Arndt-Lappe, Sabine. 2011. Towards an Exemplar-Based Model of Stress in English Noun-Noun Compounds[i]. *Journal of Linguistics* 47, 549-585.

draft version!

Abstract

English noun-noun compounds are traditionally assumed to be left-stressed. However, rightward stress (as in *morning páper* or *Madison Ávenue*) is far from exceptional and recent studies (e.g. Plag 2006, Plag et al. 2007, 2008) have shown that the Compound Stress Rule, or rule-based approaches that make use of argument structure (Giegerich 2004) or semantics (e.g. Fudge 1984), are not able to account satisfactorily for the existing variability.

In this paper it is argued that an exemplar-based approach is superior to traditional rule-based approaches. Two current implementations of exemplar-based algorithms, TiMBL (Daelemans et al. 2007) and AM::Parallel (Skousen & Stanford 2007), are compared to rule models in terms of their ability to predict stress in two large corpora of noun-noun compounds (based on CELEX, Baayen et al. 1995, and the Boston University Radio Speech Corpus, Ostendorf et al. 1996). It is shown that, in terms of their general predictive power, the exemplar-based models clearly outperform the rule models. This is consistently true for both corpora as well as for both algorithms. Furthermore, systematic testing in TiMBL and AM reveals that the reasons for the differences observed between exemplar and rule models mainly lie in their ability to incorporate detailed, non-abstract information. Thus, both TiMBL and AM make crucial

[i] This paper has benefitted greatly from comments and discussions from audiences at the *International Morphology Meeting*, Vienna 2008, and the ISLE conference, Freiburg 2008, where earlier versions were presented, as well as from helpful comments from two anonymous JoL reviewers. Special thanks are due to Ingo Plag for working with me on the predecessor of this paper, to Gero Kunter and Ingo Plag for comments, discussion and critical input, and to Kristina Kösling for helping me with the coding. Needless to say, all remaining errors are my own. Finally, I thank the *Deutsche Forschungsgemeinschaft* (Grant PL-151/5-3) for supporting this research.

use of constituent family information. By contrast, more abstract syntactic and semantic features only play a minor role. The findings present a challenge to traditional rule-based accounts of English compound stress, but are in line with with recent findings concerning the role of consitutent families in related studies of compound semi-regularity (Gagné 2001, Krott et al. 2002).

1. INTRODUCTION

Stress assignment in English compounds has got quite a long tradition in the phonological literature. This is especially true for the generative literature, where one of the most influential and well-known discussions is found in Chomsky & Halle (1968: 15ff., 91ff.). Chomsky & Halle assume that stress in English compounds is governed by the Compound Stress Rule. This rule invariably assigns primary stress to the left constituent of English binary nominal, adjectival, and verbal compounds. The result is that, on the surface level, the stress pattern of these compounds is different from that of syntactic phrases: compounds are left-prominent, phrases are right-prominent. Chomsky & Halle's example to illustrate this distinction is the difference between the compound *bláckboard* and the phrase *black bóard*. An example of a nominal compound is *chémistry laboratory* (Chomsky & Halle 1968: 92).

However, it is also well-known that not all English noun-noun compounds are left-prominent. 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
Óxford Street	Madison Ávenue

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. 2007, 2008) have shown 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. Two current exemplar-based algorithms, TiMBL (Daelemans et al. 2007) and AM (Skousen & Stanford 2007), will be used to investigate whether an exemplar-based approach to compound stress is empirically more adequate than a categorical model. We will use the data from the two Plag et al. (2007, 2008) 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).[1]

The paper is structured as follows. In section 2 we will set the stage, introducing our theoretical assumptions (section 2.1) and the BURSC and CELEX data and coding (section 2.2). After a brief outline of the architecture of the two exemplar-based algorithms (section 3), we will proceed to presenting our findings in sections 4 and 5. The paper ends with a conclusion and directions for 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. We will follow Plag 2006 in labelling them syntactic, semantic, and analogical approaches.

In syntactic approaches, differences in stress among noun-noun combinations are taken to be a reflex of differences in syntactic and/or lexical status. For example, Liberman & Sproat 1992 suggest a strictly categorical approach, arguing that left-stressed NNs are N^0 projections, whereas right-stressed NNs are N^1 projections. They also show, however, that the stress criterion is the only criterion on which this difference in syntactic status can be based – in particular, it does in their view not correlate with any other distinctive criterion, be it syntactic (i.e. interpreting modifiers in N_1 constructions as adjectives) or semantic. Another syntactic approach is found in Giegerich 2004, where he claims that argument-head compounds such as *bookseller* are lexical (i.e. non-syntactic) entities and thus generally left-stressed, whereas modifier-head compounds are generally stressed on the right. Apparent exceptions such as left-stressed *ópera glasses* are the result of the lexicalisation of an originally phrasal structure.

In semantic approaches, right stress is assumed to be triggered by specific semantic relations between left and right constituents (cf. e.g. Sampson 1980, Fudge 1984, Ladd 1984, Olsen 2000, 2001, Spencer 2003). Oft-cited relations are material, temporal, locative, and predicative relations as in *steel brídge*, *summer dréss*, *Boston hárbor*, and *woman-dóctor*. In terms of their theoretical scope, semantic approaches fall into two groups. One group comprises approaches which consider semantic factors to be subsidiary to syntactic factors, motivating exceptions to the (syntactic) Compound Stress Rule. For example, Fudge 1984 and Liberman & Sproat 1992 essentially advocate a syntactic approach, but provide a semantic classification especially of right-stressed constructs, suggesting that right stress is especially frequent in certain semantic classes. By contrast, Olsen 2000 rejects a syntactic approach and claims that semantic factors are the central determinants of compound stress. In general, the syntactic and semantic approaches formulated in the literature differ in terms of the level of generality and formality which they employ in formulating their predictions. Still, it is generally noted that the rules proposed are not exceptionless.

One oft-cited reason for the existence of exceptions is the idea that exceptional stress may be the effect of analogy (e.g. Schmerling 1971, Liberman and Sproat 1992, Giegerich 2004), with compounds having the same right or left constituent sharing the same stress pattern. Standard examples are street names, which are stressed on the left constituent if the right constituent is *street*, but stressed on the right constituent if the right constituent is *avenue* or *lane* (as in *Óxford Street* vs. *Oxford Ávenue, Oxford Láne*). Crucially, in all these approaches the role of analogy is subsidiary in nature – it is assumed that analogy 'kicks in' in those cases where the alleged regular determinants – be they syntactic or semantic – fail to predict the attested output pattern.

Syntactic, semantic, and analogical approaches have recently been empirically tested against large amounts of data by Plag et al. 2007, 2008, Plag 2010, and Plag & Kunter 2010.[2] Plag et al. 2007 investigates the relevance of syntactic, semantic, and analogical determinants of compound stress assignment in several thousands of noun-

noun compounds extracted from the CELEX corpus. CELEX (Baayen et al. 1995) contains mostly dictionary, i.e. lexicalised, data.

Plag et al. 2008 is concerned with syntactic and semantic factors in compounds extracted from a very different corpus, the Boston University Radio Speech corpus (Ostendorf et al. 1996). BURSC contains audio recordings of radio news texts. Plag 2010 and Plag & Kunter 2010 use data from both BURSC and CELEX as well as from a third corpus to investigate the effects of different types of constituent family information. Methodologically, all studies mentioned use stochastic models to test whether the stress distribution observed in the corpora can be predicted with the help of the syntactic and semantic, and, in the case of Plag 2010 and Plag & Kunter 2010, analogical determinants mentioned in the literature. In addition, for the CELEX data Plag et al. 2007 also tests the analogical approach with the help of an exemplar-based model (using the TiMBL algorithm).

An important conclusion to be drawn from this group of studies is that 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. Thus, there is no set of predictor variables derived from the literature that is able to convincingly predict stress category in the data. This does not mean, however, that the categories suggested in the literature are irrelevant. A considerable number of them does turn out to function as statistically significant predictors in a stochastic regression model. Overall, however, we still find a relatively large amount of unexplained variation. Such variation is found among compound types as well as among tokens of the same type. Furthermore, the extent of variation as well as the effects found are highly robust. Thus, Plag et al. 2007, 2008 investigate compounds from two very different corpora (BURSC and CELEX), but find largely the same effects and the same degree of variation.

On a theoretical level, then, these findings raise the question of which kind of grammatical theory would most adequately account for compound stress. We have seen that, traditionally, compound stress has been accounted for in a categorical, rule-based framework, where syntactic or semantic factors work as categorical rules, and where exceptions to the rule have a different status in the grammatical system (as irregular) than 'well-behaved' (regular) items. From an empirical perspective (Plag et al. 2007, 2008, Plag & Kunter 2010, Plag 2010), such an approach faces two serious challenges: First of all, stress in English compounds is decidedly noncategorical. Secondly, analogy seems to be more widespread than is expected by a framework in which analogical cases can only be modelled as exceptional cases.

An alternative approach is to assume a grammatical theory where the distinction between regular and irregular items is inexistent because linguistic categorisation is inherently gradient and emergent. One such theory is exemplar theory (cf., e.g., Bybee 2001, Pierrehumbert 2001, Gahl & Yu eds. 2006, Bod & Cochran eds. 2007 for linguistic applications). A key feature of exemplar-based approaches to linguistic categorisation is that there is no categorical rule mechanism; instead, linguistic generalisations emerge from similarity-based computation, comparing new input to exemplars of previously encountered linguistic experiences that are stored in the lexicon. As a consequence, regularities as found in the data are noncategorical and emergent: Since classification of new input crucially depends on the stored properties of similar exemplars in memory, these may be different for each new input.

A second feature that makes exemplar theory attractive for a model of compound stress is the fact that it offers a principled account of analogical effects. Although the exact nature of how exemplars are to be represented is still debated in the literature, it is clear that exemplar representation requires a lesser degree of abstraction than the symbolic features that linguistic rules traditionally take as their arguments. We find a growing number of evidence for the important role of such less abstract representations not only in occasional remarks on analogical effects in compound stress assignment (e.g. Schmerling 1971: 56, Ladd 1984), but also – and, more recently – in studies of other aspects of derivational morphology and compounding (e.g. Gagné 2001, Krott et al. 2002, Chapman & Skousen 2005). In exemplar theory, then, the often-noted analogical effects that we find in stress assignment between compounds with the same constituents receive a very natural interpretation: Information about individual compound constituents is readily available to the algorithm that computes similarity and, hence, to stress assignment.

The present study presents a first step towards an exemplar-based theory of English compound stress. Specifically, we will test whether the two features of exemplar theory just discussed – non-categoriality and the integration of less abstract features – help to produce a better model of English compound stress – 'better' in the sense of meeting the empirical facts more accurately than a rule-based model. In order to do this, we will set two exemplar-based algorithms, TiMBL (Daelemans et al. 2007) and AM (Skousen & Stanford 2007), to the task and compare their performance with the performance of a categorical approach.

It is important to note at this point that, focussing on non-categoriality and non-abstract features, the present study does not integrate all features of a full-fledged exemplar-based model of English compound stress. In particular, we will use type corpora as our data and will not attempt to incorporate token-related information (such as, for example, frequency information).[3]

2.2 The coding

As an empirical basis, we use the data from the Plag et al. studies, as well as their coding of the data. In order to test the effects of argument structure, semantics and analogy, Plag et al. first extracted all NN structures from BURSC and CELEX.[4] 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 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, and excluded all those types that did not have such a family. The rationale behind this choice was that we wanted to make sure that all features that could act as potential determinants of stress – argument structure, semantics, and analogy – were available for all 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 feature: argument structure, semantic categories, and semantic relations. The argument structure feature codes the syntactic argument status of the first constituent with respect to the head noun. It thus applies only to compounds in which the head is a deverbal noun. The semantic categories comprise a pool of those that have most often been cited in the pertinent literature (e.g. Fudge 1984: 144ff., Gussenhoven & Broeders 1981, Zwicky 1986, Liberman & Sproat 1992). The semantic relations comprise a set of presumably language-independent predicates most of which have been proposed by Levi (1978; cf. Plag et al. 2008 for discussion).[5] 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[6]	Example
argument-head	*computer maker*
modifier-head	*truck accident*

(3) Semantic property of constituent or compound

Property	Example
N1 refers to a period or point in time	*day care*
N2 is a geographical term	*bay area*

N2 is a type of thoroughfare	*state road*
The compound is a proper noun	*Harvard University*
N1 is a proper noun	*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. Finally, each compound in the corpus was coded for its stress category, left or right. For CELEX we relied on the classification of stress information as given in the corpus. Here stress is treated as an invariant property of the compound lemmata. For BURSC, the stress category of each compound token was determined acoustically by two independent raters who both hold a degree in English linguistics. There were 1,805 different tokens, distributed over 722 types. Raters agreed on a stress category in 1,395 (i.e. 77.3%) of the cases. We then used only those tokens for which the two raters agreed in their judgements, assigning the stress category of the majority of tokens to the pertinent compound type.[7] In cases of a tie, the compound was excluded from the corpus. Of course this procedure, combined with the requirement outlined above that every compound should have a constituent family among the data, led to a further reduction of the dataset. The final BURSC dataset to be used in the analysis is comprised of 536 compound types. A sample of a coded item is given in (5).

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

constituents	left	*clinic*
	right	*worker*

argument structure	Is N1 an argument of N2?	*no*
semantic categories	...	*no*
semantic relations	N2 FOR N1	*yes*
	N2 LOCATED at N1	*yes*
	...	*no*
stress category		*left*

3. The models – TiMBL and AM

In the present paper we will pursue a quite radical and computationally testable analogical approach to compound stress and compare this approach to a more traditional, rule-based approach. By the latter we mean a rule system that is categorical in the sense that it does not predict variation for any given input, and that it requires the ordering of subrules to be fixed. [8]

We will tease apart the different aspects that distinguish these two sets of approaches (analogical and rule-based) and see in how far they contribute to the performance of our models on empirical data. Specifically, we will focus on three important aspects:

- symbolic, rule-based models vs. exemplar-based models: which approach is empirically more adequate, measured in terms of its predictive power for the corpus data?

- abstract vs. non-abstract features in TiMBL and AM: how abstract do features have to be for a successful computation of compound stress?
- nearest neighbours vs. Analogical Set: do differences between TiMBL and AM in terms of their computational architecture have an effect on their performance?

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 same data.

The second aspect concerns the question of which features are relevant for the representation of generalisations about compound stress. We will compare compare predictions based on relatively abstract information about the semantic, syntactic, morphological, or semantosyntactic status of a compound with predictions based on less abstract constituent family information. For ease of reference, we will from now on refer to the latter as 'non-abstract' features and to the former as 'abstract' features, acknowledging, of course, that 'abstractness' is a matter of degree here.

Both TiMBL and AM are exemplar-based algorithms that 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. 2007 for more details). The classification of a new item is extrapolated from similar exemplars that are explicitly 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, or by weights which are able to distinguish between different values for a single feature. 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 given item is answered for every single new input on an individual basis. 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 feature values (termed contexts) and determines whether or not these contexts behave in a homogeneous way with respect to the target category (i.e., in our case, stress assignment). Only exemplars with homogeneous contexts will then be included in the Analogical Set (cf. Skousen 1989, 2002a, 2002b for details). Note, however, that 'homogeneous' does not necessarily mean 'deterministic'. The definition of homogeneity of a context crucially hinges on the fact that in AM related contexts are ordered hierarchically in terms of their level of generality (more general contexts act as supracontexts of more specific subcontexts). A homogeneous context is a context whose subcontexts behave identically (cf., e.g., Skousen 2002a for details). Another interesting consequence of the ordering of context relations in AM is that contexts which are more general and, hence, less similar to the context of the test item may also end up in the Analogical Set.

Apart from the large processing demands that computation of Analogical Sets in AM entails, an interesting difference between TiMBL and AM 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, which 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 so-called leave-one-out procedure. In AM, we used the same data set for both training and testing. Every context was made unique by introducing an additional dummy variable which had a different value for every compound in the database, but no relevance to the classification task. We then used the available option in AM that classification of test items is to be based only on contexts in the training file which are not identical to the test item.[9]

4. THE PREDICTIVE POWER OF RULE-BASED APPROACHES

Tables 1 and 2 provide an overview of the actual distribution of stresses in the corpora.

left stress	right stress	total
360	176	536
67.16%	32.84%	100.00%

Table 1. Distribution of stresses in the BURSC data.

left stress	right stress	total
2487	156	2643
94.10%	5.90%	100.00%

Table 2. Distribution of stresses in the CELEX data.

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, 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 (2008).

As is shown in Table 2, a major difference between the compound data from BURSC and CELEX lies in the proportion of right stresses. This difference is exactly parallel to the differences in the distribution of stress categories in the much larger compound corpora extracted from BURSC and CELEX that were used in Plag et al. 2007, 2008. As discussed in detail in Plag et al. 2007, the high proportion of left stresses in CELEX is most likely due to the fact that very many of the CELEX compounds derive from dictionaries, i.e. are highly lexicalised data. By contrast, the distribution of stress categories in the BURSC data comes very close to what is generally estimated to be the proportion of right and left stressed compounds in the literature (cf. Plag et al. 2008 for discussion). Even though, however, the CELEX data contain only some 6% of right stresses, it is clear that even the CELEX data pose a challenge to the Compound Stress Rule.

For both BURSC and CELEX empirical studies are available which test the empirical adequacy of structural and semantic approaches to compound stress (Plag et al. 2008 and Plag et al. 2007, respectively). Specifically, these studies 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., Zwicky 1986, Olsen 2000), as predictors of compound stress. For both BURSC and CELEX, 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 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. In what follows, we will show that a rule-based account along traditional lines is unsatisfactory also for the subset of the BURSC and CELEX compounds used in the present studies.[10]

According to a categorical stress rule based on argument structure, we should find in our data that compounds in which N1 is an argument of N2 are left-stressed, whereas all other compounds are right-stressed. The contingency tables in Tables 3 and 4 show how argument status of the first constituent w.r.t. the head of the compound (argument or non-argument) is distributed among left- and right-stressed compounds in our datasets. Cells for which the categorical rule just cited makes the right predictions are shaded.

		N1 is a(n)...		
		argument	non-argument	total
observed stress	left:	77	283	360
	right:	30	146	176
	total:	107	429	536

Table 3. observed stress vs. argument status of N1 w.r.t. N2 in the BURSC data (N = 536)

		N1 is a(n)...		
		argument	non-argument	total
observed stress	left:	358	2129	2487
	right:	8	148	156
	total:	366	2277	2643

Table 4. observed stress vs. argument status of N1 w.r.t. N2 in the CELEX data (N = 2643)

For the BURSC data, a rule based on argument structure would make the right predictions for only 77 + 146 = 223 of 536 compounds. This amounts to an overall predictive accuracy below chance level (41.60%). For the CELEX data, predictive accuracy is even much worse (19.14%). We also note in Tables 3 and 4 that the argument structure rule is better at predicting right stresses than it is at predicting left stresses. Thus, only 21.39% (BURSC) and 14.39% (CELEX) of all left-stressed compounds are predicted to be left-stressed, but 82.95% (BURSC) and 94.87% (CELEX) of all right-stressed compounds are predicted to be right-stressed.

These findings may seem surprising, but note that they are fully in line with what Plag et al. 2007, 2008 found for a much larger subset of BURSC and CELEX compounds. Although significant effects of argument structure were detected in the data, argument structure turned out to be the weakest of all predictors tested in that study, to the extent that it could be eliminated from the final, stochastic models without a significant loss of predictive accuracy.[11]

At this point a note is in order about how we will measure the predictive power of rules and models in this paper. The distributions in Tables 3 and 4 provide a nice illustration of why it is not always a good idea to measure predictive accuracy solely in terms of what we have called 'overall predictive accuracy' above, i.e. the percentage of items predicted correctly. First of all, we note that predictive accuracy can be very different for different target categories (i.e., in our case, left and right stress). Secondly, given that the distribution of observed stresses in both datasets is skewed towards left stress, overall predictive accuracies will reflect that uneven distribution. To take an extreme case as an example, imagine a (hypothetical) model that predicts only left stresses. For the BURSC data, this model would reach an overall predictive accuracy of 67.16%, for CELEX it would even reach a predictive accuracy of 94.1%. Conversely, a model that predicts only right stresses would reach a predictive accuracy of only 32.84% for BURSC and 5.9% for CELEX. Thus, in order to measure the predictive power of rules and models, we require a measure that somehow takes into account the uneven distribution of target categories and tells us how well our model predicts each of them, on average.

One measure commonly used in information theory is the F-score (cf., e.g., Daelemans et al. 2007: 31f., Daelemans & Bosch 2005: 48ff.; 78f. for details). F-scores are computed for each target category, in our case for left and right stress. Let's take the target category 'left stress' in the rule model exemplified in Table 3 as an example to see how F-scores work. The F-score for left stress is the harmonic mean of two ratios, both of which can assume a value between 0.0 and 1.0. The first ratio, termed 'recall', is what we have called 'accuracy for left stress' above. It gives us the number of items for which the model correctly predicts left stress divided by the number of items which are indeed left-stressed. It thus tells us how well the model is able to find left-stressed items in the corpus. According to Table 3, then, the rule model's recall for left stress is $\frac{77}{360}$ = 0.2139.

The second ratio is termed 'precision', and it gives us the number of items for which the model correctly predicts left stress divided by the number of items for which the model predicts left stress. Given the distributions in Table 3, the rule model's precision for left stress is $\frac{77}{107} = 0.7196$. The F-score for left stress is the harmonic mean of precision and recall:

$$F_{left} = \frac{2 * 0.7196 * 0.2139}{0.7196 + 0.2139} = 0.3298$$

For the target category 'right stress' in Table 3 we can likewise compute the F-score, which is 0.4826. In order to obtain one final measure of the predictive power of our model, we can now compute an average F-score for left and right stress:

$$F_{av} = \frac{0.3298 * 360 + 0.4826 * 176}{536} = 0.3800$$

Note that this averaged F-score is a weighted mean; each F-score is weighted proportionally according to the observed stresses of the relevant category in the dataset. This is termed 'micro-averaging' (cf. Daelemans et al. 2007: 32).[12] From now on, we will document the predictive power of the rules and models discussed in this paper using micro-averaged F-scores. For reasons of conceptual simplicity, predictive accuracies (i.e. recall scores) will also be given, but not discussed.

Table 5 compares the performance of the argument-structure rule for the BURSC and CELEX data.

		BURSC	CELEX
F-score			
	overall, micro-averaged	0.38	0.24
	left	0.33	0.25
	right	0.48	0.12
recall			
	overall, micro-averaged	0.42	0.19
	left	0.21	0.14
	right	0.83	0.95

Table 5. predictive accuracies and F-scores for the argument-structure rule

We see that, if measured in terms of F-scores, the rule's alleged good performance for the prediction of right stress turns out to be less convincing. For CELEX, F-scores for right stress are even worse than those for left stress. This is due to the fact that for the CELEX data, even though recall of right stress is quite high (0.95), precision is very low (0.06).

The second rule-based approach to be tested here uses the semantic characteristics of compounds as predictors. Here the question arises which of the many semantic features coded should be included in a semantic rule system. Our categorical approach will be based on a subset of the features that have been mentioned in semantic approaches that were discussed in section 2.1. The subset comprises those semantic features that have been found to show significant stress effects in the Plag et al. studies (Plag et al. 2007 for CELEX, Plag et al. 2008 for BURSC). By excluding features whose predictive power for the data at hand is poor, we thus ensure that our categorical semantic model is as powerful as it can get.

For the BURSC data, a semantic rule based on the semantic properties and relations would assign right stress if a compound exhibits at least one of the following semantic characteristics:

semantic properties

- N1 refers to a period or point in time
- N2 is a geographical term
- N1 and N2 form a proper noun
- N1 is a proper noun
- N1 and N2 form a left-headed compound

semantic relations

- N2 IS MADE OF N1
- N1 IS N2
- N2 IS LOCATED AT N1
- N2 DURING N1
- N2 IS NAMED AFTER N1

Elsewhere, the rule would assign left stress.

For the CELEX data, the semantic rule based on Plag et al.'s 2007 findings would assign right stress essentially to the same categories as the rule for the BURSC data, with the exception of the two semantic categories 'N2 is a geographical term' and 'N1 and N2 form a left-headed compound', which turned out to be insignificant in the Plag et al. 2007 study.

Tables 6 and 7 show how the stress categories predicted by the semantic rule are distributed among the observed stresses in our subsets of the BURSC and CELEX compounds. As in the previous tables, cells representing correct predictions are shaded.

		semantics, tending towards...		
		left stress	right stress	total
observed stress	left:	252	108	360
	right:	112	64	176
	total:	364	172	536

Table 6. observed stress vs. semantics in the BURSC data (N = 536)

		semantics, tending towards...		
		left stress	right stress	total
observed stress	left:	1705	782	2487
	right:	66	90	156
	total:	1771	872	2643

Table 7. observed stress vs. semantics in the CELEX data (N = 2,643)

The performance of the semantic rule is assessed in Table 8.

		BURSC	CELEX
F-score			
	overall, micro-averaged	0.59	0.76
	left	0.70	0.80
	right	0.36	0.18
recall			
	overall, micro-averaged	0.59	0.68
	left	0.70	0.69
	right	0.36	0.58

Table 8. predictive accuracies and F-scores for the semantic rule

For the semantic rule, averaged F-scores are substantially higher than those for the argument structure rule. Still, a major weakness of the semantic rule is the prediction of right stress. This is especially true for the CELEX data.[13]

An obvious question that emerges from our findings so far is whether the predictive power of the approaches based on argument structure and semantics could be improved if they joined forces. As is clear from the tables, however, we cannot expect much improvement under a rule-based paradigm. The reason is that the semantic rule predicts right stress for a subset of those compounds for which the argument-structure rule predicts right stress (i.e. modifier-head compounds which exhibit the semantic characteristics specified above). As a consequence, a combined rule can never reach better accuracies for right stress than the semantic rule. As we have seen above, however, prediction of right stress is the major weakness of that rule.

In what follows we will document how the two exemplar-based models, TiMBL and AM, are able to predict compound stress in our data. We will deal with the two corpora, BURSC and CELEX, in turn.

5. Exemplar-based modelling of the Boston data

5.1 Step 1 – all features as information source

Table 9 provides an overview of the performance of TiMBL and AM if the algorithms are given all coded features as information source. The relevant parameter settings are given in appendix A. The table gives both F-scores and, in parentheses, recall (i.e. predictive accuracy).

information source	**predictive power**		
	F-score (recall), averaged	**F-score (recall) for left stress**	**F-score (recall) for right stress**
TiMBL (with *k* = 3):			
all features	0.75 (0.76)	0.83 (0.88)	0.59 (0.53)
AM:			
all features	0.73 (0.74)	0.82 (0.85)	0.56 (0.51)

Table 9. Classification accuracies for BURSC, all features used as predictors

We see that accuracy rates for TiMBL are slightly higher (about 2%) than those for AM. What is much more striking, however, is that both TiMBL and AM clearly outperform the rule-based models discussed above if they are given all features as information source. Whereas averaged F-scores for the rule-based approaches are both 0.38 (argument-structure rule) and 0.59 (semantic rule), the exemplar-based models reach F-scores between 0.73 and 0.75. Also in terms of F-scores for left and right stress, both TiMBL and AM are superior to the rule-based models.

In order to better understand where this discrepancy in predictive power could come from, we will take a closer look at those aspects of the architecture of the two exemplar-based models which make them fundamentally different from rule-based approaches.

The first obvious difference lies in the nature of the algorithms. Rule-based models capture the syntactic or semantic predictors of compound stress in terms of both necessary and sufficient conditions which obligatorily trigger rule application. In TiMBL and AM, stress assignment is extrapolated from the stress category of the majority of the most similar items that are stored in memory. To take an example, the compound *Massachusetts company* would (correctly) be assigned right stress in both rule-based and exemplar-based approaches, but on different grounds. For the rule-based approach the only thing that is relevant is that there is a locative relation between N1 and N2, which would provide a sufficient condition for the compound to be assigned right stress.

In TiMBL and AM, by contrast, the most similar compounds that are stored in memory are *Massachusetts market* and *Massachusetts veteran*. They are similar to *Massachusetts company* in that their codings all share the value for the locative semantics ('yes'), the first constituent ('Massachusetts'), and all other semantic and syntactic categories ('no'). The stress categories of *Massachusetts market* and *Massachusetts veteran* will thus be the categories that have the greatest say in determining the stress category of *Massachusetts company*. Other locative compounds in memory will be less influential or not influential at all.

Thus, for example, the item *Massachusetts administrator* is less similar to *Massachusetts company* than the two other compounds. The reason is that, even though *Massachusetts company* and *Massachusetts administrator* both have locative semantics and the same first constituent, the coding of semantic relation between constituents in *Massachusetts administrator* has a 'yes' value not only for a locative relation, but also for a purpose relation (N2 FOR N1). Other locative compounds will be even further away from *Massachusetts company*. An example is *bank window*, which shares the locative semantics with *Massachusetts company*, but has a different first and second constituent and, in addition, a 'yes' value for a possessive semantic relation between N1 and N2 (N1 HAS N2). In TiMBL, the question of how similar an item must be in order to be incorporated into the nearest-neighbour set depends on how the distance space *k* is set. If *k* is set to 1, only *Massachusetts market* and *Massachusetts veteran* will form the nearest neighbour set of *Massachusetts company*. If *k* is set to 2, the model will additionally incorporate compounds like *Massachusetts administrator*. Models with *k* = 3 will incorporate even less similar items like *bank window,* and so forth.

In AM the question of how similar an item must be in order to be incorporated into the Analogical Set is resolved differently. The algorithm checks all contexts defined by the feature combination of a given target (*Massachusetts company*) for homogeneity in terms of stress classification. It will then take all homogeneous contexts into account in the classification task, whereby those contexts most similar to the target context will have the greatest say in the voting procedure.

Compared to the rule-based model, then, stress classification in TiMBL and AM is more local in the sense that it is usually based not on one single global categorial system, but on a restricted set of the most similar items in the lexicon. In the case of a model that incorporates all coded categories as an information source (as in Table 9), predictive power in TiMBL was optimal if *k* was set to 3.

The second major difference between the rule-based model discussed in this paper and models like TiMBL and AM is intricately linked to the characteristics just discussed. Thus, it is clear that in both rule-based and exemplar-based models classification accuracies will depend on the type of features that the models are given as an information source. In the rule-based models these features are of a relatively abstract nature, i.e. semantic or syntactic features or a combination of both. For exemplar-based models it is perfectly natural to take also less abstract features as information source; in fact, many of the existing exemplar-based theories assume that detailed phonetic representation is stored and available for grammatical classification. In the pertinent models tested in this paper, orthographic form of the compound constituents acts as a proxy for detailed levels of representation of that kind. Thus, our exemplar-based models include the compound constituents themselves (N1, N2) as information source.

Finally, rule-based models differ from TiMBL and AM in that they postulate that rules must be ordered. As a consequence, every determinant feature has the same relevance for every data item. In exemplar-based models this may be different, but to different degrees. In this respect TiMBL and AM do not behave alike. During the training phase, TiMBL uses an entropy-based algorithm to assign informational weight to the features processed. This information weight is then used to compute similarity, such that features with higher weights are given preference over features with lower weights.

Table 10 provides an overview of the information weight scores assigned to the features used in our TiMBL experiment. For reasons of readability, only the six highest scores are shown.

Weighting	**Gain Ratio**
left constituent (N1)	0.0659
semantic relation: N2 CAUSES N1	0.0646
right constituent (N2)	0.0619
semantic relation: N2 IS LIKE N1	0.0605
semantic relation: N1 USES N2	0.0547
semantic relation: N1 HAS N2	0.0343

Table 10. Gain Ratio values for BURSC

Several things are noteworthy here. First of all we see that, with the exception of one semantic relation (N2 CAUSES N1), highest Gain Ratio weights are assigned to the non-abstract features 'left constituent' and 'right constituent'. It is these features that TiMBL regards as most important in the computation of similarity among compounds in the corpus.

With respect to the four semantic relations that are listed in Table 10, it is important to see that three of them – N2 CAUSES N1, N2 IS LIKE N1, N1 USES N2 – hardly ever occur in the corpus; items coded as 'yes' for any of these relations range between one (N1 USES N2) and three (N2 CAUSES N1). By contrast, N2 HAS N1, ranked sixth in terms of Gain Ratio values, is pertinent for 149 compounds in the corpus.

The second thing that is interesting is that, among the two non-abstract features, the left constituent has a higher Gain Ratio score than the right constituent. This runs counter to what is often assumed about analogical stress assignment in English compounds, i.e. that similarity is computed on the basis of the head of the compound (cf, e.g., *street* vs. *avenue* compounds; but cf., e.g., Schmerling 1971 and Liberman & Sproat 1992 for approaches which consider both constituent families to be relevant for analogical effects).

Finally, it is important to note that the relation between Gain Ratio scores of the features and the question which compounds end up in the nearest-neighbour set of a given compound in the corpus is a complex one. For example, we have seen above that for *Massachusetts company* the most similar compounds are those that share both the first constituent and the semantics with the target. However, in an experiment in which the *k* value is set to 3 (as is the case in the experiment represented in Table 9), the nearest neighbour set also includes compounds that do not share any of the constituents with the target compound (such as *bank window*, as discussed above). For other compounds, this is not the case. An interesting example is *minority area*. Here the most similar item is *lead area*, which does not have the same first constituent as its target, but the same second constituent and a similar semantics (a locative relation between N1 and

N2). Further away in the distance space (*k* = 2) we find the neighbour *minority woman*, which shares the first constituent with the target, but differs from the target in terms of its semantics. Finally, in the space defined by *k* = 3 we find *minority voter*, whose relation to *minority area* is similar to that of *minority woman*, with the additional difference that the head is a deverbal noun. Thus, in an experiment in which *k* is restricted to 3, all nearest neighbours of *lead area* share either the first or the second constituent with the target. The way in which nearest neighbour sets are selected therefore explains why combinations of different features (e.g. syntactic, morphological, semantic) affect classification accuracy in TiMBL very differently from rule-based approaches.

In terms of the selection of relevant neighbours, AM's behaviour is even more radical than that of TiMBL. Table 11 provides an overview of the compounds in the Analogical Set for our example *Massachusetts company*.

compound	**stress category**	**no. of pointers**	**relevance in vote**
massachusetts market	stress_right	80740352	18.554%
massachusetts veteran	stress_right	80740352	18.554%
massachusetts community	stress_right	54525952	12.530%
massachusetts republican	stress_right	54525952	12.530%
massachusetts school	stress_right	54525952	12.530%
massachusetts town	stress_right	54525952	12.530%
massachusetts road	stress_right	29360128	6.747%
california company	stress_left	18874368	4.337%
polaroid company	stress_left	4194304	0.964%
massachusetts tax	stress_right	3145728	0.723%

Table 11. Analogical Set for *Massachusetts company*

As in TiMBL, the most similar items turn out to be *Massachusetts market* and *Massachusetts veteran*, i.e. items which share with the target the first constituent and all other abstract features. However, these exemplars are chosen to be most relevant on different grounds. In AM these two compounds are in the Analogical Set because they are, like all other compounds in the Analogical Set, exponents of a context that behaves in a homogeneous way with respect to stress classification (all items with this context are right-stressed). Additionally, they are more relevant than other compounds in the Analogical Set (i.e. they get 18.554% of the vote each) because among all homogeneous contexts, they represent the most specific context (i.e. they are most similar to the target). Less relevant than *Massachusetts market* and *Massachusetts veteran*, but still getting 12.53% of the vote each, we find compounds like *Massachusetts community*. They share with the target the first constituent and all other coded features, except for the fact that, in addition to a locative semantic relation, they also have a 'yes' value for a possessive semantic relation (N1 HAS N2). Now recall that at this point the Analogical Set in AM deviates from the nearest-neighbour set in TiMBL, which had *Massachusetts administrator* at a distance of *k*=2 instead, a compound that has been coded for an additional purpose relation rather than a possessive relation. This is because, on the basis of its global weighting of features for their informativity (Gain Ratio), TiMBL makes the principled assumption that a deviation from the semantic relation of the target weighs less heavily if a purpose relation is affected than if a possessive relation is affected. This is assumed for all compounds. AM does not make such global assumptions. Instead, it finds that for the context defined by the features of the target *Massachusetts company*, the subcontext that is characterised by a different second constituent and an additional possessive relation of the type N1 HAS N2 behaves homogeneously, whereas the subcontext defined by an additional purpose relation does not. This decision is, however, specific to *Massachusetts company* and may be different for other targets. Another difference between the Analogical Set in Table 11 and the nearest-neighbour set in TiMBL is that in the former we also find two compounds (*California company*, *Polaroid company*) that share with the target the right constituent, but not the left constituent.

To sum up, we have seen in this section that the two exemplar-based algorithms are superior to rule-based approaches if they are given the richest feature combination conceivable as information source. We have also seen that this could be due to basically two fundamental differences between the approaches: either the nature of the algorithm, where the exemplar models often act more locally and are somewhat more flexible in terms of relevant features, or the nature of the information source, where exemplar models may take into account more specific, non-abstract information (i.e. N1 and N2). In what follows we will investigate these two alternative explanations. If differences between rules and exemplar-based approaches are due to the nature of the algorithm, we should find that TiMBL and AM will outperform rule models also if they are given only the abstract features as information source. If differences reside in the possibility to include constituent family information, then we should find that TiMBL and AM will perform well also without the abstract features. We will set TiMBL and AM to the two tasks in turn.

5.2 Step 2 – the abstract features as information source

Table 12 provides an overview of classification accuracies of four sets of models: those which test syntactic and semantic features in isolation, one which tests them in combination, and one which uses only those features that have been found to be relevant in Plag et al.'s 2008 study.

information source	**predictive power**		
	F-score (recall), averaged	**F-score (recall) for left stress**	**F-score (recall) for right stress**
TiMBL (with *k* adjusted so that the *k*-NN set can never be the whole dataset):			
argument structure	0.54 (0.67)	0.80 (1.0)	n.a.[a]
semantics	0.59 (0.68)	0.80 (0.96)	0.17 (0.10)
argument structure and semantics	0.58 (0.61)	0.73 (0.80)	0.25 (0.20)
the semantic features found to be relevant in Plag et al. (2008)	0.62 (0.69)	0.80 (0.95)	0.24 (0.15)
AM:			
argument structure	0.54 (0.67)	0.80 (1.0)	n.a.
semantics	0.61 (0.68)	0.79 (1.0)	0.24 (0.16)
argument structure and semantics	0.60 (0.64)	0.76 (0.86)	0.27 (0.20)
the semantic features found to be relevant in Plag et al. 2008	0.57 (0.66)	0.79 (0.95)	0.12 (0.07)

[a] In this experiment, all items are predicted to be left-stressed, no right stress is predicted. As a consequence, the F-score for right stress is not defined, as the computation involves division by 0. We follow the convention applied in TiMBL and use 0 as the limit of the harmonic mean as the F-score instead. The same procedure will be applied in all other experiments in which no right stress is predicted.

Table 12. Classification accuracies for BURSC, abstract features.

Unlike the syntactic rule, TiMBL and AM predict only left stress to occur if they base classification only on argument structure information. If they are given semantic information, performance is about similar to that of the corresponding rule model (with an F-score of 0.59 and 0.61 for the exemplar-based models as compared to an F-score of 0.59 for the rule-based model). The table also shows that a combination of syntactic and semantic information sources does not considerably enhance predictive power; on the contrary, F-scores for combinations of syntactic and semantic features are often lower than those for either syntactic or semantic features in isolation. If we use a subset of the most relevant semantic features, TiMBL's performance is able to profit from this smaller number of predictors (the F-score rises to 0.62) whereas performance of AM drops to an F-score of about 0.57. What is particularly noteworthy, however, is that, no matter which combination of abstract features we use, performance of TiMBL and AM never even nearly reaches the level it reached in our first experiment, where we included non-abstract N1 and N2 as predictors, and where F-scores were 0.75 (TiMBL) and 0.73 (AM), respectively. This suggests that it is precisely this ability to include non-abstract features as predictors that makes the exemplar-based models superior to classical rule-based approaches. To see whether this is true, we now turn to the third step in our experimental series.

5.3 Step 3 – the non-abstract features as information source

We now give AM and TiMBL only the left and the right constituents of the compounds as information source. Recall that in the subset of BURSC that we are using every test item has a constituent family for both its left and its right constituent. We carried out three experiments, testing the two constituents in isolation and in combination. Table 13 provides an overview of the results.

information source	**predictive power**		
	F-score (recall), averaged	**F-score (recall) for left stress**	**F-score (recall) for right stress**
TiMBL (with *k* adjusted so that the *k*-NN set can never be the whole dataset):			
left constituent	0.75 (0.76)	0.83 (0.86)	0.60 (0.55)
right constituent	0.68 (0.70)	0.79 (0.84)	0.47 (0.40)
left and right constituent	0.76 (0.77)	0.84 (0.91)	0.59 (0.50)
AM:			
left constituent	0.76 (0.77)	0.83 (0.86)	0.62 (0.57)
right constituent	0.73 (0.74)	0.81 (0.84)	0.57 (0.53)
left and right constituent	0.80 (0.80)	0.86 (0.90)	0.67 (0.61)

Table 13. Classification accuracies for BURSC, non-abstract features.

For both TiMBL and AM, F-scores are optimal if a combination of left and right constituents is used as information source. In this combination, the models perform better than in any of the tests in which they were trained on abstract features. AM even reaches a higher F-score than in the first experimental series, where the model was trained on all features including the left and right constituent (compare F-scores of 0.73 and 0.80, respectively). This is not the case for TiMBL, where predictive power is about the same in the two series (about 0.75).

We also note that, if given left and right constituents as predictors, AM is more successful in predicting right stresses than it was in all previous test series.

In a final test series, we tried whether predictive power on the basis of constituent family could be enhanced by adding a selection of abstract features to the constituent family information. For practical reasons, we conducted this test series with TiMBL only, because in TiMBL it is possible to determine the most powerful feature combination with the help of the features' Gain Ratio scores. Thus, a series of experiments was run in which for each experiment one feature was added as an information source, starting with only the feature with the highest Gain Ratio score and proceeding in a stepwise fashion to features with lower Gain Ratio scores. The results of the experiment with the most powerful combination (measured in terms of its averaged F-score) are given in Table 14. To facilitate a comparison, the table also repeats the results of the experiments where all features were used as information source and where only constituent family was used as information source.

information source	**predictive power**		
	F-score (recall), averaged	**F-score (recall) for left stress**	**F-score (recall) for right stress**
20 features with the highest	0.77 (0.78)	0.85 (0.91)	0.61 (0.52)

Gain Ratio values			
for comparison:			
only constituent family	0.76 (0.77)	0.86 (0.90)	0.59 (0.50)
all features	0.75 (0.76)	0.83 (0.88)	0.59 (0.53)

Table 14. Optimising feature combinations in TiMBL

Predictive power was optimal if TiMBL was given the 20 features with the highest Gain Ratio scores as information source. These include the constituent families of N1 and N2 as well as a set of syntactic and semantic features.[14] Note, however, that the model is not able to benefit greatly from the inclusion of abstract features in addition to constituent family information. The difference between the optimal feature combination and the experiment that is based on constituent family only is just about no greater than 0.01.

5.4 Intermediate summary

We have seen that both TiMBL and AM are far better at predicting compound stress assignment in the Boston corpus than rule-based approaches. This is partly due to fundamental differences in the architecture of rule-based and exemplar-based approaches. However, the main factor that makes exemplar-based models more powerful than rule-based approaches is that the former are able to accomodate constituent family information as predictor. Figure 1 provides an overview of the averaged F-scores achieved in our TiMBL and AM experiments, ordered by the results of the TiMBL experiments.

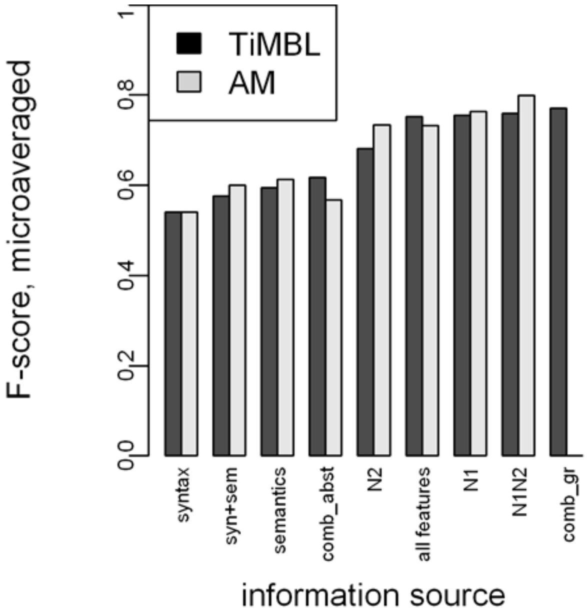

Figure 1. Results of TiMBL and AM models, BURSC

Experiments which are based solely on abstract features ('syntax', 'semantics', 'syn+sem', i.e. the combination of all abstract features, and 'comb_abst', i.e. the combination of abstract features found relevant in Plag et al. 2008) reach F-scores around 0.6 the most. If either a combination of abstract features and constituent family ('all feat' and 'comb_gr', i.e. the combination of features with the highest Gain Ratio values) or constituent family information alone ('N2', 'N1', 'N1N2') are used as information sources, F-scores rise by about 10-20%. Among the experiments including constituent family information, the selection of features chosen does not make much of a difference for TiMBL, as long as the experiment is not solely based on N2 constituent families. This is remarkable, given that N2 constituent family is the factor that has been most widely discussed in the literature on analogical effects in compound stress assignment (e.g. Schmerling 1971: 56, Ladd 1984).

For AM, differences are somewhat more pronounced, highest F-scores are reached in experiments based on N1 and N2 constituent families. If we compare TiMBL and AM, we see that despite the differences in the models' general architecture, results are remarkably similar. With the exception of two experiments, AM performance is slightly above that of TiMBL.

In the next section we will consider 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.

6. MODELLING COMPOUND STRESS IN CELEX

6.1 Step 1 – all features as information source

In the first set of experiments we again use all features, abstract and non-abstract, as information source. Table 15 summarises the results.

information source	**predictive power**		
	F-score (recall), averaged	**F-score (recall) for left stress**	**F-score (recall) for right stress**
TiMBL:			
all features	0.93 (0.94)	0.97 (0.99)	0.31 (0.21)
AM:			
all features	0.93 (0.94)	0.97 (0.99)	0.29 (0.15)

Table 15: Classification accuracies for CELEX, all features used as predictors

As in the BURSC simulations, TiMBL and AM produce very similar results which clearly outperform the rule-based approaches that were discussed at the beginning of this paper. For the rules, micro-averaged F-scores in Tables 5 and 8 ranged between 0.24 (the syntactic rule) and 0.76 (the semantic rule). By contrast, both TiMBL and AM reach F-scores of more than 0.93 when fed with all features. Table 15 also reveals a major weakness of the exemplar-based models: the asymmetry in predictive power of left and right stress, which we already saw in some of the BURSC simulations, is much more pronounced in the CELEX experiments. Thus, whereas predictive power is very strong for left stress, it is very weak for right stress. Notice, however, that this problem was also present in the stronger of the two rule-based approaches (the semantic rule: compare F-scores of 0.80 for left stress and 0.18 for right stress).

In what follows abstract and non-abstract features will again be compared in terms of how they influence the predictive power of our models.

6.2 Step 2 – the abstract features as information source

Table 16 presents the results of the set of experiments in which TiMBL and AM use only abstract features as information source.

information source	**predictive power**		
	F-score (recall), averaged	**F-score (recall) for left stress**	**F-score (recall) for right stress**
TiMBL (with *k* adjusted so that the *k*-NN set can never be the whole dataset):			
argument structure	0.91 (0.94)	0.97 (1.0)	n.a.[15]
semantics	0.91 (0.94)	0.97 (1.0)	0.02 (0.01)
argument structure and semantics	0.91 (0.94)	0.97 (1.0)	0.02 (0.01)
the semantic features found to be relevant in Plag et al. 2007	0.91 (0.94)	0.97 (1.0)	n.a.
AM:			
argument structure	0.91 (0.94)	0.97 (1.0)	n.a.
semantics	0.91 (0.94)	0.97 (1.0)	0.02 (0.01)
argument structure and semantics	0.91 (0.94)	0.97 (1.0)	0.02 (0.01)
the semantic features found to be relevant in Plag et al. 2007	0.91 (0.94)	0.97 (1.0)	0.03 (0.01)

Table 16. Classification accuracies for CELEX, abstract features used as predictors

As can be seen from Table 16, averaged F-scores are slightly below those reached when both abstract and non-abstract feature sets are given as information source (compare F-scores of around 0.93 in Table 15 with F-scores of around 0.91 in Table 16). However, when we compare the two sets of experiments in terms of how well they predict our two stress categories, left and right stress, we see that the small differences in averaged F-scores gloss over a rather big difference between the two sets of experiments. Whereas differences in F-scores for left stress are rather small (about 0.01), the two sets of experiments vastly differ in terms of how well right stress is predicted. When all features are given as information source, F-scores for right stress are about 0.31 (TiMBL) and 0.29 (AM), respectively. When, however, only abstract features are given, F-scores for right stress fall to values between 0.00 and 0.03. In the experiments using argument structure as information source, no right stress is predicted at all, neither by TiMBL nor by AM. In the experiment in which those features are used which have been found to have a significant influence on stress assignment in Plag et al.'s 2007 study of CELEX compounds, TiMBL does predict some right stress, but not for compounds which actually are right-stressed. In sum, the relatively high averaged F-scores that are reached on the basis of abstract features only (i.e. those in Table 16) are to a large extent an effect of our averaging method (micro-averaging, cf. above), in which more weight is given to target categories that are strongly represented in the corpus. For a corpus like the CELEX compounds, where actual stress distribution is strongly biassed towards left stress (compare 94.1% left vs. 5.9% right stresses), this means that differences between models in their predictions of right stress do not come out as pronounced effects in averaged F-scores.

6.3 Step 3 – the non-abstract features as information source

As in our BURSC experiments, we now feed TiMBL and AM with the left and right constituents both in isolation and in combination. The results are given in Table 17.

information source	**predictive power**		
	overall	**for left stress**	**for right stress**
TiMBL (with *k*=1 or *k*=2, respectively):			
N1	0.93 (0.94)	0.97 (0.99)	0.31 (0.22)
N2	0.92 (0.93)	0.96 (0.98)	0.25 (0.19)
N1 and N2	0.93 (0.94)	0.97 (0.99)	0.23 (0.15)
AM:			
N1	0.94 (0.95)	0.97 (0.99)	0.36 (0.26)
N2	0.94 (0.94)	0.97 (0.98)	0.41 (0.34)
N1 and N2	0.93 (0.95)	0.97 (1.0)	0.31 (0.20)

Table 17. Classification accuracies for CELEX, non-abstract features used as predictors

As in the BURSC experiments, classification accuracies in this set of simulations are above those in the experiments based on the abstract features. If fed with only N1 as information source, TiMBL reaches about the same classification accuracy as it did in the experiment in which all features were given (compare averaged F-scores of about 0.93 in Tables 15 and 17). If given N2 or both N1 and N2, there is a very slight drop in predictive power. AM, by contrast, makes slightly better predictions if given only the non-abstract features as information source than if given all features (cf. Table 15). Again, however, we find interesting differences between models when we look at how well they perform predicting right stress. Whereas TiMBL is best at predicting right stress if given only N1 or if given all features, the optimal information source for the prediction of right stresses in AM is the non-abstract features only, with highest F-scores for right stress in the experiment that uses solely N2 as information source. Here F-scores for right stress in AM experiments are generally higher than those in TiMBL experiments.

In a final step, we test whether a combination of the most informative features from both the abstract and the non-abstract set would optimise predictive power in TiMBL beyond the levels reached in previous experiments. We use the same procedure as in the BURSC experiments (cf. section 5.3) and restrict ourselves to TiMBL as the computational algorithm.

information source	**predictive power**		
	overall	**for left stress**	**for right stress**
19 features with the highest Gain Ratio values	0.93 (0.95)	0.97 (0.99)	0.33 (0.22)
for comparison:			
only N1	0.93 (0.94)	0.97 (0.99)	0.31 (0.22)
all features	0.93 (0.94)	0.97 (0.99)	0.31 (0.21)

Table 18. Optimising feature combinations in TiMBL

Predictive power was optimal when the 19 features with the highest Gain Ratio scores were used as information source. We see, however, that not much is to be gained from the optimisation procedure. Thus, the overall F-score hardly rises above the level it reaches when only N1 is used as information source.[16] Note, however, that for right stress improvement is slightly better, with an F-score difference of some 0.02.

6.4 Intermediate summary

Figure 2 summarises the experiments discussed in this section. For better illustration, however, notice that we deviate here from our usual method of averaging F-scores. Thus, up until now F-scores were micro-averaged, i.e. F-scores for left and right stress were combined in such a way that weighing of F-scores reflects biasses in the distribution of actual left and right stresses in the data. In figure 2 now we show macro-averaged F-scores, which are simply calculated as the harmonic mean of left and right F-scores. Thus, F-scores for left and right stress are treated as equals in the formula. For the CELEX data, this has the advantage that differences between models stand out visually more clearly than if we use micro-averaged F-scores. Note, however, that the choice of the averaging method does not affect the general outcome of the comparison of models made in this paper.

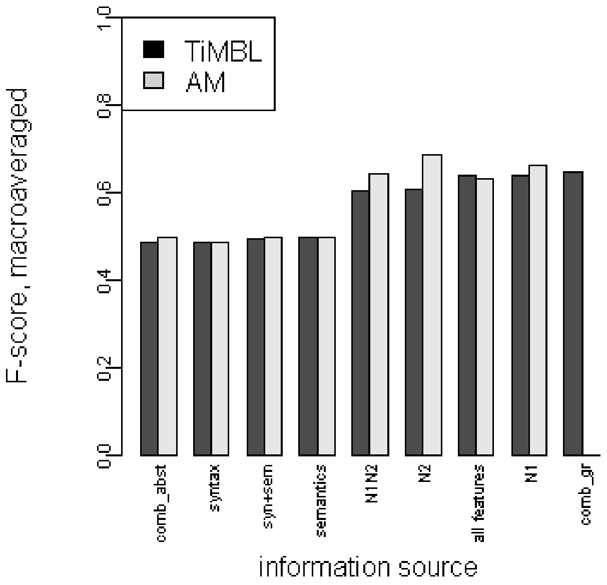

Figure 2. Results of TiMBL and AM models, CELEX

In the CELEX experiments we find that, although the corpus is very different from the Boston corpus in terms of its internal composition, results almost exactly mirror those of the Boston experiments. Thus, both TiMBL and AM are superior to rule-based models in predicting stress distributions. Furthermore, the most useful information source turns out to be constituent family information, which may or may not be combined with other abstract features. Crucially, models using only abstract features as information source are significantly worse than those including non-abstract features, no matter which combination of abstract features we use.

The one point in which the CELEX experiments differ substantially from the Boston experiments is 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.

7. CONCLUSION

The present study of English compound stress has brought to light strong evidence in favour of an exemplar-based model of compound stress. In terms of predictive power, 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 right-hand stress improves if the model is trained on the left and right constituents. 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 constituent 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 large differences in overall performance. However, we also note that AM is consistently a little better than TiMBL in most experiments.

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 categorical models, they still produce a considerable amount of classification errors. At this point it is important to note that our TiMBL and AM simulations do not incorporate token information (cf. section 2.1 for discussion). The incorporation of such information is desirable, because it would allow us to model two types of effect that are expectable in an exemplar-based model of compound stress assignment: within-type variability of stress and token frequency effects. For the existence of within-type variability, we have evidence from case studies from the BURSC corpus. Kunter 2009 shows that, contrary to prevalent tacit assumptions in most of the pertinent literature, there is inter- as well as intraspeaker 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. Unfortunately, neither the BURSC nor the CELEX data are suitable to assess token variability in compound stress beyond the case studies analysed in Kunter 2009.

Token frequency effects are expected in an exemplar-based model because, among stored exemplars, we would expect highly frequent types to exert a stronger influence in classification than less frequent types. The issue, however, how exactly token frequency is to be implemented in an exemplar-based model in general is still under debate in the literature (cf. esp. Bybee 2001: 50ff., 138ff. for an overview discussion), where proposals differ in terms of how they model the relation between type and token representations. In view of the theoretical debate as well as the practical considerations discussed in section 2.1, we leave it to future research to see whether the incorporation of tokens will yield a stronger exemplar-based model than the implementation introduced in the present paper.

Appendix A – Parameter settings in TiMBL and AM for the BURSC experiments

In TiMBL, highest classification accuracies were reached if similarity was computed using a simple overlap metric for the left and right compound constituent and the Jeffrey Divergence metric for all other features. Feature weighting was based on Gain Ratio values. Using a distance-weighted similarity metric for left and right constituents 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 $k = 3$. 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 simple majority voting.

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).

Appendix B – Parameter settings in TiMBL and AM for the CELEX experiments

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. The distance space *k* was set to 3, except in experiments in which fewer than three features were used as information source. Gain Ratio was used to weigh features.

In AM we used the same parameters as in the BURSC experiments. We also included a dummy variable that is unique for each item where appropriate.

REFERENCES

Baayen, Harald R., Richard Piepenbrock & Leon Guilkers. 1995. *The CELEX Lexical Database (CD-ROM)*. Philadelphia: Linguistic Data Consortium.

Bod, Rens & David Cochran (eds.). 2007. *Proceedings of the workshop on exemplar-based models in language acquisition and use*. Dublin: ESSLLI 2007.

Boersma, Paul & Bruce Hayes. 2001. Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry* 32, 45–86.

Booij, Geert, Janet DeCesaris, Angela Ralli & Sergio Scalise (eds.). 2003. *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.

Booij, Geert & Jaap van Marle (eds.). 2001. *Yearbook of Morphology 2000*. Dordrecht: Kluwer.

Bybee, Joan. 2001. *Phonology and language use*. Cambridge: Cambridge University Press.

Bybee, Joan & Paul Hopper (eds.). 2001. *Frequency effects and the emergence of lexical structure*. Amsterdam / Philadelphia: John Benjamins.

Chapman, Don & Royal Skousen. 2005. Analogical modeling and morphological change: the case of the adjectival negative prefix in English. *English Language and Linguistics* 9(2), 333–357.

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

Daelemans, Walter, Steven Gillis & Gert Durieux. 1994. The acquisition of stress: a data-oriented approach. *Computational Linguistics* 20(3), 421–451.

Daelemans, Walter & Antal van den Bosch. 2005. Memory-based language processing. Cambridge: Cambridge University Press.

Daelemans, Walter, Jakub Zavrel, Ko van der Sloot & Antal van den Bosch. 2007. *TiMBL: Tilburg Memory Based Learner, version 6.0, Reference Guide*. LK Technical Report 04-02. Tilburg: ILK.

Eddington, David. 2002. A comparison of two analogical models: Tilburg Memory-Based Learner versus Analogical Modeling. In Royal Skousen, Deryle Lonsdale & Dilworth B. Parkinson (eds.), *Analogical Modeling*, 141–156. Amsterdam / Philadelphia: John Benjamins.

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

Gagné, Christina L. 2001. Relation and lexical priming during the interpretation of noun-noun combinations. *Journal of Experimental Psychology: Learning, Memory and Cognition* 27, 236–254.

Gahl, Susanne & Alan C. L. Yu (eds.). 2006. Special issue on exemplar-based models in linguistics. *The Linguistic Review* 23.

Gibbon, Dafydd & Helmut Richter (eds.). 1984. *Intonation, accent and rhythm*. Berlin: Mouton de Gruyter.

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

Guion, Susan G., J. J. Clark, Tetsuo Harada & Ratree 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.

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

Krott, Andrea, Harald R. Baayen & Rob Schreuder. 2001. Analogy in morphology: Modeling the choice of linking morphemes in Dutch. *Linguistics* 39, 51–93.

Krott, Andrea, Rob Schreuder & Harald R. Baayen. 2002. Analogical hierarchy: exemplar-based modeling of linkers in Dutch noun-noun compounds. In Royal Skousen, Deryle Lonsdale & Dilworth B. Parkinson (eds.), *Analogical Modeling,* 181–206. Amsterdam / Philadelphia: John Benjamins.

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

Kunter, Gero. 2010. *Perception of prominence patterns in English nominal compounds*. Paper presented at Speech Prosody 2010, Satellite Workshop on Prosodic Prominence: Perceptual and Automatic Identification. May 10, 2010. Chicago.

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

Levi, Judith N. 1978. *The syntax and semantics of complex nominals*. New York: Academic Press.

Liberman, Mark Y. & Richard Sproat. 1992. The stress and structure of modified noun phrases in English. In Ivan A. Sag & Anna Szabolcsi (eds.), *Lexical Matters*, 131–181. Stanford: CSLI Publications.

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

Olsen, Susan. 2001. Copulative compounds: a closer look at the interface between syntax and morphology. In Geert Booij & Jaap van Marle (eds.), *Yearbook of*

Morphology 2000, 279–320. Dordrecht: Kluwer.

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

Parkinson, Dilworth B. 2002. Running the Perl/C version of the Analogical Modeling Program. In Royal Skousen, Deryle Lonsdale & Dilworth B. Parkinson (eds.), *Analogical Modeling*, 365–383. Amsterdam / Philadelphia: John Benjamins.

Pierrehumbert, Janet. 2001. Exemplar dynamics: word frequency, lenition, and contrast. In Joan Bybee & Paul Hopper (eds.), *Frequency effects and the emergence of lexical structure*, 137–157. Amsterdam / Philadelphia: John Benjamins.

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 Sprachwissenschaft* 29(2).

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

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.

Sag, Ivan A. & Anna Szabolcsi (eds.). 1992. *Lexical matters*. Stanford: CSLI Publications.

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

Schmerling, Susan F. 1971. A stress mess. *Studies in the Linguistic Sciences* 1, 52–65.

Skousen, Royal. 1989. *Analogical Modeling of language*. Dordrecht: Kluwer.

Skousen, Royal. 2002a. An overview of Analogical Modeling. In Royal Skousen, Deryle Lonsdale & Dilworth B. Parkinson (eds.), *Analogical Modeling*, 11–26. Amsterdam / Philadelphia: John Benjamins.

Skousen, Royal. 2002b. Issues in Analogical Modeling. In Royal Skousen, Deryle Lonsdale & Dilworth B. Parkinson (eds.), *Analogical Modeling*, 27–48. Amsterdam / Philadelphia: John Benjamins.

Skousen, Royal, Deryle Lonsdale & Dilworth B. Parkinson (eds.). 2002. *Analogical Modeling*. Amsterdam / Philadelphia: John Benjamins.

Skousen, Royal & Thereon Stanford. 2007. *AM: Parallel*. Provo, UT: Brigham Young University.

Spencer, Andrew. 2003. Does English have productive compounding? In Geert Booij, Janet DeCesaris, Angela Ralli & Sergio Scalise (eds.), *Topics in Morphology. Selected Papers from the Third Mediterranean Morphology Meeting (Barcelona, September 20–22, 2001)*, 329–341. Barcelona: Institut Universitari de Lingüística Applicada, Universtitat Pompeu Fabra.

Zwicky, Arnold. 1986. Forestress and afterstress. *Ohio State University Working Papers in Linguistics* 32. Columbus: Ohio State University.

Author's Address: *Fachbereich Sprach-, Literatur- und Medienwissenschaften*

Universität Siegen, Germany

Adolf-Reichwein-Str. 2

D-57068 Siegen

arndt-lappe@anglistik.uni-siegen.de

FOOTNOTES

[1] In this paper we will use the term 'compounds' as a convenient label to refer to noun-noun 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.

[2] For a discussion of how these approaches were operationalised in terms of a coding of the corpus data, cf. section 2.3 below.

[3] This move constitutes a departure from the focus of much exemplar-based modelling as employed in the phonological literature (as, e.g., in Pierrehumbert 2001, the papers in Gahl & Yu eds. 2006). One crucial assumption in exemplar-based models of linguistic categorisation is that memory traces of individual linguistic experiences, i.e. of tokens, are part of exemplar representations stored in memory (cf., e.g., Bybee 2001: 50ff., 138ff. for theoretical discussion). Many exemplar-based approaches in phonology focus on the role of acoustic detail, as it is remembered in individual tokens, in phonological categorisation. The focus of the present study is, however, different, as it concerns remembered characteristics of exemplars on a different level of analysis: syntax, semantics, and constituent family. Given that tokens of the same compound type do not differ with respect to these characteristics, the use of tokens would have the potential to obscure their effects, or at least to render them unconvincing. Thus, in an exemplar-based model of compound stress, stress assignment is based on stress as stored in the most similar exemplars in the lexicon. If, then, we use tokens instead of types, every entry in this lexicon is a token. Tokens of the same type, however, will be more similar to any given new compound than tokens of a different type. As a consequence, classification by the algorithm will almost always be based on exemplars that represent the same compound type, and it will be impossible to tease apart effects of tokens of the same type and other types of effect.

[4] While this is straightforward for the lexical database CELEX, all texts from 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.

[5] The categories coded comprise a pool of all semantic relations and categories that have been mentioned as relevant in the literature cited. Since we used Levi's set as a basis for the coding of semantic relations, our coded semantic relations include more relations than mentioned in the stress literature. The rationale behind this was that we left it to the empirical facts to answer the question which subset of categories and relations coded would turn out to be significant, rather than using a preselected set of categories.

[6] Apart from plain argument structure, Plag et al. (2007, 2008) also coded the morphological category of the deverbal head noun of argument-head compounds. Thus, e.g., *maker* in *computer maker* was coded as –er derivative, whereas *reform* in *budget reform* was coded as a case of conversion. In their statistical models, they found an argument structure effect for only one the morphological categories, –er derivatives, in BURSC as well as in CELEX. In the present investigation we will, however, ignore the morphological coding. The reason is that, given the rather small numbers of pertinent data, differentiating between morphological categories does not do much to enhance predictive accuracies. The point to be made in this paper can be made without introducing yet another set of coded categories.

[7] As a reviewer points out, inter-rater agreement in rating stress category is an interesting issue for research in its own right. In the present study, excluding ambiguous items seemed the cleanest practical solution. Not only is it unclear which status ambiguous tokens should have in the computation of a stress category for the pertinent compound type; it has also recently been shown (cf. Kunter 2010) that rating stress category in compounds is subject to a host of complexities which arise from both the proficiency of the rater and the acoustic properties of the compound. The interaction of these factors is not yet fully understood.

[8] It is important to note that our interest in the exact formulation of the rules in question is restricted to these two properties. We chose rules and not, say, optimality-theoretic constraints for ease of exposition. Note, however, that, in their traditional form, both rule and constraint systems are categorical in the sense outlined above. Of course there are rule-based or constraint-based models on the market which are geared more specifically to dealing with variation (such as, e.g., Stochastic Optimality Theory, as described in Boersma & Hayes 2001). How they fare in dealing with compound stress variation is not in the focus of the present paper.

[9] Thanks to David Eddington, p.c., for drawing my attention to this possibility.

[10] Note that the scope of the present study is limited to a comparison of rule-based and exemplar-based models of the pertinent distributions. A detailed analysis and discussion of the pertinent distributions of syntactic and semantic categories in BURSC and CELEX has been published elsewhere (cf. Plag et al. 2007 for CELEX and Plag et al. 2008 for BURSC).

[11] Note that the question of the empirical accuracy of the argument structure rule is more complex than the findings in Tables 3 and 4 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 the tables only appears because lexicalisation has not been taken into account. Plag et al. (2007) tested the lexicalisation hypothesis, using type 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 would-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.

[12] A second possibility of computing an averaged F-Score is to simply use the arithmetic mean of the two F-scores for left and right stress, respectively (termed macro-averaging). Conceptually, this means that the same weight is given to the two target categories, irrespective of the uneven distribution that we find in the training data. In this paper we will use micro-averaging as the general method, mainly because this method is intuitively closer to the way in which predictive accuracies are traditionally computed. Note, however, that nothing substantial hinges on the choice of the averaging method: The general differences between models to be discussed in this paper remain, no matter which averaging method we choose.

[13] Notice that for the BURSC data, F-scores closely resemble recall scores. This is an effect of the fact that mispredictions are quite balanced (cf. Table 6 above). Contra to what Table 8 seems to suggest, F- and recall scores are not fully identical. The fact that numbers are identical in the table is an effect of rounding.

[14] Specifically, these are: N2 CAUSES N1, N2 HAS N1, N1 HAS N2, N2 MAKES N1, N1 MAKES N2, N2 IS MADE OF N1, N2 USES N1, N1 USES N2, N1 IS LIKE N2, N2 LOCATED at/in N1, N1 LOCATED at/in N2, N2 IS NAMED AFTER N1, the compound has an argument-head structure, N1 is a period or point in time, N2 is a thoroughfare, N1 is a proper noun, the compound is a proper noun, N2 has a deverbal suffix.

[15] In this experiment, all items are predicted to be left-stressed, no right stress is predicted. As a consequence, the F-score for right stress is not defined, as the computation involves division by 0. We follow the convention applied in TiMBL and use 0 as the limit of the harmonic mean as the F-score instead. The same procedure will be applied in all other experiments in which no right stress is predicted.

[16] There is indeed a slight increase in the averaged F-score (by 0.0012), which is not visible in Table 18.