The Variability of Compound Stress in English: Towards an Exemplar-Based Alternative to the Compound Stress Rule

Sabine Lappe
University of Siegen
lappe@anglistik.uni-siegen.de

Ingo Plag
University of Siegen
plag@anglistik.uni-siegen.de

Abstract

Recent research has shown that the Compound Stress Rule cannot adequately handle the variable stress behaviour of noun-noun constructs in English (e.g. Plag, 2006, Plag et al., 2006a, 2006b). In this paper we present an analysis of compound stress assignment in exemplar-bases models, using data from two corpora, CELEX (Baayen et al., 1995) and the Boston University Radio Speech Corpus (Ostendorf et al., 1996). The data were coded according to a number of semantic and structural criteria taken from the literature on compound stress. Two different algorithms are tested, TiMBL 5.1 (Daelemans et al., 2004) and AM::Parallel (Skousen et al., 2004), and it turns out that both analogical algorithms are superior to previous, categorical approaches.

Credits

This research was funded by the Deutsche Forschungsgemeinschaft (Grant PL 151 / 5-1). Special thanks go to Gero Kunter for his splendid management of the CELEX and BURSC data, and to Maria Braun for helping us code the data.

1 Introduction

The Compound Stress Rule (Chomsky & Halle, 1968) is one of the most wellknown stress rules in English. It states that in English noun-noun compounds, the left-hand constituent is more prominent than the right-hand constituent. However, it is also wellknown that not all English noun-noun compounds abide to the Compound Stress Rule. Examples of both left-stressed and right-stressed noun-noun compounds are provided in (1). The most prominent syllable is marked by an acute accent.

(1)	*ópera glasses*	*steel brídge*
	wátch-maker	*morning páper*
	clássroom	*silk tíe*
	Òxford Street	*Madison Àvenue*

Rightward stress is far from exceptional, and the nature of the observable variability is still rather unclear. Recent investigations (e.g. Plag, 2006, Plag et al., 2006a, 2006b) have shown, however, that categorical approaches (such as the Compound Stress Rule) are unsuccessful in making correct predictions about compound stress assignment. This paper will present an analysis of the variation in compound stress assignment in exemplar-based models. We will use two current exemplar-based algorithms, TiMBL (Daelemans et al., 2004) and AM (Skousen et al., 2004), to investigate whether an exemplar-based approach to compound stress is empirically more adequate than a categorical or probabilistic model. As data we will use the data from the two Plag et al. (2006a, 2006b) studies. These data comprise all noun-noun compounds extracted from the Boston University Radio Speech Corpus (BURSC, Ostendorf et al., 1996) and all compounds extracted from the CELEX lexical database (Baayen et al., 1995).

The paper is structured as follows. In section 2 we will set the stage, introducing our theoretical assumptions (section 2.1), the BURSC and CELEX data and coding (section 2.2), and the methodological principles that guided our TiMBL and AM experiments (section 2.3). Sections 3 and 4 will then report our findings. The paper ends with a conclusion and outlook onto future research.

2 Setting the Stage

2.1 English compound stress

Whereas it is still general common ground that the Compound Stress Rule is the major predictor of compound stress, there are three types of approaches to deal with the observable variation in compound stress assignment, which we may label syntactic, semantic, and analogical. Right stress has been claimed to be an effect of the syntactic, i.e. phrasal nature of a construction, whereas left stress is associated with lexical, i.e. morphological, items. A recent proponent of this structure-based approach is Giegerich (2004), who claims that argument-head compounds such as *bookseller* are morphological entities and thus generally left-stressed, while modifier-head compounds are generally stressed on the right. Apparent exceptions, such as left-stressed *opéra glasses* are the result of the lexicalization of an originally phrasal structure.

Other people (e.g. Sampson, 1980, Fudge, 1984, Ladd, 1984, Liberman and Sproat, 1992, Olsen, 2000, 2001, Spencer, 2003) have claimed that right stress is triggered by specific semantic relations between left and right constituents (such as, for example, material or locative relations as in *steel brídge* or *Boston hárbor*). Finally, exceptional right stress has been explained as the effect of analogy (e.g. Schmerling, 1971, Liberman and Sproat, 1992), with compounds having the same right or left constituent sharing the same stress pattern. Standard examples of the latter phenomenon are street names, which are stressed on the left if the right constituent is *street*, but right-stressed if the right constituent is *avenue* or *lane* (as in *Óxford Street* vs. *Oxford Ávenue*, *Oxford Láne*).

These different approaches have recently been tested against large amounts of empirical data by Plag et al. (2006a, 2006b), who investigated the phonetic properties and the determinants of compound stress assignment in several thousands of compounds in two corpora, CELEX and BURSC. The CELEX database (Baayen et al., 1995) contains mostly dictionary, i.e. lexicalised, data, while the Boston Radio Speech Corpus (BURSC, Ostendorf et al., 1996), contains audio recordings of radio news texts. In their studies Plag et al. found a number of new interesting insights about the nature of compound stress, the most relevant of which are:

- Neither the Compound Stress Rule nor any of the syntactic or semantic or analogical factors proposed in the literature can adequately explain compound stress in a categorical fashion.
- Statistical analysis reveals a relatively large amount of unexplained variation. Variation is found among compound types as well as among tokens of the same type.
- In spite of the fact that compounds from CELEX and BURSC exhibit striking differences with respect to their status of lexicalisation, the two corpora yield strikingly similar, almost parallel findings.

For the present study in exemplar-based modeling we used the data from the Plag et al. studies, as well as their coding of the data, to which we now turn.

2.2 The coding

In order to test the effects of argument structure, semantics and analogy, Plag et al. first extracted all NN structures from BURSC and CELEX.[1] In what follows we will use the term 'compounds' as a convenient label to refer to these structures. We thus remain deliberately agnostic with respect to the question of whether or not some of these structures should be attributed a phrasal status. It should be emphasised, however, that all our compounds are of a kind to which is attributed word status, not phrasal status, in the general descriptive literature. The total number of NN constructions extracted from BURSC is 4410 tokens, which are distributed among 2476 different types. The total number of NN compounds extracted from CELEX is 4491 (types).

In the present study we used only types, not tokens. Furthermore, we used subsets of the two corpora which comprise those compounds that have a constituent family, i.e. for which there are other compounds that share the same right or left constituent. The rationale behind this choice was that we wanted to make sure that all sources of information were available for all compounds.

[1] While this is straightforward for the lexical database CELEX, all texts from the BURSC had to be manually annotated for all sequences consisting of two (and only two) adjacent nouns, one of which, or which together, functioned as the head of a noun phrase. From this set proper names such as *Barney Frank* and those with an appositive modifier, such as *Governor Dukakis* were eliminated. The exclusion of these two types of structures was based on two considerations. First, we would expect these to show consistent rightward stress, second we know of no claims that these structures would be regarded by anyone as compounds.

For reasons of consistency, we used the same subsets in all experiments. For BURSC the total number of compounds in our subset is 722. The corresponding number for CELEX is 2643. All compounds were coded in terms of

- the orthographic representation of their left and right constituents
- the structural and semantic features held to be responsible for stress assignment in the literature
- the stress category (left or right)

For each compound the structural and semantic features were coded independently by two raters. The coding categories are given in (2) – (4). We broadly distinguish between three types of features: argument structure, semantic categories, and semantic relations. In what follows we will use the terms 'argument structure', 'semantic categories', and 'semantic relations' as convenient labels to refer to these sets of features.

(2)

Argument Structure	**Example**
argument-head	*computer maker*
modifier-head	*truck accident*
morphology of head: -er	*computer maker*
morphology of head: -ing	*arts funding*
morphology of head: -ion	*habitat acquisition*
morphology of head: conversion	*budget cut*

(3) **Semantic property of constituent or compound**

N1 refers to a period or point in time
example: *day care*
N2 is a geographical term
example: *bay area*
N2 is a type of thoroughfare
example: *state road*
The compound is a proper noun
example: *Harvard University*
N1 is a proper noun
example: *Mapplethorpe controversy*

(4) **Semantic relation between the constituents of the compound**

Relation			**Example**
N2	CAUSES	N1	*retirement age*
N1	CAUSES	N2	*drug war*
N2	HAS	N1	*school district*
N1	HAS	N2	*state inspector*
N2	MAKES	N1	*computer company*
N1	MAKES	N2	*university research*
N2	IS MADE OF	N1	*paper drum*
N2	USES	N1	*biotech industry*
N1	USES	N2	*police effort*
N1	IS	N2	*jail facility*
N1	IS LIKE	N2	*crime wave*
N2	FOR	N1	*consumer advocate*
N2	ABOUT	N1	*health law*
N2	LOCATED at/..	N1	*neighborhood school*
N1	LOCATED at/..	N2	*school district*
N2	DURING	N1	*lifetime*
N2	NAMED AFTER	N1	*Mapplethorpe show*

Due to the well-known fact that many compounds are ambiguous and can be interpreted as belonging to more than one of the above semantic categories, each of the semantic categories had to be coded individually for each compound as a binary category (with 'yes' and 'no' as values). In addition to the categories in (2) – (4), the data were coded according to left and right constituents. These were given in their orthographic form. A sample of a coded item is given in (5).

(5) **The coding of *clinic worker* (BURSC)**

constituents	left	*clinic*
	right	*worker*
argument structure	argument?	*no*
	morphology of right constituent?	*-er*
semantic categories	...	*no*
semantic relations	N2 FOR N1	*yes*
	N2 LOCATED at N1	*yes*
	...	*no*
stress	stress	*left*

At this point a note is in order with regard to the coding of the target category, stress. For CELEX we relied on the classification of stress information as given in the corpus. For BURSC, the classification into left and right stress was based on the items' acoustic cues, using the algorithm proposed in Kunter & Plag (2007). The basis of this classification comprises both measurement and perception data on the acoustic correlates of stress: pitch (f0), intensity, duration, and jitter (as a correlate of creaky voice) in the right constituent. Kunter & Plag (2007) and Plag et al. (2006a) have shown that this model is highly reliable in predicting listeners' perceptions of prominence relationships in compounds from the BURSC corpus.

Nevertheless, it is important to note that our classification of stress in BURSC and CELEX is subject to two sources of error: First, the classification ignores the within-speaker, across-speaker and within-type (i.e. token) variability that actually exists in compound stress assignment (Kunter, 2007). In CELEX stress is treated as an invariant property of the compound lemmata, so that we have no information about how robustly left- or right-stressed a particular item is. In BURSC, we classified a given type as left- or right-stressed if the majority of tokens of that type had left or right stress. Second, the automatic classification of stresses according to acoustic cues inevitably generates some error. In spite of these problems, however, it turns out that the analysis of the compound data from the two corpora produces very robust findings with respect to the determinants of compound stress. Thus, despite all differences between the two corpora, the study based on BURSC (Plag et al., 2006a) and the study based on CELEX (Plag et al., 2006b) produced strikingly similar results. The same holds for the experiments in the present paper. We thus have good reason to believe that the only influence that classification errors may have on our study lies in slightly lower rates of predictive accuracy than we may expect without these additional sources of error.

2.3 The models -- TiMBL and AM

In sections 3 and 4 we will present the results of a series of tests in which we examine in how far two computational implementations of exemplar-based models, TiMBL (version 5.1, Daelemans et al., 2004) and AM::Parallel (Skousen et al., 2004), are able to predict the actual distribution of left and right stresses in BURSC and CELEX. Both algorithms classify new items on-line by comparing a new test item with similar items that are stored in memory.

TiMBL is a k-Nearest Neighbour (k-NN) system that encompasses several different implementations of memory-based learning techniques (cf. Daelemans et al., 2004 for details). The classification of a new item is extrapolated from similar exemplars that are explicitely stored in memory (the item's *nearest neighbours*), via a majority vote (which may optionally be weighted according to the distance between a given neighbour and the test item). Nearest neighbours are selected from a distance space k. The experimenter can manipulate k, so that she has some control over how narrow this space should be for a given experiment. Also for the computation of similarity, the experimenter can choose from a variety of different similarity measures, which conceptually fall into two different classes. In one class of measures, similarity is computed in terms of the simple number of matching values for all features given in the new input (*overlap metric*). Alternatively, the degree to which matches between input features and features of stored exemplars are relevant for the computation of similarity is influenced by feature weights (*Information-gain feature weighting* and others), or by weights which are able to distinguish between different values for a single feature (*MVDM, Jeffrey divergence metric*). Again, there are different ways in which feature weights may be computed. Crucially, however, the features weights are computed on the basis of the whole dataset that the system is given as training data. As a consequence, if feature weights are used in an experiment, they will be the same for every new input that is to be classified. Thus, for example, if the algorithm has found that in the training set the right constituent is more informative than the syntactic relation, it will assign to the right constituent of every exemplar in memory a higher weight value than to the syntactic relation. For every new input, similarity will thus be computed using the same feature weights.

AM treats feature weights differently. In AM, the relevance of features for the classification of a givent item is answered for every single new input on an individual basis (cf. Skousen, 1989, 2002a, b for details). The set of exemplars that is relevant for classification of a given input is termed the exemplar's *analogical set*. For every input, the algorithm checks all conceivable combinations of features (termed *contexts*) and determines in how far the set of exemplars in memory that match that combination behave in a homogeneous way with respect to the target category (i.e., in our case, stress assignment). Only homogeneous contexts will then be taken into account for the analogical set (cf. Skousen, 1989, 2002a, 2002b for details). In this process, also contexts which are more general and, hence, less similar to the context of the test item may be taken into account. Apart from the huge processing demands that this kind of procedure entails, an interesting difference between TiMBL and AM therefore lies in the degree to which different features may play a different role for different inputs. Here AM seems to be more flexible.

In all experiments in the present study, we tested the corpus on itself. That means that every item in the corpus was classified on the basis of all other items present in the same corpus. Both

TiMBL and AM provide parameter settings that can be used to implement this kind of experimental setup. In TiMBL the relevant procedure is the *leave-one-out* procedure. In AM, we used the same data set for both training and testing and set the parameters in such a way as to ensure that those items in the training set that are identical to those in the training set are excluded during classification.

In our experiments we focus on three important aspects:

- symbolic, rule-based models vs. TiMBL, AM: Which models are empirically more adequate, measured in terms of predictive accuracy?
- abstract vs. non-abstract features in TiMBL, AM: How abstract do the features have to be for a successful computation of compound stress?
- nearest neighbours vs. analogical set: Do the differences between TiMBL and AM in which exemplars they consider for classification result in different predictive accuracies?

The first aspect concerns the question of whether TiMBL and AM are empirically more accurate than categorical, rule-based models. To this end, we will compare the two models' predictive accuracies with the predictive accuracies that the pertinent rule-based models reach for the data in BURSC and CELEX.

The second aspect concerns the question of which features are relevant for the representation of generalisations about compound stress. The coding of data in Plag et al. (2006a, 2006b) gives us the opportunity to compare three different claims about the nature of linguistic representation, along which recent theories are divided. On the one hand traditional accounts have claimed that stress is computed on the basis of quite abstract information about the semantic, syntactic, morphological, or semantosyntactic status of a compound. We will use the term 'abstract features' from now on as a descriptive label to refer to the pertinent features. On the other hand, however, recent studies have shown that the level of representation which is relevant for the computation of stress may not be all that abstract. This position is found, for example, in occasional remarks in the literature about analogical effects that occur with compounds with specific left or right constituents or, more radically, in exemplar-based theories, where analogical effects with previously encountered, 'non-abstract' representations of stored exemplars is held to be the rule, rather than the exception. Whereas work on other aspects of derivational morphology and compounding has produced growing numbers of evidence for this position (e.g. Gagné, 2001, Krott et al., 2002, Chapman & Skousen, 2005), this view has never been tested for English compound stress. The BURSC and CELEX data, which have been coded both in terms of the pertinent abstract as well as in terms of the 'non-abstract' lexical representation of the left and the right constituent, provide an ideal testing ground to look at abstractness of representation.

We will do this with the help of three different experimental series. In series 1 we test the simple and, admittedly, overly simplistic hypothesis that it is either argument structure, semantic categories, or semantic relations that can predict compound stress. The computational model is thus provided with only one set of the pertinent features. In the second experimental series we test in how TiMBL and AM are successful in predicting stress if fed with the most informative combination of abstract features conceivable. Finally, in series 3 we feed TiMBL and AM only with the 'non-abstract' features, the left and right constituents of each compound.

3 Modelling Compound Stress in BURSC

3.1 The data, or: the demise of the rule-based approach

Table 1 provides an overview of the actual distribution of stresses in the corpus (N = 722).

left stress	right stress
358	364
49.58%	50.42%

Table 1. Distribution of stresses in BURSC.

The distribution in table 1 shows that the prediction of left stress as expected by the Compound Stress Rule does not correspond to the data. We will see in the analyses below that there is a huge amount of variation in the data, no matter according to which of the predictor categories we subdivide the data. As a consequence of that variation, not only the Compound Stress Rule, but also other traditional rule-based approaches to compound stress fail to account for the data. This has been shown in detail in Plag et al.'s BURSC study (2006a), whose main findings we will briefly summarise below.

Plag et al. test the predictive accuracy of approaches using argument structure (as proposed in Giegerich, 2004), semantic categories, and semantic relations (as proposed, e.g., in Fudge, 1984: 144ff., Liberman & Sproat, 1992, Zwicky, 1986), as predictors of compound stress. The central finding that emerged is that none of these three sets of predictors can adequately account for the variation that we find in the BURSC data. This is true for each set in isolation (argument structure, semantic categories, semantic relations) as well as for a combination of features from the three sets. This generalisation is independent of whether the features are conceptualised as predictors in a categorical rule model or in a logistic regression model. For reasons of space, we will limit the present discussion to rule-based approaches.

Table 2 provides an overall summary of the predictive accuracies for rule-based models that are reported in Plag et al. (2006a). We distinguish here between overall predictive accuracies and predictive accuracies for right and left stresses.

	predictive accuracies		
a rule system based on...	**overall**	**for left stress**	**for right stress**
argument structure:	53.8%	18.0%	85.3%
semantic relations and categories:	54.5%	30.3%	76.8%

Table 2. Predictive accuracies of categorical rules for BURSC (from Plag et al. 2006a).

It is important to note that the two approaches cannot be compared directly because, due to methodological constraints, the figures could not be computed from exactly the same set of the data (for the argument structure approach: N = 4091, for the semantic approach: N = 2027). However, in terms of the general distribution of all relevant predictor categories, the two subsets behave almost identically. The same is true for the subset of 722 items used in the present study.

Neither of the two approaches reaches an overall predictive accuracy that is significantly beyond a prediction by chance. If we look at predictive accuracies for left and right stresses in isolation, we see that both approaches are good at predicting right stress, but very poor in their predictions of left stress. This does not come as a surprise, given that both approaches have come into existence as an attempt to explain why there are so many exceptions to the Compound Stress Rule, i.e. why there are so many right-stressed compounds in English. What the table shows, however, is that they go too far in their predictions, grossly overpredicting right stress. Note that this also means that neither the syntactic nor the semantic approach can be saved if we combine them with the Compound Stress Rule. They would still predict right stress for the majority of items while their left-stress predictions converge with the predictions made by the Comound Stress Rule.

The question of the empirical accuracy of rule-based models is, however, made more complex than the findings in table 2 may suggest. Thus, Giegerich (2004) has noted that for modifier-head compounds left stress may arise as a consequence of lexicalisation. Following this line of reasoning, one could argue that overprediction of right stress as seen in table 2 only appears because lexicalisation has not been taken into account. Two remarks are in order here. Plag et al. (2006a) tested the lexicalisation hypothesis, using token frequencies and orthography as two different indicators of lexicalisation. The analyses of both indicators converged on showing that indeed there is a lexicalisation effect. However, the size of the effect is very small, and, more crucially, the effect is not restricted to the categories that are predicted to be right-stressed in Giegerich's (2004) account. The problem of lexicalisation is also highly interesting from an exemplar-based perspective. Given that under such an approach all items are ‚stored', both the woulde-be regular left-stressed items and the would-be irregularly right-stressed items would be available in memory and could thus serve as exemplars for analogical processes. Hence, we would expect lexicalization effects in both directions, and not only in the direction of left stress, as Giegerich would have it.

Yet another obvious question that emerges from table 2 is whether predictive accuracies of the approaches based on argument structure and semantics could be improved if they joined forces. As is clear from the table, we cannot expect much improvement under a rule-based paradigm. Both approaches underpredict left stress, and most of those compounds for which we do predict left stress based on argument structure are already included in the set of those compounds for which we predict left stress based on semantics.

3.2 The TiMBL and AM experiments - parameter settings

In TiMBL, highest classification accuracies were reached if similarity was computed using a simple overlap metric for the left and right member and the Jeffrey Divergence metric for all other features. Using a distance-weighted similarity metric for left and right members proved disadvantagous for the classification task, presumably because these features are represented in our corpora only as one single orthographic form. Potentially informative phonological characteristics like the number of syllables, syllable structure, rhythmic patterns were not represented. The system was thus not given suitable information that would allow it to establish similarities between different values for left and right constituents.

In experiments in which the algorithm was presented with a large number of features, classification was most successful if the distance space over which nearest neighbours were defined was set to k = 5. In experiments in which fewer features were used, k was adjusted so as to make sure that the nearest neighbour set never included the whole training corpus. The voting procedure that produced best results was Inverse Distance voting. Thus, neighbours that were closer in similarity to a given test item had a greater say in classification than more distant exemplars.

In our AM experiments we had the algorithm compute analogical sets using pointers, not occurrences (cf. Parkinson, 2002 for a general outline of the options provided by AM). All items were included in the classification process. Furthermore, it is problematic in AM to use only few and abstract features to test the corpus on itself. The reason is that no *leave-one-out* procedure is implemented. The experimenter is only given the option to exclude a data item from consideration during classification if its context (i.e. the features) is identical to the test item. However, in a data set in which we give the system only very few features, we expect the context of our test items to be represented in the dataset. This will give us the opportunity to test, for example, whether all argument-head compounds ending in –er are successful if used as predictors in the model. Therefore, we allowed AM to include every context in the evaluation, even if this context is given in the dataset. However, we expect predictive accuracies to be higher in this case, and, thus, the result not to be directly comparable to results in the other series or to the predictive accuracy reached by TiMBL.

3.3 Series 1 – only one set of abstract features as predictor

Table 3 provides an overview of classification accuracies for the first series of experiments.

We see that, in spite of the differences between the parameter settings, TiMBL and AM produce very similar classification results on the three codings. Like the rule-based accounts described in section 3.1, also exemplar-based models, if trained on argument structure and semantic categories, produce accuracies of classification hardly above chance level. Unlike rule-based accounts, however, TiMBL and AM are much better at predicting left stress than they are at predicting right stress.

information source	**accuracy**		
	overall	**for left stress**	**for right stress**
TiMBL (with k adjusted so that the k-NN set can never be the whole dataset):			
argument structure	52.35%	88.27%	13.19%
semantic categories	52.22%	92.46%	12.64%
semantic relations	56.37%	56.42%	56.32%
AM (every test item is also a member of the data set):			
argument structure	52.49%	93.3%	12.36%
semantic categories	52.77%	92.74%	13.46%
semantic relations	65.37%	65.36%	65.39%

Table 3. Classification accuracies for BURSC, series 1.

Finally, it is interesting to note that semantic relations feature differently from the other sets of abstract features. In both TiMBL and AM simulations they are the best information source. The difference between predictive accuracy of argument structure and semantic categories on the one hand and semantic relations on the other hand is statistically significant for AM (argument structure vs. semantic relations: Yate's $\chi^2 = 24.219$, p = 0.0000; semantic categories vs. semantic relations: Yate's $\chi^2 = 23.201$, p = 0.0000), but not for TiMBL.

3.4 Series 2 – fine-tuning the pool of abstract features

In this series we test whether we can reach better predictive accuracies if we select among the abstract features those that are most informative to the classification task and ignore the less informative ones. In order to determine which features are most informative, we pursue two different strategies. In their analysis of the BURSC data, Plag et al. (2006a) have found that only some of the abstract features produced statistically significant effects, whereas others did not. Thus, we selected these features for our experiments. They are

- argument status: only if the second constituent ends in *–er*
- semantic categories: 'N2 is a geographical term', 'N1 is a proper noun'
- semantic relations: 'N2 IS LOCATED AT N1', 'N2 DURING N1', 'N1 IS N2'

In a second experiment, we selected those features that TiMBL, if given all features, finds most informative for the classification task in its training phase. Unlike in AM, in TiMBL the classification of test items is preceded by a training phase, in which the model organises the features that it is trained on in terms of how informative they are for the classification task. Two different informativity measures are computed: Gain Ratio and Information Gain (for a detailed description cf. the TiMBL manual, Daelemans et al. 2004). Since Information Gain feature weighting tended to produce better classification accuracies, we used this measure as a reference. TiMBL was then trained only on the ten most informative features. They are given in table 4, together with the relevant Information Gain values.

Feature	Example	InfoGain value
N1 IS LIKE N2	*crime wave*	0.1021
N2 MAKES N1	*computer company*	0.0368
the compound is a proper noun	*Harvard University*	0.0317
N2 is a thoroughfare	*state road*	0.0206
N1 LOCATED at/in N2	*minority area*	0.0191
N1 is a time	*day care*	0.0164
N1 is a proper noun	*Mapplethorpe controversy*	0.0123
N1 HAS N2	*state inspector*	0.0077
N2 DURING N1	*lifetime*	0.0051
N2 LOCATED at/in N1	*neighborhood school*	0.0048

Table 4. The10 highest IG values in BURSC.

AM was given the same two sets of abstract features. Again, the parameters were set in such a way that all contexts were used, even if the context in the test item occurs in the dataset. Thus, a direct comparison between accuracies of classification in TiMBL and AM is not possible. The results are given in table 5.

As in series 1, the two models are very similar in their performance: Both sets of features lead to very similar predictive accuracies. In both sets left stress is much better predicted than right stress. The two models differ, however, in terms of how their predictions differ from those in series 1. TiMBL does not show any significant improvement compared to the results in series 1. AM, by contrast, is considerably better in series 2 than it was in two of the experiments in series 1 (those based on argument structure and on the semantic categories). The difference in predictive accuracy between the worse of the two tests from series 2 and the semantic categories test from series 1 is just below the level of statistical significance (Yate's $\chi^2 = 3.631$, p = 0.0567); for the argument structure test from series 1, the same difference is significant (Yate's $\chi^2 = 4.044$, p = 0.0443). However, even in AM predictive accuracies in series 2 are still not better than it was in series 1 when given only the semantic relations as features in series 1.

	accuracy		
information source	**overall**	**for left stress**	**for right-stress**
TiMBL (k = 5)			
the features found relevant in Plag et al. 2006a	52.21%	63.69%	40.93%
the 10 features with the highest InfoGain values	54.16%	61.73%	46.70%
AM (every data item is also a member of the test set)			
the features found relevant in Plag et al. 2006a	60.25%	68.44%	52.20%
the 10 features with the highest InfoGain values	57.76%	67.60%	48.35%

Table 5. Classification accuracies for BURSC, series 2.

3.5 Series 3 – the non-abstract features as predictors

We now give AM and TiMBL only the left and the right constituents of the compounds as information source. Recall that the subset of BURSC that we are using here is set up in such a way that every test item will have a constituent family for both its left and its right constituent in the test set.

In TiMBL we carried out three experiments, testing the two constituents in isolation and in combination. In the former case, k was set to 1, in the latter case, it was set to 2, to ensure that the algorithm never used the whole training set as nearest neighbour set. Given that there is no *leave-one-out* procedure in AM, we had AM test only the combination of the two constituents. Since this combination is unique in the dataset, we excluded identical contexts from consideration. Thus, no item was classified on the basis of an identical context. Classification accuracies may now be directly compared between TiMBL and AM. Table 6 summarises the results.

For TiMBL, overall accuracy of prediction is optimal if a combination of left and right constituents are used as information source. In this combination, TiMBL performs better than in any of the tests in which it was trained on abstract features. Statistically, that difference in performance is significant for all combinations of abstract features tested in series 1 and 2 except for the test employing the semantic relations as predictors (for the comparison of the constituents experiment and the best combination of abstract features in series 2: Yate's $\chi^2 = 4.988$, $p = 0.0255$). Recall that for AM, a direct comparison between series 1, 2, and 3 is not possible because in series 1 and 2 the analogical set for each test item included the item itself. Nevertheless, none of these tests is significantly better than the test that used the two constituents as information source, in spite of the fact that the tests with the abstract features had the 'advantage' that the test item was included in the dataset (Yate's $\chi^2 = 0.149$, $p = 0.6996$, comparison of the constituents experiment and the best experiment involving abstract features).

We also note that, if given left and right con-

Information source	**Accuracy**		
	overall	**for left stress**	**for right-stress**
TiMBL (k = 1 / 2)			
left constituent	59.14%	53.35%	64.84%
right constituent	54.71%	46.37%	62.91%
left and right constituent	60.11%	56.7%	63.46%
AM (data items which are identical to a test item are excluded during classification)			
left and right constituent	64.27%	61.73%	66.76%

Table 6. Classification accuracies for BURSC, series 3.

stituents as predictors, both TiMBL and AM are more successful in predicting right stresses than they are at predicting left stresses. This was different in most of the experiments in series 1 and 2, where for the two models' predictions of right stress were generally much worse than predictions of left stress. An exception is again the set of experiments based on semantic relations in series 1, where predictive accuracies for right and left stresses were quite balanced. For TiMBL, the difference between accuracies for right stress predictions in the semantic relations experiment and in the constituents experiment (left and right constituent) is just below the level of significance (Yate's $\chi^2 = 3.574$, p = 0.0586).

A comparison of TiMBL and AM in terms of general predictive accuracy in the constituents test in series 3 shows that the AM experiment yields slightly higher accuracies both in terms of overall accuracy as well as in terms of accuracies for right and left stress. However, the difference is not significant (Yate's $\chi^2 = 2.477$, p = 0.1155).

3.6 Intermediate summary

The experiments described in this section yielded a variety of important insights with respect to the nature of compound stress in English. First of all, although general predictive accuracies are not too impressive, our best TiMBL and AM simulations reached higher predictive accuracies than any of the categorical models proposed in the previous literature. The simulations employing constituent family (series 3) are signifcantly better than the best of the rule-based approaches introduced in section 3.1 (TiMBL: Yate's $\chi^2 = 6.542$, p = 0.0105; AM: Yate's $\chi^2 = 20.269$, p = 0.0000). Nevertheless, we should note that predictive accuracies are far from satisfactory. In this context it is interesting to note that Kunter (2007) has shown for the BURSC data that compound stress involves additional sources of variability, such as token variability within individual types, as well as type variability between different speakers. Whereas this type of variability is a challenge to all kinds of traditional, rule-based models, it is expected under an exemplar-based approach, even if it was not tested in the experimental series presented in this section.

The second thing that we can learn from the results in the previous sections is that compound stress may be computed in an exemplar-based model without assuming that abstract features are involved in the computation. A model that takes into account only left and right constituents is empirically just as adequate as a model that takes into account semantic relations, and significantly better than any other set or combination of abstract features. With respect to an evaluation of the significance of semantic relations, more research is called for. What is interesting, however, is that even this representational abstraction is not necessary, given that the 'non-abstract' constituents can do the same job equally well.

Finally, we have learned that the compound stress data provide inconclusive evidence as to which of the two algorithms, TiMBL or AM, is empirically more adequate. AM consistently performs better than TiMBL, but this difference never reaches statistical significance. Interestingly, the difference is most pronounced if the models are trained on the most relevant, non-abstract features in series 3, whereas they perform more alike in the tests involving the 'less important' sets of abstract features in series 1. This is true in spite of the fact that in series 1 AM has the advantage that every test item is also a member of the data set.

In the next session we report on a parallel set of experiments that were conducted using the CELEX lexical database. We will see that, although the database is very different, the results point into the same direction that we have seen in the BURSC data.

4 Modelling Compound Stress in CELEX

4.1 The data, or: yet another demise of a rule-based approach

As in the BURSC experiments, we used only those compounds from the corpus that have a constituent family. The coding method was the same as for the BURSC data, except for the stress classification, which is simply a given in CELEX. As for the BURSC data, also for the CELEX compound data there exists an in-depth empirical study (Plag et al., 2006b), to which the interested reader is referred for the statistical details concerning the distributional characteristics of determinants of compound stress in the full corpus. The overall distribution of stresses is given in table 7 (N = 2643).

left stress	right stress
2487	156
94.10%	5.90%

Table 7. Distribution of stresses in CELEX.

A major difference between the compound data from BURSC and CELEX lies in the proportion of right stresses. Whereas left and right stresses are almost equally distributed in BURSC, we find only very few right-stressed compounds in CELEX (cf. Plag et al., 2006b for discussion). As in their BURSC study, Plag et al. (2006b) tested the predictive accuracy of previous approaches on the CELEX data. The results mirror those of the BURSC study: None of the syntactic or semantic categories proposed in the literature is successful in adequately predicting the stress distribution in the corpus. Again, this is true no matter whether we implement these categories as predictors in a rule-based model or in a probabilistic, logistic regression model. Nor may predictive accuracy be significantly enhanced if features from the syntactic and the semantic approaches are combined.

For reasons of space, we will focus here only on the rule-based approaches for illustration. Table 8 summarises the predictive accuracies that Plag et al. find for the pertinent approaches, if implemented in a rule-based model.

Note that, due to methodological considerations, the figures are not based on exactly identical datasets. However, all datasets have been shown to pattern alike with respect to the pertinent features. So does the subset to be employed in the present study. As in section 3, we distinguish between overall accuracy on the one hand and accuracy for left and right stresses in the corpus on the other hand.

For argument structure we see the same effect that we saw in BURSC: The model overpredicts right stress; overall predictive accuracy does not reach chance level. For semantic relations and categories, we see an interesting difference between the predictions for BURSC and CELEX. In CELEX the approach is better at predicting left stress than it is at predicting right stress. The opposite was true for the BURSC data. The overall predictive accuracy based on semantic relations and categories is better than for argument structure.

Given the very low proportion of right stresses in the data, the CELEX data provide an extremely difficult information source for an exemplar-based model to learn stress assignment in compounds. In particular, it will be very difficult to learn the distribution of right stress on the basis of the abstract features. That this is indeed the case will be shown in our TiMBL and AM experiments in the next section.

4.2 The TiMBL and AM experiments – parameter settings

As in the BURSC simulations, TiMBL achieved best results with the Jeffrey Divergence metric as a distance metric and Inverse Distance voting among nearest neighbours. Unlike the BURSC simulations, setting the distance space to k = 3 produced better results than a higher k value. Gain Ratio was used to weight features. Again we adjusted k in experiments in which fewer features were used so that the nearest-neighbour set never comprised the whole dataset.

In AM we used the same parameters as in the BURSC settings. As in the BURSC experiments, identical contexts could be excluded from the dataset only in experiments with the two constituents as information source (series 3).

4.3 Series 1 – only abstract features

As in the BURSC simulations, we investigated whether each set of abstract features would serve as an adequate information source for TiMBL and AM. Again we distinguish between overall accuracy, accuracy for compounds that are classified as left-stressed in the corpus, and those that are classified as right-stressed. Due to the differences in parameter settings discussed in section 3.2, accuracies in TiMBL and AM cannot be compared directly.

	predictive accuracies		
a rule system based on...	**overall**	**for left stress**	**for right stress**
argument structure:	49.0%	46.9%	72.4%
semantic relations and categories:	78.7%	85.0%	30.0%

Table 8. Predictive accuracies of categorical, rule-based models for the CELEX data. (fro

	accuracy		
information source	**overall**	**for left stress**	**for right stress**
TiMBL (with k adjusted so that the k-NN set can never be the whole dataset):			
argument structure	94.10%	only left stress predicted	
semantic categories	94.06%	99.96%	0.00%
semantic relations	93.95%	99.46%	1.28%
AM (every test item is also a member of the data set):			
argument structure	94.10%	only left stress predicted	
semantic categories	94.14%	99.36%	0.64%
semantic relations	94.51%	100.00%	7.05%

Table 9. Classification accuracies for CELEX, series 1.

Accuracy of classification is in general much higher than with the BURSC corpus. It is also much higher than that of the rule-based approaches sketched in section 4.1. (for the difference in predictive accuracy between the best rule and the weakest experiment in series 1: Yate's $\chi^2 = 222.311$, $p = 0.0000$). However, it is clear that this effect is largely due to the predominance of left stresses in the training data. Both TiMBL and AM are led to predict left stress for the overwhelming majority of the data, and, because the test set is identical to the training set, high levels of accuracy may be reached just by predicting left stress. By contrast, prediction of right stresses is very weak. What this test series shows very clearly, thus, is that, even if fed the same features, an exemplar-based model may differ considerably in its predictions from a rule-based account.

It is, furthermore, interesting to note that, in spite of all problems, the tendencies that show up in the CELEX experiments are very similar to those we have seen in the BURSC experiments. In both the CELEX and the BURSC experiments, the semantic relations simulations stand out, where both TiMBL and AM manage to predict at least some right stress correctly.

4.4 Series 2 – fine-tuning the pool of abstract features

As in the BURSC experiments, we illustrate the effect that fine-tuning of the set of features given as an information source has on predictive accuracies, with two experiments. In one experiment we chose those features that proved to be significant predictors in Plag et al.'s (2006b) probabilistic model of the CELEX data. These are

- argument structure, but only for compounds ending in –er
- semantic categories: the compound is a proper noun
- semantic relations: N2 has N1, N1 has N2, N2 is made of N1, N1 is like N2, N2 for N1, N2 is located at N1, N2 is named after N1

In the second experiment we used the 10 abstract features with the highest levels of informativity for TiMBL. Since in the CELEX simulations Gain Ratio was the most successful feature weighting measure, we used the Gain Ratio values to select our features. They are given in table 10.

feature	**example**	**Gain Ratio**
N2 IS MADE OF N1	*stone wall*	0.0286
N1 is a proper noun	*India paper*	0.0270
the compound is a proper noun	*labour day*	0.0220
N1 LOCATED at N2	*telephone booth*	0.0183
N2 FOR N1	*writing paper*	0.0169
N2 MAKES N1	*silk worm*	0.0130
N2 DURING N1	*spring tide*	0.0128
argument-head structure	*fire fighter*	0.0062
N2 is a thoroughfare	*service road*	0.0061
N2 NAMED AFTER N1	*guinea fowl*	0.0039

Table 10. The 10 highest GR values in CELEX.

Table 11 summarises the results of our experiments. As in the BURSC experiments, we do not see a considerable improvement of accuracies if we combine abstract features from

	accuracy		
information source	**overall**	**for left stress**	**for right stress**
TiMBL (with k adjusted so that the k-NN set can never be the whole dataset):			
the features found most relevant in Plag et al. (2006b)	94.17%	100.00%	1.28%
the 10 most informative features (TiMBL)	93.98%	99.88%	0.00%
AM (every test item is also a member of the data set):			
the features found to be most relevant in Plag et al. (2006b)	94.17%	99.96%	1.93%
the 10 most informative features (TiMBL)	94.29%	100.00%	3.21%

Table 11. Classification accuracies for CELEX, series 2.

different sets, even if we choose the most informative of them. We now turn to series 3, where we use the non-abstract features, left and right constituent, as information source to feed TiMBL and AM.

4.5 Series 3 – the non-abstract features

In parallel to section 4.5, we feed TiMBL with the left and right constituents both in isolation and in combination. Due to its limitations in parameter settings, AM is given only the combination of features. Since this combination is unique for every compound in the dataset, we could exclude identical contexts from consideration during classification.

In parallel to section 4.5, we feed TiMBL with the left and right constituents both in isolation and in combination. Due to its limiations in parameter settings, AM is given only the combination of features. Since this combination is unique for every compound in the dataset, we could exclude identical contexts from consideration during classification. The results are given in table 12.

Unlike the BURSC experiments, the constituents experiments for CELEX do not yield significantly higher predictive accuracies than those from series 1 and 2 (cf., e.g., the difference between the best experiment from series 3 and the worst experiment from series 1 and 2: Yate's $\chi^2 = 1.893$, $p = 0.1688$). Nevertheless, the constituents experiments differ substantially from the previous series: Thus, they are the only experiments in which both TiMBL and AM manage to predict a considerable number of right stresses. The differences between the number of correct predictions of right stress between the constituents experiment and the experiment with the most predictions of right stresses on the basis of abstract features is highly significant for both TiMBL (Yate's $\chi^2 = 17.394$, p= 0.0000) and AM (Yate's $\chi^2 = 9.932$, $p = 0.0016$). The latter is true, in spite of the fact that the experiment involving 'abstract' features had the advantage that the test item itself was included in the dataset and thus enhanced predictive accuracy.

A comparable level of predictive accuracy for

	accuracy		
information source	**overall**	**for left stress**	**for right stress**
TiMBL (with k adjusted so that the k-NN set can never be the whole dataset):			
left constituent	94.32%	98.87%	21.79%
right constituent	93.27%	97.91%	19.23%
left and right constituent	94.32%	99.32%	19.23%
AM (the test item is never a member of the data set):			
left and right constituent	94.85%	99.56%	19.87%

Table 12. Classification accuracies for CELEX, series 3.

right stress is only found in the rule-based approach based on semantic relations and categories that was described in section 4.1. (30.0% accuracy for right stress). The categorical model, however, reaches its high level of accuracy by overpredicting right stress, which results in a relatively poor predictive accuracy for left stress (85.0%) and, as a consequence, an overall predictive accuracy that is significantly lower than the simulations in series 3 (cf., e.g., Yate's $\chi^2 = 237.595$, $p = 0.0000$, comparison between the TiMBL simulation with two constituents and the rule-based model based on semantic categories and relations).

Finally, as in all other experiments, we do not find a significant difference between the performances of TiMBL and AM in series 3.

4.6 Intermediate summary

In terms of overall predictive accuracy, our exemplar-based models significantly outperform the categorical, rule-based approaches that are proposed in the literature. In this respect, the CELEX experiments very much resemble the BURSC experiments presented in section 3. However, the CELEX experiments differ from the BURSC experiments in that in most simulations, both TiMBL and AM grossly overpredict left stress. Strikingly, the one series of experiments in which both models predict at least a substantial amount of right stresses is the series that uses only the two 'non-abstract' features, the constituents, as an information source. In this respect, the experiments in series 3 are more successful than those employing abstract features (series 1 and 2), although, due to the large proportion of left stresses in the data, these differences do not result in statistically significant differences in terms of overall predictive accuracy.

With respect to the question of whether differences between nearest neighbour selection in TiMBL and analogical set composition in AM in terms show in experiments on compound stress, we are presented with a similar picture as in the BURSC experiments: AM' s levels of overall accuracy are consistently above those reached by TiMBL, but this difference is well below statistical significance.

5 Conclusion and Outlook

The corpus-based empirical study of English compound stress has brought to light a wealth of evidence in favour of an exemplar-based model of compound stress. In terms of predictive accuracy, we have shown that, if trained solely on the right and left constituents of each compound, computational models like TiMBL or AM are more successful in predicting the locus of stress than any categorical rule that has been proposed in the literature, including the Compound Stress Rule. More specifically, the predictive accuracy of – the would-be ill-behaved - right-hand stress improves if the model is trained on the left and right constituents only. Thus, our findings suggest that the level of abstraction needed to compute stress in compounds over stored exemplars does not require abstract syntactic and semantic features. A representation of left and right constituents suffices. This is in line with recent findings concerning the role of consitutent families in compound semantics and compound morphology (e.g. Gagné, 2001, Krott et al., 2001, 2002). It is also paralleled by other recent studies, which argue that word stress assignment is influenced by stress of phonologically similar items in the lexicon (cf., e.g., Daelemans et al. 1994, Eddington 2002 for computational models, Guion et al. 2003 for experimental evidence on English), but contrary to much work in metrical phonology, where it is generally assumed that stress is part of the lexical representation of individual items only in cases of 'exceptional' stress assignment.

Comparing the performance of TiMBL and AM, we see that there are no significant differences in predictive accuracies. However, we also note that AM is consistently a little better than TiMBL in almost all experiments. Interestingly, at least for the BURSC simulations that difference is most pronounced in those experiments in which the model is given features as information source that it finds useful. Conversely, the difference almost disappears if the model is given less useful features (e.g. argument structure, semantic categories), although in these experiments even an exact copy of the test item is included in the dataset. If these observations can be substantiated in more detailed testing, the observed differences between AM and TiMBL may provide additional support for our claim that the the non-abstract constituents are better predictors of compound stress than the pertinent abstract features. Studies comparing TiMBL and AM (e.g. Eddington, 2002, Krott et al., 2002) have mentioned that TiMBL is less influenced by noise or irrelevant features in the data than AM. If abstract features are irrelevant, then we expect AM to have problems if fed with these features, and this is indeed what we find.

Finally, we need to emphasise that research in exemplar-based modelling of compound stress cannot stop here. In this paper we have presented only the first, necessary step. Although TiMBL and AM outperform the categorical and statistical models discussed in Plag et al. (2006a, 2006b), they still produce a considerable amount of classification errors. At this point it is important to note that our TiMBL and AM simulations have neglected two aspects which many exemplar-based approaches to linguistic generalisation consider a vital ingredient of exemplar-based modelling, but whose incorporation is far beyond the scope of this paper.

- the continuous classification of test items
- the influence of frequency factors

In our experiments we have set TiMBL and AM to the task of categorically classifying each test item as either left-stressed or right-stressed. This, however, is an idealisation of the real facts. Using case studies from the BURSC corpus, Kunter (2007) shows that, contrary to prevalent tacit assumptions in most of the pertinent literature, there is inter- as well as intra-speaker variability of stress assignment in compounds. Crucially, compounds differ in terms of the extent to which they exhibit such variability. These facts support an exemplar-based approach to compound stress where different individual realisations of a single compound are assumed to be stored in memory. The question that arises, then, is if we can enhance predictive accuracy by having our test items classified continuously. Whereas both TiMBL and AM provide parameters which allow us to get continuous classification as outputs, however, neither the BURSC nor the CELEX data are suitable to asses token variability in compound stress beyond the case studies analysed in Kunter (2007).

Secondly, we have exclusively relied on types in our experiments. However, in line with much of the literature in exemplar-based modelling, we may expect different types of frequency to play a role in determining the strength of individual exemplars. Among the candidates to be tested are the token frequencies of each compound as well as the family size of the constituents of our compounds. Starting from where this paper ends, it is a task for future research to test whether the integration of both variability and frequency factors into our exemplar-based model of compound stress considerably improves classification.

References

Baayen, H. R. H., R. Piepenbrock, and L. Guilkers (1995), *The CELEX Lexical Database (CD-ROM).* Philadelphia: Linguistic Data Consortium.

Chapman, D., and R. Skousen (2005), Analogical Modeling and Morphological Change: the Case of the Adjectival Negaive Prefix in English. *English Language and Linguistics* 9.2: 333-357.

Chomsky, N., and M. Halle (1968), *The Sound Pattern of English*. New York: Harper & Row.

Daelemans, W., S. Gillis and G. Durieux (1994), The acquisition of Stress: a Data-Oriented Approach. *Computational Linguistics* 20 (3), 421-451.

Daelemans, W., J. van der Sloot, and A. van den Bosch (2004), *TiMBL: Tilburg Memory Based Learner, version 5.1, Reference Guide*. ILK Technical Report 04-02. available from http://ilk.uvt.nl/timbl.

Eddington, D. (2002), A Comparison of Two Analogical Models: Tilburg Memory-Based Learner versus Analogical Modeling. In: Skousen et al. (eds.), *Analogical Modeling*. Amsterdam: John Benjamins, 141-156.

Fudge, E. C. (1984). *English Word-Stress*. London: George Allen & Unwin.

Gagné. C. (2001), Relation and Lexical Priming During the Interpretation of Noun-Noun Combinations. *Journal of Experimental Psychology: Learning, Memory and Cognition* 27: 236-254.

Giegerich, H. (2004), Compound or Phrase? English Noun-Plus-Noun Constructions and the Stress Criterion. *English Language and Linguistics* 8: 1-24.

Guion, S., J.J. Clark, T. Harada, and R. P. Wayland (2003), Factors Affecting Stress Placement for English Nonwords Include Syllabic Structure, Lexical Class, and Stress Patterns of Phonologically Similar Words. *Language and Speech* 46 (4), 403-427.

Krott, A., H. R. Baayen, and R. Schreuder (2001), Analogy in Morphology: Modeling the Choice of Linking Morphemes in Dutch. *Linguistics* 39: 51-93.

Krott, A., R. Schreuder, and H. R. Baayen (2002), Analogical Hierarchy: Exemplar-Based Modeling of Linkers in Dutch Noun-Noun Compounds. In: Skousen et al. (ed.), *Analogical Modeling*. Amsterdam: John Benjamins, 181-206.

Kunter, G. (2007). Within-Speaker and between-Speaker Variation in Compound Stress Assignment. Paper to be presented at the Second International Conference on the Linguistics of Contemporary English ICLCE2, July 2-4, Université de Toulouse.

Kunter, G., and I. Plag (2006). What is Compound Sress? Paper to be presented at the International Congress of Phonetic Sciences, University of Saarbrücken, 6-10 August 2007.

Ladd, D. R. (1984). English Compound Stress. In Gibbon, D. & H. Richter (eds.) *Intonation, accent and rhythm*. Berlin: Mouton de Gruyter, 253-266.

Liberman, M. & R. Sproat (1992). The Stress and Structure of Modified Noun Phrases in English. In Sag, I. A. & A. Szabolcsi (eds.) *Lexical Matters*. Stanford: Center for the Study of Language and Information. 131-181.

Olsen, S. (2000). Compounding and Stress in English: A Closer Look at the Boundary between Morphology and Syntax. *Linguistische Berichte* 181: 55-69.

Olsen, S. (2001). Copulative Compounds: A Closer Look at the Interface between Syntax and Morphology. In Booij, G. E. & J. van Marle (eds.) *Yearbook of Morphology 2000***.** Dordrecht/Boston/London: Kluwer. 279-320.

Ostendorf, M., P. Price, and S. Shattuck-Hufnagel (1996), *Boston University Radio Speech Corpus*. Philadelphia: Linguistic Data Consortium.

Parkinson, D. B. (2002), Running the Perl/C Version of the Analogical Modeling Program, in: Skousen et al. (eds), 365 – 383.

Plag, I. (2006), The Variability of Compound Stress in English: Structural, Semantic, and Analogical Factors. *English Language and Linguistics* 10.1, 143-172.

Plag, I., G. Kunter, S. Lappe, and M. Braun (2006a, submitted for publication) *Testing Hypotheses about Compound Stress Assignment in English: a Corpus-Based Investigation*. a revised version to appear in*: Corpus Linguistics and Linguistic Theory*, 2007.

Plag, I., G. Kunter, S. Lappe, and M. Braun (2006b, submitted for publication) *Modeling Compound Stress in English.*

Sampson, R. (1980). Stress in English N+N Phrases: A Further Complicating Factor. *English Studies* 61: 264-270.

Skousen, R. (1989), *Analogical Modeling of Language*. Dordrecht: Kluwer.

Skousen, R. (2002a), An Overview of Analogical Modeling. In: Skousen et al. (ed.), *Analogical Modeling*. Amsterdam: John Benjamins, 11-26.

Skousen, R. (2002b), Issues in Analogical Modeling. In: Skousen et al. (ed.), *Analogical Modeling*. Amsterdam: John Benjamins, 27-48.

Skousen, R., D. Lonsdale, and D. B. Parkinson (eds, 2002), Analogical Modeling – An Exemplar-Based Approach to Language. Amsterdam: John Benjamins.

Skousen, R., and Th. Standord (2004), *AM:: Parallel.* Brigham Young University. available from http://humanities.byu.edu/am.

Schmerling, S. F. (1971). A Stress Mess. *Studies in the Linguistic Sciences* 1: 52-65.

Spencer, A. (2003). Does English Have Productive Compounding? In Booij. G. E., J. DeCesaris, A. Ralli & S. Scalise (eds.). *Topics in Morphology. Selected Papers from the Third Mediterranean Morphology Meeting (Barcelona, September 20—22, 2001).* Barcelona: Institut Universitari de Lingüística Applicada, Universtitat Pompeu Fabra. 329—341.

Zwicky, A. M. (1986) Forestress and Afterstress. *Ohio State University Working Papers in Linguistics* 32, 46-62.