativity, while at the same time taking constituent identity into account. The present study will do exactly this.

2.5 Length

Another factor which has been hypothesized to influence compound stress assignment is the length of the constituents. Jespersen (1909: 485), for example, remarks that longer right constituents have a greater tendency to be stressed. This factor has been empirically verified by Bell (2013) and Bell & Plag (2012), the only studies to date that have systematically investigated it. Although not the main focus of our interest, the present study will include length as a covariate, using the same measure as Bell & Plag (2012).

3 Methodology

3.1 Data

In this study, we will use a set of compounds from the BU corpus. In contrast to the British English data of Bell (2013) and Bell & Plag (2012), the BU corpus represents American English. It contains over seven hours of professionally read radio news scripts and includes speech from seven FM radio news speakers associated with the public radio station WBUR, four male and three female. The BU corpus is well suited for testing hypotheses about compound stress assignment. It contains a large number of compounds, it provides high-quality recordings, and the speakers, being trained news announcers, produce relatively standard, error-free speech. Furthermore, because many compounds in the corpus are repeated, both by the same speaker and by different speakers, the data exemplifies the within-speaker and cross-speaker variation that is characteristic of English compound stress (Ladd 1984: 256, Bauer 1983: 103, Plag 2006, Kunter 2011: 174-201, Bell & Plag 2012).

Compounds from the BU corpus have been investigated in a number of previous

studies of compound stress, including Plag *et al.* (2008), Plag (2010), Plag & Kunter (2010), Kunter (2011) and Arndt-Lappe (2011). The reader is referred to these works for full discussion of the corpus data, including the role of discourse factors (Plag *et al.* 2008) and the determination of prominence based on the acoustic signal (Kunter 2011). In the present study, we make use of data from Plag *et al.* (2008) and Plag (2010). For convenience, we summarize below how this data was collected and coded.

From the BU corpus, Plag *et al.* (2008) 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' (*ibid.*: 767). From this set, personal names and those sequences with an appositive modifier, such as *Governor Dukakis*, were eliminated. The resulting noun-noun compounds were then coded for a number of semantic, structural and phonological features. For a subset of these compounds, Plag (2010) elicited human judgements of stress pattern: two trained listeners classified each token on the basis of their auditory impression, and tokens were excluded from subsequent analysis if the two judgements differed. In 77.3 percent of cases, however, the two judgements were the same, producing a dataset of 1154 tokens with agreed prominence. This set of 1154 tokens is the one we will use here. We will use the existing codings from the previous studies mentioned above, complemented by measures of length and informativity. Length and informativity will be measured using the same variables as those employed by Bell & Plag (2012). Due to the fact that so many compounds occur repeatedly, the data can be analyzed either by compound type or by individual token. In this paper we present token-based analyses, since these most accurately reflect the within-type variability found in the data.

3.2 Categories coded

With regard to semantics, Plag *et al.* (2008) coded each token for five properties of a constituent or the compound as a whole, e.g. N1 IS A PROPER NOUN, as well as eighteen relations between constituents, e.g. N1 HAS N2. Each of these semantic categories was coded as a binary factor with the levels yes and no, so that a single token could be coded for more than one possible interpretation. Two trained linguists independently coded each token, taking into account its context, and tokens were excluded from further analysis if these two codings differed. Because we are using only a subset of the Plag *et al.* (2008) data, namely those tokens for which agreed human stress ratings are available, an additional restriction to items with agreed semantic codings would reduce the dataset to a mere 506 tokens. Furthermore, within these 506 tokens, only 8 of the 26 semantic classes have more than ten representatives. This greatly reduces the power of any statistical analysis involving these variables, whose effects are in any case well established, and we therefore decided not to include them in the present study. This allows us to use a larger set of data, focusing on the effects of informativity, length and constituent identity. Table 2 summarizes the predictors we used, and we will discuss each group in turn.

Insert table 2 here

3.2.1 Constituent identity

Plag (2010) computed the tendency of a particular noun in either N1 or N2 position to be associated with a particular stress pattern. Using the same set of 1154 tokens as we are using, and the same stress ratings, the proportion of left stresses within each positional family was calculated for each token in the dataset. To give an example, consider the compound *advertising business*, which is represented in our dataset by a

single token. There are six more tokens in the data that have the same left constituent, one each of *advertising agency*, *advertising battle*, *advertising commentator*, *advertising costs*, *advertising days* and *advertising dollars*. Of these seven tokens, only the token of *advertising battle* is rightward-stressed. For the token of *advertising business*, the left constituent bias for leftward stress is therefore calculated as 5/6, i.e. 0.833, because five of the other six tokens have leftward stress. Similarly, there are two more tokens in the data that have business in N2 position, one each of *biotechnology business* and *computer business*. Of the three tokens that share this constituent, only *computer business* is coded for rightward stress. For the token of *advertising business*, the right constituent bias for leftward stress is therefore calculated as 1/2, i.e. 0.5, because only one of the other two tokens has leftward stress. Note that by using this procedure, the stress of the token in question is not taken into account when computing the family biases for that token. This is done in order to avoid the problem of predicting the stress of an item on the basis of stress information gleaned from that very item. Plag (2010) transformed these proportional biases into a categorical variable with the values left bias, right bias and neutral, however we will follow Plag & Kunter (2010) in using the untransformed proportions, in order not to lose information about the degree of variability in the data. Plag & Kunter (*ibid.*) standardized these proportions to reduce the danger of collinearity adversely affecting their models. We also did this, but it turned out to make no difference to any of the analyses, and so the models we describe here use the raw biases.

3.2.2 Lexicalization

As a measure of lexicalization, Plag *et al.* (2008) coded the data for compound frequency. Because many of the compounds in the BU corpus are low frequency items

relating to particular news stories, they do not occur in standard corpora, and so it was necessary to obtain frequencies from the internet. This was done using Query Google (Hayes & Ma 2003-) to extract word counts automatically using the Google search engine, with the searches restricted to English language webpages. The Google frequencies provided by Plag *et al.* (*ibid.*) are lemmatized, i.e. summed over all inflectional forms, and include all spelling variants (one word, hyphenated and two words). Plag *et al.* (2007) demonstrate that frequencies obtained in this way are reliably correlated with frequencies obtained from the Cobuild corpus, a balanced corpus of 18 million words.

3.2.3 Informativity

As described in Section 2.4, Bell & Plag (2012) used two types of measure to estimate the informativity and relative informativity of N2: semantic specificity, based on synset counts, and probability of occurrence, based on positional family sizes. To ensure direct comparability with that study, we will use the same measures here. Synset counts were manually extracted from the WordNet index file for nouns, for all values of N1 and N2 in the data. In the few cases where a constituent did not appear in WordNet, any compound with that constituent was removed from the data: this reduced the dataset to 1075 tokens. For the estimation of family sizes we used COCA, The Corpus of Contemporary American English (Davies 2008-), accessed in the fall of 2009. We searched for spaced noun-noun collocates in which the relevant values of N1 or N2 occurred in first or second position respectively. Such an automated procedure generates many items that are not actually compounds, but time constraints made it impossible to check every hit. However, as shown in Bell (2013) and Bell & Plag (2012), using a manually corrected sample, the numbers of hits

returned by such automated searches are highly correlated with actual family sizes obtained by manual checking. The number of types returned by a search was therefore taken to represent the positional family size for that constituent. There were a few constituents that did not have families in COCA, and so compounds with these constituents were excluded from subsequent analysis, further reducing the dataset to 1056 tokens.

3.2.4 Length

Plag *et al.* (2008) coded the syllable structure of every noun-noun compound token from the BU corpus: each syllable was classed as either W or S, where S represented the main stressed syllable of a constituent (secondary stress was not coded). For one token of *deputy superintendent*, for example, N1 (*deputy*) was coded as SWW and N2 (*superintendent*) as WWWSW. For the present study, we used this information to calculate a number of measurements of constituent or compound length. These were: the number of syllables in N1, the number of syllables in N2, the total number of syllables and the number of syllables after the main stressed syllable of N1. Initial exploratory analyses indicated that all of these measures were predictive of prominence placement, but that the latter was the most successful predictor. Bell & Plag (2012) also found that, for their BNC data, the length measure that produced the best and most interpretable models was the number of syllables after N1 main stress. For left-stressed compounds, this variable can be conceptualized as a measure of the length of the unaccented 'tail' following the main stress (see Bell & Plag 2012: 510). To illustrate the coding, consider the token of *deputy superintendent* mentioned above. For this token, the number of syllables after N1 main stress is seven. However, because each token was coded separately, the number is not necessarily constant for all tokens

of the same compound. In the case of *magazine subscription*, for example, some tokens are produced with the main stress of *magazine* on the first syllable, while others have it on the final syllable. The number of syllables after N1 main stress for *magazine subscription* is therefore three for some tokens and five for others.

3.3 Statistical analysis

For the statistical analysis we carried out mixed effects regression modeling (e.g. Baayen *et al.* 2008) using the lme4 package (Bates *et al.* 2011) in the R statistical programming environment (R Development Core Team 2011). In regression analysis, the outcome of a dependent variable, in this case stress pattern, is predicted on the basis of independent predictor variables, in this case informativity, length and so on. Multiple regression has the advantage of showing the effect of one predictor while holding all others constant, which is especially welcome for an investigation like this one, where many different variables seem to play a role at the same time. The further advantage of mixed effects modeling is that both fixed and random effects can be included as predictors. A fixed effect is one that can be generalized from a sample to a population: for example, if it is shown that the more frequent a compound in our dataset, the more likely it is to have left stress, then we can predict the stress pattern of new compounds on the basis of their frequency. Random effects, on the other hand, are not generalizable: for example, certain speakers may have a tendency to prefer one stress pattern over another, but knowing the idiosyncratic preferences of our particular subjects would not enable us to predict the preferences of a new group. In our analyses we therefore include SPEAKER as a random effect, with the variables described in Section 3.2 as fixed-effect predictors. Including speaker as a random effect instructs the algorithm to take into account the relative tendencies of the

different speakers towards one stress pattern or the other. As a precaution against the potentially harmful effects of extreme values on our statistical models, compound frequencies and all measures of informativeness were first logarithmatized. We then applied the usual procedure of step-wise elimination of non-significant variables to arrive at our final models.

4 Results

In this section we will present the results of various analyses, in which different constellations of variables are used as predictors. This is done in order to explore the individual contributions of the different predictors and the relationships between them. The variables we are most interested in are informativity and constituent identity, and these are therefore the main focus of our discussion. In addition, all the analyses include length and lexicalization measures as control variables. Although we do not discuss these in great detail, we do find effects of length and of lexicalization in all our models, and in the predicted directions. We will first present the results of an analysis that has, apart from these control variables, only informativity measures as predictors, in order to attempt to replicate the results of Bell & Plag (2012).

4.1 Informativity as a predictor of compound stress

Our initial model included the following informativity measures: family size of N2, conditional probability of N2, and synset counts for N1 and N2. We find main effects for the two probability-based measures but no significant effect for the synsets, and no interactions. The final model, after elimination of the insignificant variables, is given in table 3. Positive coefficients indicate a tendency towards rightward stress, negative ones towards left stress. *C* is a measure of the discriminative power of a logistic

regression model, i.e. of the degree of agreement between observed outcomes and the probabilities computed by the model. The possible values of *C* range from 0.5 to 1.0, where 1.0 indicates that the model always assigns higher probability to the outcome actually observed than to the alternative outcome. Standardly, values of 0.9 or more indicate an excellent fit between the model and the data, and values between 0.8 and 0.9 indicate a good fit (see, for example, Kutner *et al.* 2005 for details). Our figure of 0.821 therefore indicates that the model is quite successful in its predictions. In fact, this *C*-value is remarkably similar to that of the best model reported in Plag (2010), *C* = 0.828, even though the latter was based on a different set of predictors, namely constituent identity, semantics and spelling.

Insert table 3 here

The effects are plotted in Figure 1. In the top row we find the effects of our control variables, as predicted. The longer the compound, the higher the probability of it receiving an accent on N2, and the more frequent the compound, the more prone to left stress it will be. The bottom two plots in Figure 1 show the informativity effects. The bottom left plot and the bottom right plot show that the more expected N2 is, either conditioned by N1 or by itself, respectively, the less likely it is that N2 receives stress. This is in accordance with the hypothesis that prominence is at least partly determined by informativity.

Place figure 1 about here

Notably, in the presence of the frequency-based informativity measures, semantic specificity as gauged by the synset counts is not significant. In order to investigate the role of semantic specificity and the relationship between the two kinds of informativity variables a bit further, we fitted a model from which the family size measures were excluded. After model simplification, the two control variables from above and both measures of semantic specificity were highly significant in the predicted directions, with no interaction (for N1: p=0.000151, z=3.789, for N2: p=0.005188, z=-2.795). The two semantic specificity effects in this model are plotted in Figure 2. The left-hand panel shows that the larger the N1 synset count, i.e. the less specific N1, the higher the probability of an accent on N2. This is consistent with the hypothesis that as N1 becomes less informative, N2 becomes more informative relative to N1, and is therefore more likely to be accented. The right-hand panel shows that the larger the synset count of N2, i.e. the less specific N2, the less likely it is to receive an accent. Again, this is consistent with the hypothesis that prominence is determined by the informativeness of N2. This model is not quite as successful as the one based on family sizes, but is still quite good in its predictions (C=0.792). The fact that the synset counts do not emerge as significant predictors in the presence of the probability-based variables indicates that the synset effects are preempted by the family sizes. This in turn suggests that the two types of measure account for the same portion of variation in the dependent variable, and may indeed represent the same underlying phenomenon, namely informativity.

Place figure 2 about here

To summarize, these analyses have provided robust evidence for an effect of informativity on compound stress assignment. This nicely replicates Bell & Plag's (2012) findings with a different dataset. Let us now turn to the question of whether informativity effects survive when constituent identity is also included as a predictor.

4.2 Constituent identity and informativity as predictors

To test whether constituent identity and informativity are predictive of stress pattern in the presence of each other, we fitted a model with both types of predictor in addition to the same control variables as above. The final model is documented in table 4. There are significant main effects for length, compound frequency, both constituent biases and one measure of informativity, namely the conditional probability of N2. Figure 3 shows these significant effects. As before, we see that longer compounds are more likely to have right prominence while more frequent compounds are more prone to left prominence. The effect of the conditional probability of N2 is also unchanged: the less likely N2 given N1, the more likely it is to receive stress. Both constituent bias effects also work in the expected direction: as the bias of either constituent for left stress increases, the probability of an accent on N2 falls. With the inclusion of the constituent biases, the predictive power of the model is improved (*C*=0.855).

Insert table 4 here

Place figure 3 about here

Taken at face value, these results suggest that a constituent's informativity and its bias for a particular stress pattern are to some extent independent of one another. However, a moment's thought reveals that this would be surprising. Let us first

consider the nature of the constituent identity effect. In our models, it is based on the constituent family bias, which we compute using orthographic representations of the compounds in our dataset. In other words, it is the tendency of a particular orthographic string in either N1 or N2 position to be associated with a particular stress pattern. Most obviously, orthographic strings represent phonological strings, but there is ample psycholinguistic evidence that in the mental lexicon each string is also associated with many other properties including, for example, semantics, length, frequency and positional family sizes (e.g. Schreuder & Baayen 1997, Bertram *et al.* 2000, Moscoso del Prado Martín *et al.* 2004). In other words, constituent identity can be seen as a proxy for a bundle of variables, some of which may have an influence on stress assignment. For example, as shown by Plag *et al.* (2008), geographical terms in N2 position predispose a compound to right stress, and geographical terms will therefore tend to have relatively low N2 biases for left stress. Similarly, as shown in this and the preceding section, a noun that modifies few other nouns, i.e. has a low N1 family size (resulting in high conditional probability of N2), predisposes a compound to left stress when it occurs in N1 position. Nouns with low N1 family sizes will therefore tend to have relatively high N1 biases for left stress. In fact, any property of a constituent that is predictive of stress placement will automatically contribute to that constituent's bias towards a particular pattern.

Since informativity is taken here to be a property of individual constituents, and is computed using constituent-based variables in our models, any effect of informativeness on stress will contribute to and even give rise to a constituent identity effect: more informative constituents in N1 position will have greater N1 biases for left stress, while more informative constituents in N2 position will have lower N2 biases for left stress. Whether the constituent family bias indeed incorporates other

constituent-related measures can be tested empirically. Thus, it should be possible to predict the stress bias of a given constituent on the basis of other known properties of that constituent, including its informativity. This will be done in the next subsection.

4.3 Constituent identity and informativity as related measures

In order to test whether informativity and other constituent properties underlie the constituent identity effect, we fitted two regression models: one with the constituent family bias of N1 as the dependent variable, the other with the constituent family bias of N2 as the dependent variable. The independent variables were all coded predictors related to N1 or N2, respectively. For N2 bias, the predictors were the length of N2 in syllables, the positional family size of N2, and its synset count. For N1 bias, the predictors were the number of syllables in N1 after the main stressed syllable, the conditional probability of N2 (i.e. 1/positional family size of N1), and the synset count of N1. We expected to find significant effects for at least some of these predictors. Our final models are shown in table 5 for N1, and table 6 for N2, with the corresponding graphs in figure 4.

Insert tables 5 and 6 here

Place figure 4 about here

It can be seen that for both constituents, length and informativeness are significant predictors of stress bias. For each constituent, an increase in length is associated with a decrease in bias for left stress, reflecting the fact that longer compounds tend to be right stressed. In the case of N1, the effect of informativity is represented by conditional probability: the more likely is N2 given N1, the greater the N1 bias for left

stress, as predicted by our hypothesis. For N2, the effect of informativity is represented by its synset count: the greater the synset count, i.e. the less semantically specific N2 is, the greater its bias for left stress, as expected. The low values of R-squared for these models indicate that they would not be very successful at predicting the stress bias of new items, presumably because various other significant predictors are missing from the analyses. At the very least, semantic information would need to be added since, as discussed in section 2.1, it is clear that some semantic categories are likely to affect bias. Nevertheless, despite the fact that some other predictors are missing, there is clear evidence that both length and informativity contribute significantly to the effect of constituent identity on compound stress.

If indeed it does subsume all other constituent-based predictors, then it is not surprising that constituent identity is consistently found to be the most reliable and significant predictor of compound stress. More surprising, if informativeness is one of the factors underlying bias, is the fact that conditional probability of N2 emerges as a significant predictor of stress, even when the constituent biases are also included in the model. However, it should be remembered that the bias and informativity measurements used in our analyses are only very rough approximations to what might be found in the mind of any given speaker. Our constituent family biases, in particular, were calculated on the basis of only 1154 tokens (those for which we have agreed stress ratings), and may therefore be even less reliable than the family sizes, which were computed from the much larger families extracted from COCA.

In order to further explore the constituent identity effect, we fitted a model with N1 and N2 as random effects (Baayen *et al.* 2008). In the same way that including speaker as a random effect allows us to jointly model a range of factors including sex, age, L1 dialect, education and so on, including N1 and N2 as random effects allows us

to combine a host of variables, including but not restricted to semantics, phonology, structure, distribution and frequency. In a model of compound stress that includes N1 and N2 as random effects, the model takes into account the tendencies of each compound's N1 and N2 to be associated with particular stress patterns in the data. And although this is effectively what the constituent family bias also does, we might expect the model with random effects to be even more successful in its predictions. This is because the possible values of family bias are limited by the number of compounds in the family: in a family of five compounds, for example, the possible values are 0, .25, .5, .75 and 1.0. This means that constituents with the same family size are restricted to the same possible values of bias. The random effects, in contrast, can be different for every N1 and N2 in the data, so that the resulting model can fit the data much more closely.

Based on these considerations we fitted a mixed effects model to our data, with stress position as the dependent variable and speaker, N1 and N2 as random effects. The resulting model has an extremely high *C* value of 0.960, which is not improved by the addition of any of the fixed effects discussed above. Furthermore, removing the random effect of speaker only slightly reduces the *C*-value, to 0.956. This suggests that nearly all the stress variation in our data can be accounted for in terms of the tendencies of the two constituents to be associated with one stress pattern or the other. In other words, for the model to correctly assign stress to almost any compound in the dataset, the only information needed is how other compounds are stressed that share a constituent with the compound in question.

The success of the random-effects model supports our hypothesis that the effect of constituent identity on stress subsumes those of family size, synset count and length. However, it also begs the question as to whether, in the mental lexicon too, stress is

assigned purely on the basis of constituent identity: in other words, whether the observed effects of informativity and semantics are simply side-effects of the constituent identity effect. In fact, this is very unlikely to be the case. If stress were assigned only on the basis of constituent identity, there is no particular reason why constituents with similar levels of informativity or with similar semantics should cluster together, and so there would be no way of explaining the well-established correlations between semantic classes and stress position, or of explaining the effects of informativity that we have confirmed in this paper. While an effect of constituent informativity on stress will automatically give rise to constituent identity effects like the family stress bias, as discussed in section 4.2, the reverse is not true: biases based only on constituent identity would not automatically produce an informativity effect.

5 Discussion and Conclusion

In this paper we have investigated the role of informativity as a determinant of compound stress assignment in English and explored its relationship to other predictors. The results of our analyses replicate the informativity effects found by Bell & Plag (2012) for a different set of data, and therefore lend further support to the idea that informativity is indeed predictive of stress. Compounds with a relatively informative second constituent are more likely to be right stressed than compounds with a less informative second constituent. This result, while in line with that of Bell (2013) and Bell & Plag (2012), contradicts the findings by Plag & Kunter (2010), who used the same set of compounds as we have done, but did not find an informativity effect. However, they used only one informativity measure, family size, and this was based on constituent families derived from a rather small corpus: our results suggest that this corpus was too small to produce useful estimates of family size. In contrast,

COCA provides more representative family sizes, which, as we have demonstrated, lead to the expected informativity effects in the statistical models.

The present study also investigated the relationship between constituent identity effects and informativity. Our analysis confirmed previous findings that constituent identity is the strongest single predictor of compound stress pattern (e.g. Plag 2010, Arndt-Lappe 2011). However, we also saw that that the two types of measure are not independent of each other, and that a constituent's stress bias can be partially predicted by its informativeness. This raises the question as to how we can understand the relationship between constituent identity and informativity as predictors of compound stress.

Let us start with the general principle that more informative elements are stressed. Given this principle, it is possible to hypothesize how other observed effects, in particular the constituent identity effect, might arise as a result. Recall that any constituent whose informativity exerts an influence on prominence will exert this influence not only in the compound one happens to look at, but in all its compounds, i.e. in its whole family. Thus, informativity directly translates into a constituent family bias towards a particular kind of stress.

This in turn raises the question as to why, if a constituent's stress bias is based on informativity, it should be more highly predictive than the property from which it derives. At this point we need to remind ourselves that the informativity measurements we are using can only be imperfect approximations to the information content of any particular token in the mind of a speaker. For example, using synset count to estimate semantic specificity can only provide a crude estimate of what is likely in reality to be a highly complex property. Furthermore, our family sizes, though based on a fairly large corpus, can only approximate to actual probabilities of

occurrence, not least because they only take into account bigram probabilities. In reality, the informativity of a particular constituent token will depend not only on its immediate neighbour in the compound, but also on that constituent's distribution with respect to the wider syntactic and discourse context, as well the encyclopedic knowledge of the speaker and listener. So it could be that constituent identity is actually a better approximation to informativity than our rather imperfect measurements of semantic specificity and family size.

Constituent identity embodies a whole range of other lexical properties, some of which are known to be predictive of stress placement and may themselves be related to informativity. For example, our results demonstrate that constituent length is a strong predictor both of compound prominence and of constituent stress bias; but constituent length itself can be predicted on the basis of informativity. Piantadosi *et al.* (2011) show that, across a range of languages, information content is a very reliable predictor of word length, much better even than frequency, whether length is measured in terms of letters, syllables or phonemes. The longer a word the more informative it is. It is therefore possible that the very strong effects of length that we see in our data, which replicate those found by Bell (2013) and Bell & Plag (2012) for the BNC, are in fact related to informativity.

Where the length of N2 is concerned, it is clear how its effect on prominence could result from informativity: the longer N2, the more informative it is and therefore the more likely to be stressed. However, recall that, although the length of N2 is a strong predictor of prominence, we found that the number of syllables after the main stress of N1 was an even better predictor. This measure clearly makes reference to the phonological structure of N1 as well as to the length of N2, and its effect on prominence is therefore more difficult to explain in terms of informativity. Another

possible explanation is that there is a phonological constraint against long strings of unaccented syllables (cf. Ladd 2008: 244) and we do not preclude the possibility that such an effect operates in addition to an informativity-mediated effect of N2 length. How exactly the relationship between these effects can be modeled and better understood is a matter for future research.

The informativity principle might also underlie the known semantic effects on compound stress. Let us first consider the semantic properties of N1 and N2. Tarasova (2012: 55-56) shows that, for English simplex nouns, their positional family size in N1 position is inversely correlated with their positional family size in N2 position. In other words, some nouns tend to occur in compounds mainly as the left-hand constituent, while others occur mainly as the right-hand constituent (see also Baayen 2010 for a similar finding). It is plausible that the preference of a noun for one position or the other is associated with its semantic properties.

Fanselow (1981: 156, 174ff, 192ff) argues for a semantic classification of compounds into those that show basic relations (GRUNDRELATIONEN) and those that show stereotypical relations (STEREOTYPENRELATIONEN). Basic relations arise from basic properties common to all things: size, shape, location, material etc., whereas stereotypical relations arise from the stereotypes represented by particular nouns. We might hypothesize that nouns that represent basic properties such as material, size or location are likely both to have a preference for N1 position and to have large N1 family sizes: their semantics as fairly general modifiers will make it possible for them to modify a large number of head nouns. If it is true that semantic classes associated with basic relations do have large N1 family sizes and also tend to be associated with rightward stress, then informativity can account for the correlation of semantic class with stress type. This correlation has been recognized as a conundrum in earlier

studies, and the informativity approach opens up new avenues of research into this problem.

Informativity could also underlie the effects on stress of particular semantic relations between N1 and N2. Tarasova (2012: 57-59) shows that the greater a noun's family size, either in N1 or N2 position, the greater its tendency to be associated with a particular semantic relation. What is not clear, is whether nouns with large N1 families tend to favor a particular set of relations, while those with large N2 families favor a different set. If it turns out that large N1 family size is generally associated with relations favoring right prominence, while large N2 family size is associated with different relations, then informativity might account for the effects of semantic relation on stress. Semantic relations known to predict right prominence would do so because they are associated with N1 constituents that have large family sizes and thus render N2 relatively informative. Similarly, relations known to favor left prominence would do so because they are associated with N2 constituents that have large families and are therefore relatively uninformative. On the other hand, if there sometimes turns out to be a mismatch between the stress pattern predicted by the family size and that predicted by the semantic relation, then these conflicting pressures could give rise to some of the within-type variability found in natural data.

A limitation of our models is that our measures of informativity are based on orthographic representations of the constituents. This makes them rather crude, since they can take no account of the polysemy or even homonymy of individual constituents. For example, *savings bank*, *merchant bank, river bank* and *canal bank* would all be counted as members of the same N2 positional constituent family, even though *bank* clearly has a different meaning in *savings bank* and *merchant bank* than it does in *river bank* and *canal bank*. However, with more sophisticated measures of

informativity, it would be possible to distinguish between the different readings of a given constituent, for example by calculating family sizes based on particular readings.

Such more sophisticated measures might also be able to account for well-known pairs such as *tóy factory* vs. *toy fáctory*, where different stress patterns coincide with different semantic interpretations: *tóy factory* is generally taken to mean 'a factory for making toys' while *toy fáctory* is 'a model factory that is a toy' (see Bell 2013 for empirical evidence of this). When *toy factory* means 'a factory for making toys', the constituent *factory* has its core meaning of 'a building with machinery for the manufacture of goods'. With this meaning it can occur in N2 position with a wide variety of nouns in N1 position, reflecting the wide variety of goods that can be manufactured in a factory. This large N2 family makes *factory* relatively uninformative, so the left-stress pattern associated with this reading is to be expected. On the other hand, when *toy factory* means 'a model factory that is a toy', *factory* has undergone a metonymic shift to mean 'a model of a factory'. With this metonymic meaning one would assume that it would have a much smaller family size, since world knowledge suggests that there are fewer potential values of N1 that would make sense. This small family would make *factory* with this metonymic reading relatively informative, so that the attested rightward stress is again expected. In other words, the two stress patterns would fall out from informativity.

An exploratory empirical analysis suggests that this explanation is on the right track. Again using COCA, we checked in context every token of the fifty most frequent NN compounds with *factory* as N2. In forty-six of these types, the constituent *factory* was used exclusively with its core meaning, e.g. *shoe factory*, *paint factory*, *munitions factory*. Only in the remaining four types, namely *dream factory, hit factory,*

idea factory and *soul factory*, did it occur with a different reading. These attested minority readings all show the same kind of semantic extension of *factory* to mean 'institution producing N1', and there is not a single attestation of the metonymic reading 'model of a factory'. This patterning of the data constitutes evidence for the idea developed here that a semantically more fine-grained measure of informativity could indeed account for minimal stress pairs that show a semantic contrast.

Let us turn to another implication of our findings. The fact that informativity is highly influential in determining compound prominence speaks for an analysis of compound stress as an intonational phenomenon, rather than one of lexical stress assignment. This means that, at the theoretical level, compound prominence needs to be integrated into a more general account of the larger prosodic organization of utterances. This integration could be framed in different ways. Using the terminology of Calhoun (2010: 2), two alternative frameworks would be the PITCH-ACCENTING APPROACH (e.g. Bolinger 1965, Rochemont 1986, Selkirk 1995, Gussenhoven 2004) and the METRICAL STRESS APPROACH (e.g. Calhoun 2010), but the details of any such account would need to be worked out in future research. It should be noted, however, that accounts based on intonation would not necessarily be incompatible with one based on lexical stress. In psycholinguistically plausible models of the mental lexicon (e.g. Libben 2010), with rich and redundant information storage, recurrent patterns of prominence would naturally lead to memorized stress patterns for individual compounds. Recently, Schweitzer *et al.* (2011) have provided evidence that prosodic realization can indeed be subject to lexicalized entrenchment.

This in turn has wider implications for general accounts of lexical stress. For many words, compound or non-compound, stress information seems to be part of the lexical entry (e.g. Cutler 1984). This information, that the verb *devíse* has stress on the

last syllable, for example, can be overridden if informativity considerations arising from the surrounding discourse require it. Contrastive stress, on compounds or other words, means that in the given context a particular string is extremely informative and is thus pronounced with high prominence (perhaps *modulo* certain metrical constraints, see Calhoun 2010), irrespective of its normal level of informativity. Utterances such as *I said dévise, not révise*, or *It was an apple cáke, not an apple píe*, would be cases in point. Returning to our informativity measures, such measures as family size should be seen as approximating to a kind of baseline informativity from which language use starts. On top of this, discourse can build the actual informativity of tokens in context, on the basis of which pitch accents are distributed over utterances. A picture therefore emerges in which prosodic realizations based on frequent patterns of relative informativeness can become entrenched as lexical (i.e. memorized) stress, while at the same time being over-ridable should relative informativeness change in a particular context.

An informativity-based account of compound stress would naturally extend to compounds that have more than two constituents. As shown, for example, by Giegerich (2009) and Kösling & Plag (2009), stress in triconstituent nominal compounds is also highly variable and can fall on any of the three nouns involved. Even though they speak of factors other than informativity, Kösling & Plag (2009) Kösling (2012) and Kösling *et al.* (2013) explicitly propose that the same mechanisms regulate stress assignment in triconstituent compounds as in those compounds that have only two constituents. Bell (2013) builds a variety of models of compound stress using a data set that contains both binomial and trinomial compounds. In her models these larger compounds do not behave differently from simple NN compounds and show the same kinds of informativity effects. It is a matter for future research to

replicate these effects for even larger compound structures.

References

Arndt-Lappe, Sabine 2011. Towards an exemplar-based model of English compound stress. *Journal of Linguistics* 47(3): 549-585.

Baayen, R. Harald, Doug J. Davidson & Douglas M. Bates 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59(4): 390-412.

Baayen, R. Harald 2010. The directed compound graph: an exploration of lexical connectivity and its processing consequences. *Linguistische Berichte* Sonderheft 17: 383-402.

Bates, Douglas, Martin Maechler & Ben Bolker 2011. *lme4: Linear mixed-effects models using s4 classes*. R package version 0.999375-41.

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 J. 2011. At the boundary of morphology and syntax. In Alexandra Galani, Glyn Hicks & George Tsoulas (eds.), *Morphology and its interfaces*, vol. 178 Linguistics Today. Amsterdam: John Benjamins Publishing Company. 137-167.

Bell, Melanie J. 2013. *The English noun noun construct: its prosody and structure*. Cambridge: University of Cambridge dissertation.

Bell, Melanie J. & Ingo Plag 2012. Informativeness is a determinant of compound stress in English. *Journal of Linguistics* 48(3): 485-520.

Bertram, Raymond, R. Harald Baayen & Robert Schreuder 2000. Effects of family size for complex words. *Journal of Memory and Language* 42(3): 390-405.

Bolinger, Dwight 1972. Accent is predictable (if you're a mind-reader). *Language* 48: 633-644.

Calhoun, Sasha (2010). The Centrality of Metrical Structure in Signaling Information Structure: A Probabilistic Perspective, *Language* 86(1): 1-42.

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

Cutler, Anne 1984. Stress and accent in language production and understanding. *Intonation, Accent and Rhythm: Studies in Discourse Phonology* 8: 76-90.

Davies, Mark 2004-. British national corpus. http://corpus.byu.edu/bnc/.

Davies, Mark. 2008-. The Corpus of Contemporary American English: 425 million words, 1990-present. http://corpus.byu.edu/coca/.

Fanselow, Gisbert 1981. *Zur Syntax und Semantik der Nominalkomposition*, vol. 107 Linguistische Arbeiten. Tübingen: Niemeyer.

Fellbaum, Christiane 1998. *Wordnet: an electronic lexical database*. Cambridge, MA: Bradford Books.

Giegerich, Heinz J. 2009. The English compound stress myth. *Word Structure* 2(1): 1-17.

Gussenhoven, Carlos 2004. *The phonology of tone and intonation*. Cambridge: Cambridge University Press.

Hayes, Bruce & Timothy Ma 2003-. Query Google. http://www.linguistics.ucla.edu/people/hayes/QueryGoogle/.

Jespersen, Otto 1909. *A modern English grammar on historical principles: part 1, sounds and spelling*. Heidelberg: Carl Winter's Universitätsbuchhandlung.

Jones, Daniel 1922. *An outline of English phonetics*. 2nd edn. Cambridge: W. Heffer & Sons.

Kösling, Kristina 2011. *Prominence assignment in English triconstituent compounds.* PhD thesis, Universität Siegen.

Kösling, Kristina & Ingo Plag 2009. Does branching direction determine prominence assignment? *Corpus Linguistics and Linguistic Theory* 5(2): 201-239.

Kösling, Kristina, Gero Kunter, Harald Baayen & Ingo Plag 2013. Prominence in triconstituent compounds: Pitch contours and linguistic theory. To appear in *Language and Speech*.

Kunter, Gero 2011. *Compound stress in English: the phonetics and phonology of prosodic prominence*. Berlin: de Gruyter.

Kunter, Gero & Ingo Plag 2007. What is compound stress? In Jürgen Trouvain & William J. Barry (eds.), *Proceedings of the 16th international congress of phonetic sciences*. Saarbrücken. 1005-1008.

Kutner, Michael H., Christopher J. Nachtsheim, John Neter & William Li 2005. *Applied linear statistical models*. 5th edn. Boston, MA: McGraw-Hill Irwin.

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

Ladd, D. Robert. 2008. *Intonational phonology*, 2nd ed. Cambridge: Cambridge University Press.

Libben, Gary 2010. Compound words, semantic transparency, and morphological

transcendence. *Linguistische Berichte* Sonderheft 17: 317-330.

Marchand, Hans 1969. *The categories and types of present-day English word formation: a synchronic-diachronic approach*. München: Beck'sche Verlagsbuchhandlung.

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

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

Pan, Shimei & Julia Hirschberg 2000. Modeling local context for speech accent prediction. In *Proceedings of the 38th annual meeting of the association for computational linguistics*. 233–240.

Pan, Shimei & Kathleen R. McKeown 1999. Word informativeness and automatic pitch accent modeling. In Pascale Fung & Joe Zhou (eds.), *Proceedings of EMNLP/VLC'99*. 148–157.

Payne, John & Rodney D. Huddleston 2002. Nouns and noun phrases. In Rodney D. Huddleston & Geoffrey K. Pullum (eds.), *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press. 323-523.

Piantadosi, Steven T., Harry Tily & Edward Gibson 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences* 108(9): 3526-3529.

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 2010. Compound stress assignment by analogy: the constituent family bias. *Zeitschrift für Sprachwisenschaft* 29(2): 243-282.

Plag, Ingo & Gero Kunter 2010. Constituent family size and compound stress assignment in English. *Linguistische Berichte* Sonderheft 17: 349-382.

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): 199-233.

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): 760-794.

Plag, Ingo, Gero Kunter & Mareile Schramm 2011. The phonetics of primary and secondary stress in North American English. *Journal of Phonetics* 39: 362-374.

Moscoso del Prado Martín, Fermin, Raymond Bertram, Tuomo Häikiö, Robert Schreuder & R. Harald Baayen 2004. Morphological family size in a morphologically rich language: the case of Finnish compared with Dutch and Hebrew. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 30(6): 1271.

R Development Core Team 2011. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing Vienna, Austria. http://www.Rproject.org/. ISBN 3-900051-07-0.

Rochemont, Michael S. 1986. *Focus in generative grammar*. Amsterdam: John Benjamins.

Selkirk, Elisabeth. 1995. Sentence prosody: Intonation, stress and phrasing. In John Goldsmith (ed.), *The handbook of phonological theory.* Oxford: Blackwell. 550–69.

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

Schreuder, Robert & R. Harald Baayen 1997. How complex simplex words can be.

Journal of Memory and Language 37(1): 118-139.

Schweitzer, Katrin, Michael Walsh, Sasha Calhoun & Hinrich Schütze 2011. Prosodic variability in lexical sequences: intonation entrenches too. In Wai-Sum Lee & Eric Zee (eds.), *Proceedings of the international congress of phonetic sciences*. Hongkong. 1778–1781.

Shannon, Claude E. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27: 379-423.

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

Sweet, Henry 1892. *A new English grammar: logical and historical part 1, introduction, phonology and accidence*. Oxford: Clarendon Press.

Tarasova, Elizaveta 2012. *Nominal compounds in English and Russian: cognitive perspective.* Victoria University of Wellington dissertation.

Tables

Table 1: Semantic categories found to influence compound prominence (Plag et al. 2007, 2008)

Semantic category	**Direction of influence**
Semantic property of constituents	
N1 refers to a period or point in time	rightward stress
N2 is a geographical term	rightward stress
N1 and N2 form a proper noun	rightward stress
N1 is a proper noun	rightward stress
N1 and N2 form a left-headed compound	rightward stress
Semantic relation between constituents	
N1 has N2	rightward stress
N2 is made of N1	rightward stress
N1 is N2	rightward stress
N2 located at N1	rightward stress
N2 during N1	rightward stress
N2 is named after N1	rightward stress
N2 for N1	leftward stress
N2 uses N1	leftward stress

Table 2: Predictors initially present in the analysis

constituent identity:	N1 constituent family stress bias
	N2 constituent family stress bias
lexicalization:	compound frequency
informativity:	positional family size of N2
	conditional probability of N2 based on family size of N1
	synset count of N2
	synset count of N1
length:	number of syllables following the main-stressed syllable of N1

Table 3: Final model, informativity only, *N* = 1056

Random effects					
	Variance	Std.Dev.			
Speaker (Intercept)	0.32425	0.56943			
Fixed effects					
	Estimate	Std. Error	*z* value	Pr(>|*z*|)	
(Intercept)	-4.31557	1.09091	-3.956	7.62e-05	***
Syllables after N1 main stress	0.63776	0.06663	9.572	< 2e-16	***
Log frequency of NN	-0.24864	0.03455	-7.197	6.18e-13	***
Log family size of N2	-0.18441	0.09681	-1.905	0.0568	•
Log conditional probability of N2	-0.88428	0.11064	-7.992	1.32e-15	***
Model fit					
C	Dxy				
0.8210719	0.6421439				

Significance codes: *** p < .001; ** p < .01; * p < .05; • marginal

Table 4: Final model, informativity and family bias, *N* = 1056

Random effects					
	Variance	Std.Dev.			
speaker (Intercept)	0.26034	0.51024			
Fixed effects					
	Estimate	Std. Error	*z* value	Pr(>|*z*|)	
(Intercept)	-0.52151	0.97552	-0.535	0.593	
Syllables after N1 main stress	0.35795	0.07455	4.801	1.58e-06	***
Log frequency of NN	-0.17075	0.03733	-4.574	4.73e-06	***
Log conditional probability of N2	-0.49533	0.11748	-4.216	2.48e-05	***
N1 bias for left stress	-1.75454	0.29579	-5.932	3.00e-09	***
N2 bias for left stress	-2.25271	0.27438	-8.21	< 2e-16	***
Model fit					
C	Dxy				
0.8550325	0.7100651				

Significance codes: *** p < .001; ** p < .01; * p < .05; • marginal

Table 5: Final model of N1 bias for left stress as predicted by constituent-related variables, adjusted R-squared = 0.2404

	Value	Std. Error	*t* value	Pr(>|*t*|)	
(Intercept)	1.7269	0.07639	22.61	<2e-16	***
N1 Syllables after N1 main stress	-0.1792	0.01089	-16.47	<2e-16	***
Log conditional probability of N2	0.1275	0.01045	12.20	<2e-16	***

Table 6: Final model of N2 bias for left stress as predicted by constituent-related variables, adjusted R-squared = 0.1753

	Value	Std. Error	*t* value	Pr(>|*t*|)	
(Intercept)	0.83868	0.03149	26.631	<2e-16	***
N2 Syllables	-0.10949	0.00960	-11.405	<2e-16	***
Log synset count of N2	0.06382	0.01169	5.461	5.907e-08	***

Figures

Figure 1: Partial effects in final model, informativity only

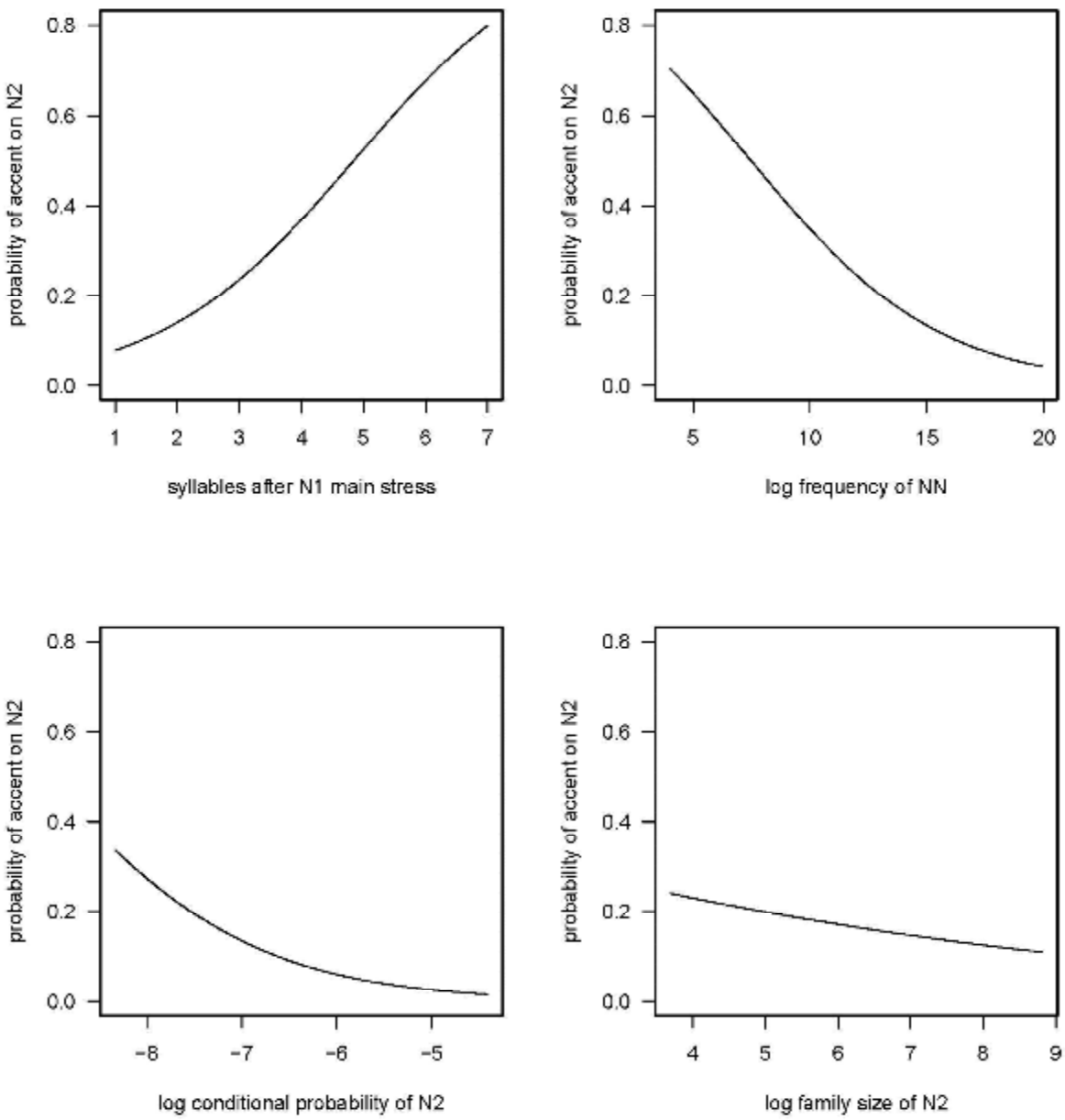

Figure 2: Partial effects of synsets, model under exclusion of probability measures

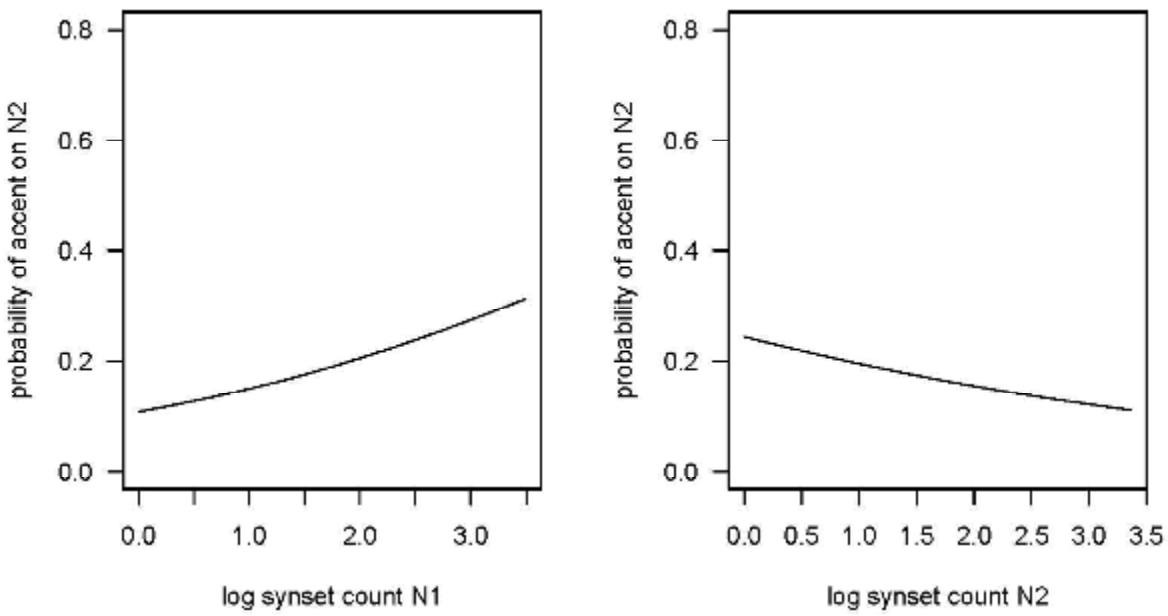

Figure 3: Partial effects for the model with informativity and family bias

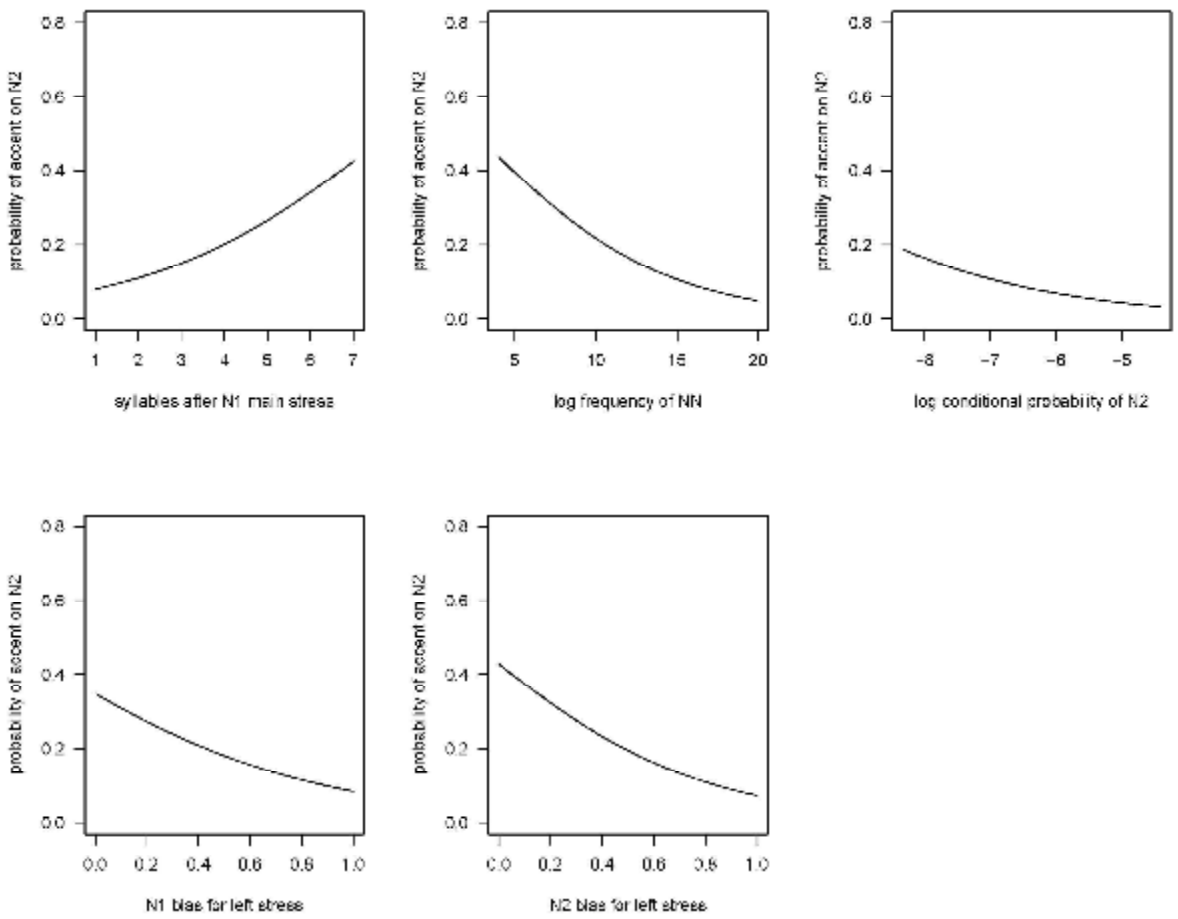

Figure 4: N1 and N2 family biases, as predicted by the respective constituent-related variables