DRAFT VERSION, May 2013, submitted for publication

Analogy and the nature of linguistic generalisation: locality, generality, and variability in English compound stress[1]

Sabine Arndt-Lappe
Heinrich-Heine-Universität Düsseldorf

Melanie J. Bell
Anglia Ruskin University

1. Introduction

Analogical reasoning, the ability to recognise and reproduce patterns of similarity between entities and events, is all pervasive in human cognition and has even been proposed as the most fundamental difference between human cognition and that of other intelligent species (Gentner & Smith 2012). Although most research on analogical reasoning has involved reasoning with awareness, there is also evidence that human subjects use the mechanism unconsciously (e.g. Day & Gentner 2007), and this gives rise to the possibility that analogical reasoning could be the unifying process underlying human grammars.

In linguistics, the idea that analogy constitutes a core mechanism of human grammar has a long tradition, especially in classical and neo-grammarian approaches (e.g. Becker 1990, Anttila 2003 for a summary). Within generative approaches, however, analogy has come to be interpreted as exactly the opposite, namely as the source of exceptions to a rule-based core. A central assumption of such approaches is that the language experienced by speakers is insufficient to act as a template for the great diversity of language produced, and that some part of language must therefore be already present in speakers' minds at birth. It is postulated that there is a central, innate grammar, conceptualised as a system of abstract symbolic rules, that precedes exposure to language and is therefore independent of actual linguistic instances (Chomsky 1975, discussed in Blevins & Blevins 2009; cf. Itkonen 2005). A logical consequence of this view is that whenever there is clear evidence of linguistic generalisation based on specific exemplars, and hence whenever there is clear evidence of analogy, such generalisations are interpreted as having a different status from the grammar proper. An example of this is the group of 'dual-route' proposals that emerged in the 1990s to account for

[1] Acknowledgements to be added.

the productivity of irregular past tense formation in English. Prasada & Pinker (1993), for example, argue that irregular past tense forms can be created from appropriate present tense forms by analogy with irregular present-past pairs stored in memory, whereas regular forms are generated by 'a default suffix concatenation process capable of operating on any verb', i.e. by a rule system in the sense of generative grammar (see also Pinker & Prince 1988, Pinker & Prince 1991, Pinker 1991, Kim et al. 1991, Marcus et al. 1992).

Unlike generative approaches to linguistics, usage-based frameworks have embraced the idea that memory of individual instances can give rise to linguistic generalisation (Bybee 2010 and references therein). Yet despite a common emphasis on the potential relevance of specific exemplars, such models do not agree on the status of analogy. The reason seems to be a disagreement about the role of abstraction in generalisation. Whereas construction-based frameworks generally attribute much importance to schematisation mechanisms that constitute abstractions from individual instances (e.g. Booij 2010), radical exemplar-based approaches, which include most analogical models, conceptualise linguistic generalisation as emerging directly and probabilistically from the set of exemplars in memory (e.g. Derwing & Skousen 1989, Skousen 1989, Daelemans et al. 1999, Daelemans & van den Bosch 2005, Gahl & Yu 2006, van den Bosch & Daelemans 2013). Note, however, that the issue of abstraction also arises here, because a degree of symbolism is involved in the lexical representation of exemplars. For example, it would be an abstraction to represent an exemplar as a string of phonemes, since the phoneme is an abstraction from actual sound waves. It seems incontravertible that some level of abstraction occurs in natural language, otherwise we would not be able to recognise as such the same word spoken on different occasions, let alone by different speakers. But at present, it is unclear what kind of evidence could distinguish between abstraction at the level of lexical representation and abstraction in the form of higher order schemas, or if indeed there is any principled difference.

In the present paper, we set out to argue in favour of an analogy-based view of linguistic generalisation. We employ the term 'linguistic generalisation' as a more neutral variant of the term 'grammar' and the term 'analogy-based' to refer to a single-route model in which memory of individual instances forms the basis of all linguistic generalisation. We remain agnostic as to whether generalisation is also mediated by higher-level schemas as advocated in construction theory: the models presented in this paper neither rule out such schemas nor provide positive evidence for their existence. The key point we wish to make is that an analogical view of linguistic generalisation provides an empirically adequate account of two properties of language that in our view have convincingly emerged from recent quantitative work in linguistics. These are:

1. linguistic generalisations involve various levels of generality;
2. linguistic generalisations are often probabilistic.

Using a computational simulation of a testbed phenomenon, we will show how these properties emerge in a straightforward way from an analogical theory.

Computational implementations of analogical theories, such as TiMBL (Daelemans et al. 1999 et seq.) and AM::Parallel (Skousen 1989, Skousen et al. 2002 et seq., Skousen & Stanford 2007), have successfully been used to replicate many aspects of human language: morphological and phonological, syntactic and semantic, synchronic and diachronic (e.g. Skousen et al. 2002, Daelemans & Van den Bosch 2005, Chapman & Skousen 2005, Keuleers 2008, Krott et al. 2007, Krott 2009, Soskuthy 2013, Van den Bosch & Daelemans 2013). Our particular contribution is to show not only that an analogical model is successful, but also how this success is achieved. We show that a single analogical mechanism can give rise to rule-like, almost categorical, linguistic patterns as well as the observed variability in linguistic phenomena, and can account for the relevance of different degrees of abstraction in linguistic generalisation. At the same time we show that, contrary to some earlier conceptions of the role of analogy in grammar, analogy can function as a restrictive central mechanism of a grammatical system. The prerequisite is an independent, invariant procedure that constrains similarity-based reasoning. Such a system is implemented in the analogical algorithm AM::Parallel (Skousen & Stanford 2007), henceforth AM.

The testbed for our model is the phonological form of English noun-noun compounds: some have leftward stress, e.g. *ápple juice*, others have rightward stress, e.g. *apple píe*, and others are variable, e.g. *íce cream* and *ice créam*, where stress is marked by an acute accent on the prominent syllable. Although a number of recent corpus studies have shown that, *inter alia*, a compound's semantics and the identities of its constituent nouns are significant probabilistic predictors of stress placement (e.g. Plag et al. 2007, Plag et al. 2008, Plag 2010), it is still unclear why some types are more variable than others. Furthermore, although statistical models show that certain variables can be used to predict stress patterns, they say nothing about the mechanism that might be involved in creating new forms. This paper fills these gaps.

We use AM to successfully model stress assignment in a corpus of 486 nominal compounds, experimentally elicited from multiple speakers and carefully rated for stress placement (Bell 2013). In an analogical theory, stress is assigned to new compounds in accordance with the stress pattern of similar compounds previously encountered and stored in the lexicon. In the AM model presented here, degree of similarity is calculated using the compounds' constituents and semantic properties: those with a sufficient level of similarity are included in the set of potential analogues. Because the algorithm does this for each compound independently, in such a way as to maximise the certainty of its predictions, the type and degree of similarity that is relevant can vary from compound to compound. A number of compounds with particular semantic properties tend to cluster into a large, strong 'analogical gang' in the lexicon, influencing stress assignment in new compounds with the same semantic properties, and overriding potential constituent effects. Because a large number of compounds with similar semantics have the same stress behaviour, the result resembles the effect of a categorical rule. For many other compounds, however, stress assignment is based on more local analogues: in these cases, the identities of the constituents are key features in the computation of similarity. But in all cases, the underlying mechanism is the same.

The paper is structured as follows. In section 2 we discuss the empirical evidence for incorporating different levels of generality and different types of variability into a theory of linguistic generalisation, providing an overview of the pertinent literature (section 2.1) and discussing in detail pertinent evidence with respect to our testbed phenomenon (section 2.2). In Section 3, we then introduce the basic mechanisms of an analogical theory of morphophonology as implemented in AM. Section 4 describes the data used for our empirical models, which are described in Sections 5 and 6. The paper ends with a summary and conclusion in Section 7.

2. Locality and variability in linguistic generalisation

2.1 The general picture

The idea that different levels of generality are involved in linguistic generalisation is articulated especially, but not exclusively, in constructionist and exemplar-based work. The crucial insight is that linguistic generalisations can be based on different degrees of abstraction, ranging from very 'local' to very general. In this context, 'local' means that the relevant properties of exemplars or, depending on the framework employed, the inputs to linguistic rules are very specific, applying only to small groups of highly similar words or expressions; 'general', on the other hand, means that the relevant properties are more abstract, pertaining to larger groups of less similar words or expressions. An example of a 'local' generalisation is productive ablaut formation in the English past tense, which in nonce-word experiments has been shown to occur predominantly in words which are phonologically highly similar to existing words that are ablaut-forming (Bybee & Moder 1983, Prasada &

Pinker 1993). An example of a 'general' generalisation is the use of the suffix *–s* to mark the third person singular present in all lexical verbs in English.

The evidence that linguistic generalisation involves reference properties at different levels of abstraction comes from empirical work in a variety of linguistic sub-fields. In the realm of word-formation, Booij (2010) provides ample evidence of different degrees of generality in word-formation patterns, which lead him to postulate a layered, hierarchical structure of schematisation, where schemas differ in terms of their level of generality. From a slightly different theoretical perspective, Albright & Hayes (2003) and Albright (2002, 2009b) have argued in a series of simulation studies using the *Minimal Generalization Learner* that different levels of generality are needed to account for various phenomena in inflectional morphology. For phonology, it is argued for example in Pierrehumbert (2001) and in Albright (2009a) that both fine-grained phonetic properties of individual exemplars as well as more abstract phonological representations influence generalisation. Further evidence comes from the relevance of local as well as more abstract generalisations not only in the acquisition of syntax (Tomasello 2003 et seq.), but also in constructionist theories of adult grammar (Goldberg & Jackendoff 2005, Goldberg 2006 et seq).

The issue of variability and gradience has also been the focus of much recent work, especially in quantitatively oriented linguistics (e.g. Bod et al. 2003, Hay & Baayen 2005, Bresnan 2007, Coetzee 2009, Coetzee & Pater 2011, Ernestus 2011). The usage of the terms 'variability' or 'gradience' is largely theory-dependent. 'Variability' tends to be used to refer to outputs of linguistic generalisation mechanisms (such as rules or constraints) whereas 'gradience' is used to refer to the structure of representational categories. Which of the two terms is used, then, often depends on the assumed division of labour between the generalisation mechanism and the lexical representation. One interesting point of general consensus seems to be that variability does not occur equally across the board, but that linguistic generalisations differ in terms of how much variability they allow. Thus, most scholars agree that there are some categorical, in the sense of exceptionless, generalisations, but that at the same time there are also generalisations which are probabilistic in nature and which produce variable outputs. Unlike in earlier generative treatments, where a principled distinction was drawn between categorical and gradient phenomena, there is growing evidence that the distinction is not absolute, but a continuum (cf. Ernestus 2011 for an insightful overview discussion of the issue in phonological theory).

Another variability-related issue, which is subject to less consensus, is the question of the scope of variability. In particular, the term 'variable' is sometimes used to refer to grammatical processes which affect different lexical types in different ways. These are cases that are traditionally described as involving subregularities. However, the term is also used to refer to grammatical processes which produce variable outputs for the same lexical item, which in turn may occur either between different speakers or even within a single speaker (for examples, see the discussion of lexically conditioned variation in Coetzee & Pater 2011, section 5).

In the next section we will introduce our testbed phenomenon. We will show that stress assignment in English compounds is subject to generalisations at different levels of generality, which are probabilistic in nature to different degrees. Thanks to a series of recent large-scale studies (Plag 2006, Plag et al. 2007, Plag et al. 2008, Lappe & Plag 2007, Kunter & Plag 2009, Kunter 2007, Kunter 2009, Giegerich 2009, Bell 2013), the empirical facts of English compound stress are well understood. Furthermore, it has been shown that computational analogical algorithms can model the phenomenon quite successfully (Plag et al. 2007, Arndt-Lappe 2011). For both these reasons, English compound stress is an ideal vehicle for the elaboration of a detailed analogical account of linguistic generalisation.

2.2 English compound stress

In what follows, we first describe the variation in English compound stress, and then discuss how determinants with different levels of generality interact to produce this variation. Note that we will not review here the theoretical debate about the status of pertinent constructs as compounds or phrases (Chomsky & Halle 1968, Liberman & Sproat 1992). As has been shown in an extensive body of literature, stress placement in English noun-noun constructions is not reliably correlated with any morphosyntactic criterion (Giegerich 2009, Bell 2013 for detailed discussion). All the data to be analysed in this paper consist of pairs of nouns in which the first noun (N1) is the semantic modifier of the second (N2). For ease of exposition, and without theoretical commitment, we will refer to them all as compounds.

2.2.1 Variability

English compound stress displays variability of two kinds (cf. Kunter 2007, Kunter 2009: ch. 8 for detailed discussion). The first, which we will refer to as between-type variability, describes a situation where different lexical types behave in different ways. An example of stress variability between types can be found in the compounds *ápple cake* and *apple píe*. The former is usually left-stressed whereas the latter is usually right-stressed. Most of the literature on determinants of compound stress variability has focused on such between-type variability (e.g. Sampson 1980, Fudge 1984, Ladd 1984, Liberman & Sproat 1992, Olsen 2000, Olsen 2001, Giegerich 2004, Giegerich 2009). In studies that are based on a categorical rule or constraint system, the implication is that the stress pattern of a given type will be determined by the value of a particular predictor: in the examples given above, if N2 = *cake* then the compound will be left stressed, whereas if N2 = *pie*, the compound will be right stressed.

Apart from between-type variability, however, compound stress has also been observed to be subject to within-type variability, meaning variability within a single lexical type (Bauer 1983a, Bauer 1983b, Kunter 2011, Bell 2013). An example is *ice cream*, for which both left stress and right stress are attested. Such variation can occur both between speakers, where speakers consistently produce a particular pattern but not all speakers produce the same pattern, or within speakers, where the same speaker produces different patterns on different occasions. An interesting thing about within-type variability in compound stress is that it does not occur to the same extent across the board. Rather, different lexical types seem to be affected to different degrees. Recent evidence for this comes from Kunter (2011) and Bell (2013). Kunter's study investigates between-speaker variability among compounds with multiple occurrences in the *Boston Radio Speech Corpus* (BURSC, Ostendorf et al. 1996). Although the dataset used in the study was rather small (comprising 64 compound types), the analysis clearly suggests that within-type variability is a phenomenon that affects a minority of compounds (about 1/3 of the dataset investigated in the study), and that lexical types differ in terms of how much within-type variability they display. For example, Kunter finds that in his data the compounds *retírement age*, *bírth control*, *home phónes*, and *key detérrent* are not variable, whereas *búdget deficit* / *budget déficit*, *státe trooper* / *state tróoper*, *áccess barriers* / *access bárriers*, and *tásk force* / *task fórce* are highly variable (Kunter 2011: 192f.). Very similar findings emerge from Bell's (2013) study of within-type variability, which used a much larger corpus and a different variety of English. In a dataset of 802 compounds, each produced by four different speakers of British English, Bell (ibid.) found between-speaker variation in 21% of all lexical types.

In a rule-based grammar, within-type variation is difficult to account for, even in terms of exceptions to a rule. It therefore has to be relegated to the realm of 'performance' or possibly pragmatics: in other words, to some domain outside the grammar proper. In an analogical grammar, however, within-type variation is expected and arises from the same central mechanism as more fixed effects.

2.2.2 Different levels of generality

In a number of recent, large-scale empirical studies of English compound stress variation, it has been shown that five general types of factor act as important predictors of stress position:

- semantic factors (Plag et al. 2007, Plag et al. 2008, Bell 2013),

- constituent family (Plag 2006, Plag 2010, Arndt-Lappe 2011)
- lexicalisation (Plag et al. 2007, Plag et al. 2008)
- informativeness (Bell 2013, Bell & Plag 2012)
- length (Bell 2013, Bell & Plag 2012, Bell & Plag 2013)

Many of these factors had already been proposed to play a role in compound stress in the earlier theoretical and descriptive literature (esp. Bloomfield 1933, Lees 1963, Marchand 1969, Payne & Huddleson 2002, Spencer 2003, Giegerich 2004, Giegerich 2009, Sampson 1980, Fudge 1984, Ladd 1984, Liberman & Sproat 1992, Sproat 1994, Olsen 2000, Olsen 2001; cf. Plag et al. 2008 for discussion). However, in terms of grammatical theory, the findings of the recent larger scale empirical studies pose a double challenge to traditional accounts of the phenomenon.

The first challenge is that all pertinent studies show that English compound stress is probabilistic. There are clear tendencies, but none of the above factors determines the locus of stress in a categorical manner. The second challenge is that stress assignment involves the interaction of determinants at different levels of generality. This difference in generality can be seen by comparing semantic effects, such as the tendency for compounds where N1 is a proper name to be right stressed, with constituent family effects, such as the tendency for compounds with N2 = *cake* to be left stressed. Semantic effects are relatively general, requiring analysis on a relatively abstract level and applying to a relatively large number of compound types. Constituent family effects, on the other hand, are less general and hence more local, since they apply to compounds involving specific lexemes; they therefore require analysis on a less abstract level and apply to relatively fewer compound types. The challenge is that traditional grammatical systems, which are based on categorical rules, can conceptualise such lexeme-specific effects only in terms of exceptions, and do not attribute the same systematic status to both general and local factors.

Semantic categories that are relevant in English compound stress can broadly be classified into two types. The first type comprises categories that describe the relation between compound constituents. An example is found in compounds in which the first constituent expresses a material of which the second constituent is made, e.g. *gold earring*, *metal cage*, etc. The second type comprises categories that only pertain to one of the constituents. An example is found in compounds where the first constituent (N1) is a term denoting a time period, e.g. *summer holidays*, *Christmas present*, etc. Needless to say, there is a strong overlap between semantic relations and categories of N1 or N2. For example, in many cases where N1 denotes a time, the relation between N1 and N2 will be temporal: N2 occurs at time N1.

Empirical studies have consistently shown that certain semantic categories strongly correlate with rightward stress in English compounds. Particularly important categories are material, temporal and locative semantic relations, and compounds with N1 belonging to the semantic classes material, time, location, and proper name (Plag et al. 2007, Plag et al. 2008, Bell 2013). But even though these semantic categories have been shown to be quite successful as probabilistic determinants of right stress, one crucial argument against an account of compound stress purely in terms of these categories is that they pertain to only a minority of compounds in the lexicon. In other words, corpus studies have generally found that there are many right-stressed compounds which do not belong to any of the categories identified as favouring right stress (cf. e.g. Plag et al. 2008: 784 for discussion).

Constituent family effects, on the other hand, have often been cited as reasons for the existence of exceptional stress patterns (Schmerling 1971, Liberman & Sproat 1992, Giegerich 2004). Traditionally, such effects have been attributed to analogy, where analogy is used in the sense of denoting lexical patterns that, crucially, are not considered to be within the realm of grammatical rules (cf. e.g. Arndt-Lappe in press for discussion). Standard textbook examples are street names, which are stressed on N1 if N2 = *street*, but stressed on N2 if N2 = *avenue* or *lane* (as in *Óxford Street* vs. *Oxford Ávenue*, *Oxford Láne*). However, recent studies have made it abundantly clear that constituent family effects happen on a much larger scale than would be expected if they were exceptional formations (cf. e.g. Plag 2006, Plag et al. 2007, Plag 2010, Arndt-Lappe 2011, Bell 2013). Plag (2006) used an experimental study to compare constituent family effects and semantic effects for novel compounds denoting classical music pieces (e.g. *Twilight Sonata* vs. *Kauffman Sonata*, *Christmas Symphony* vs. *Lieberman Symphony*). He found consistent effects of the right compound constituent, which were independent of semantic effects. Corpus-based evidence for the importance of constituent family effects in stress assignment comes from studies by Plag and collegues (2007), Plag (2010), and Arndt-Lappe (2011). Plag (2007) presents a study of stress in compounds extracted from the CELEX lexical database (Baayen et al. 1995), which also involved computational modelling with the help of an analogical algorithm (TiMBL, Daelemans et al. 2007). Analysis of the analogical model revealed that constituent family information acted as the most important information source for the model. This insight received strong support from the studies reported in (Arndt-Lappe 2011), where stress assignment in compounds from two different corpora (CELEX and the *Boston University Radio Speech Corpus*, BURSC, Ostendorf et al. 1996) was modelled with the help of two different analogical algorithms (TiMBL and AM). Again, the identities of both the first and the second constituent of the compounds tested emerged as the empirically most useful information source in all experiments. In particular, constituent family information was found to be far more important than other types of information that are known to correlate with stress assignment, including semantic information. Finally, the study reported in Plag (2010) shows for data from three different corpora that stress biases present in the constituent family also emerge as highly important predictors of stress assignment in a regression analysis.

A question that emerges from the findings of the studies just discussed is how the two types of effect – semantic and constituent family effects – interact in the grammar of English. So far, the only attempt to integrate both types of effect in an analogical model has been made in Arndt-Lappe (2011). There it was shown that the constituent family acts as a much stronger predictor than the pertinent semantic categories, to the extent that incorporating semantic information into the model did not lead to a significant improvement of the model's predictive power. It is, however, unclear how these findings can be reconciled with the observation made in many other studies that semantics does play a role in compound stress assignment. As Plag et al. (2008) note in the conclusion to their study of semantic effects published in this very journal: '[a]nalogical models may raise the question of why one should find robust semantic effects'(Plag et al. 2008: 787). In other words, if constituent family was really the only relevant determinant, it would be difficult to explain why semantically similar compounds display the same stress pattern even if they do not share any constituents.

Another consideration is that constituent family effects appear to be at least partly semantic in nature. In the case of *street* compounds, for example, the tendency to left stress only applies if the compound is the name of a thoroughfare, and not in other cases such as *London stréet* in the context *Whenever I walk in a London street*. Constituent families may even be associated with different stress patterns depending on the semantics of the construction involved (Bell & Plag 2013). For example, two clear patterns can be discerned for compounds where N1 = *toy*. When *toy* is used attributively to indicate that N2 is a replica of something else, it tends to be associated with right stress, e.g. *toy bóat*, *toy trúmpet*, *toy fáctory* (a model factory for playing with). On the other hand, when *toy* denotes playthings in general it tends to be associated with left stress, e.g. *tóy shop*, *tóy box*, *tóy factory* (a factory that makes toys, cf. Faiß 1981 for the example).

Clearly, neither form-based constituent families alone nor semantic categories alone can account for the empirical facts of English compound stress. In this paper, we will therefore present a model that incorporates both semantic and constituent family information, and use it to show how factors with different levels of generality interact in an analogical system. In contrast to categorical rule or constraint-based grammars, an analogical grammar does not require that less general effects are treated as exceptions.

Apart from semantic effects and constituent family effects, three other kinds of variable have been shown to be influential in compound stress assignment. These are lexicalisation (Plag et al. 2007, Plag et al. 2008), informativeness (Bell 2013, Bell & Plag 2012, Bell & Plag 2013) and length (Bell 2013, Bell & Plag 2012, Bell & Plag 2013). These features will not be included as explicit predictors in our model, but will be present implicitly and are therefore briefly discussed in the following paragraphs.

Firstly, a high degree of lexicalisation, measured in terms of compound frequency, listedness in a dictionary or tendency to be written as a single orthographic word, has been shown to correlate in English compounds with a tendency towards left stress (e.g. Plag et al. 2007, Plag et al. 2008, Bell 2013). Such a lexicalisation effect may at a first sight seem difficult to reconcile with the basic idea of computational analogical models that *all* language encountered leaves a trace in the lexicon. However, we assume that the measures of lexicalisation used in these studies are in fact indicative of semantic lexicalisation or institutionalisation (Bauer 2001: 45-46): in other words, that they are correlates of semantic opacity. Our data is semantically coded in a way that attempts to take this factor into account.

Secondly, it has been shown that the greater the infomativity of a compound constituent, measured in terms of either expectedness or semantic specificity, the more likely it is to be stressed (Bell 2013, Bell & Plag 2012, Bell & Plag 2013 for recent empirical support and discussion). Furthermore, a constituent’s informativity, as indicated by these measures, can be used to predict its bias for one stress pattern or the other (Bell & Plag 2013). This suggests that the informativeness effect is closely related to the constituent family effect and, as Bell & Plag argue, may even underlie it. In our model, compound constituents will be used as predictors and, since each constituent will have a certain level of informativeness, this factor will be implicitly present.

Finally, longer compounds (in terms of number of syllables) have a tendency to be right stressed (Bell 2013, Bell & Plag 2012, Bell & Plag 2013). Again, by including the identities of the compound constituents as predictors, we are implicitly including length, since each constituent has a certain number of syllables. However, the effects of informativity and length will not be seen in our models independently of other constituent effects, since we do not explicitly code for them: both variables are numeric in nature, and the AM algorithm works best with non-numeric data.

3. An analogical theory of morphophonology

In an analogical theory of morphophonology, new word forms (analogical targets) are created on the basis of existing forms in the lexicon (analogical bases). To be sustainable, such a theory therefore needs to spell out how analogical bases are selected from the lexicon in a systematic and constrained way so as to produce forms that closely mirror those produced by humans. An example of such a mechanism is provided by the computational algorithm AM (Skousen 1989, Skousen 1992, Skousen et al. 2002, Skousen & Stanford 2007). In this section we describe how AM assigns stress to new compounds in a way that accounts both for the observed variability in output and for the different levels of generality amongst known predictors. For further explanation of the algorithm and its underlying theory the interested reader is referred especially to Skousen (1992), Skousen et al. (2002), Skousen (2005), Skousen (2009), Eddington (2000), Eddington (2002), Eddington (2004: ch. 5).

Figure 1 provides an illustration of how AM assigns stress, using the compound *fish pie* as an example. We assume in this example that *fish pie* is a novel compound, encountered by the speaker for the first time. Previously-encountered compounds are stored in the lexicon, with each exemplar represented as a structured set of coded features. In this example, and in our models, exemplars are coded for just five different features, which represent different levels of generality. Features 1 and 2 are the specific identities of the first and second compound constituents. Features 3 and 4, on the other hand, are more general semantic features: respectively, the semantic relation between N1 and N2, and the semantic class of N1. Finally, Feature 5 is the stress category, which will function as the target category for the classification of new items: in other words, the model will try to predict the value of this feature on the basis of information in the lexicon and the other features of the new item. Figure 1 shows only a small subset of all items in our dataset.

If a new compound is to be classified for stress, the system extracts from the lexicon a group of exemplars that are similar to the new compound in terms of the other coded features. This group of exemplars is known in the AM literature as the analogical set. The exemplars in the analogical set serve as analogues on the basis of which stress will be assigned to the new compound. A crucial problem for an analogical theory is how to determine which exemplars in the lexicon are included in the analogical set. This problem underlies much of the criticism mounted against traditional analogy-based models of grammar (cf. e.g. Bauer 2001: 75-97 for an overview of pertinent arguments), and the solution has remained often opaque in studies using computational analogical models. The problem has two different dimensions. Firstly, it is unclear what degree of similarity is necessary for an exemplar to be selected for the analogical set. For example, in Figure 1 the analogical set comprises three exemplars that share three features with the new word (namely N2 lemma and both semantic features) as well as one exemplar which shares only two features with the new word (namely N1 lemma and the class of N1). Secondly, it is unclear which features, if any, are most important in selecting members of the analogical set. For example, in Figure 1 both *fish shop* and *chocolate raisin* share two features with the target word, but the former is included in the analogical set while the latter is excluded.

exemplars in the lexicon

Feature 1: N1 lemma	**Feature 2: N2 lemma**	**Feature 3: Semantic relation**	**Feature 4: N1 semantic class**	**Feature 5: Stress**
lemon	pie	N1 ingredient in NN	material	right
steak	pie	N1 ingredient in NN	material	right
mince	pie	N1 ingredient in NN	material	right
fish	shop	no rightward relation	material	left
county	caterer	N2 located in N1	location	right
south	avenue	no rightward relation	location	right
silver	ware	N2 made of N1	no rightward class	left
cat	food	no rightward relation	no rightward class	left
chocolate	raisin	N1 ingredient in NN	material	right
coffee	jar	no rightward relation	no rightward class	left
sport	centre	no rightward relation	no rightward class	left

analogical set

Feature 1: N1 lemma	**Feature 2: N2 lemma**	**Feature 3: Semantic relation**	**Feature 4: N1 semantic class**	**Feature 5: Stress**
lemon	pie	N1 ingredient in NN	material	right
steak	pie	N1 ingredient in NN	material	right
mince	pie	N1 ingredient in NN	material	right
fish	shop	no rightward relation	material	left

3 x right
1 x left

new item

Feature 1: N1 lemma	**Feature 2: N2 lemma**	**Feature 3: Semantic relation**	**Feature 4: N1 semantic class**	**Feature 5: Stress**
fish	*pie*	*N1 ingredient in NN*	*material*	*???*

stress: right (majority choice)

Figure 1. The basic architecture of an analogical model[2]

Although the reason for the inclusion or exclusion of particular exemplars may not be immediately apparent, an analogical algorithm such as AM in fact bases selection of the analogical set on a single principled mechanism. AM selects analogues in such a way as to maximise its certainty about the value of the target variable. The way it does this will be explained using Figure 2. In this and similar diagrams elsewhere in the paper, the central oval shows the features of the target item, in this case *fish pie*, and the three outer ovals contain all items in the lexicon that share at least one of these features. Each concentric zone represents a set of so-called ‘contexts’ which are similar to the target item to the same degree. A context is

[2] Note that the coding will be explained in section 4 in detail. The present discussion focuses on the central mechanism.

a constellation of features shared with the target item. For each context, we represent the set of exemplars that belongs to this context as a block of words. Thus, exemplars in the outermost zone in Figure 2 share at least one feature with the target item, those in the middle zone share at least two features with the target item, and those in the inner zone share three features with the target item. The assumption is that *fish pie* is a newly encountered compound, and so no previous exemplars have been encountered that share all four features. Contexts that constitute the analogical set for *fish pie* are shaded.

AM starts analogue selection from those exemplars in the lexicon that are maximally similar to the target item, considering all logically possible contexts, i.e. constellations of shared features. A context is defined in terms of particular features shared with the target item, irrespective of whether or not other features are shared. This means, for example, that any exemplar sharing two features with a target item will automatically occur in three contexts: shared feature 1, shared feature 2, shared features 1&2. Likewise, any exemplar sharing three features with the target will occur in seven contexts: shared feature 1, shared feature 2, shared feature 3, shared features 1&2, shared features 1&3, shared features 2&3, shared features 1&2&3.

Starting from each context that is maximally similar to the target, the model's strategy is to also include contexts that share fewer features with the target along each similarity dimension, provided these less similar contexts do not reduce certainty about the outcome. By 'similarity dimension', we mean the series of subcontexts that can be derived from each maximally similar context. For example, the following three contexts represent one dimension: shared features 1&2, shared feature 1, shared feature 2. As described above, any context that involves three shared features will lead to a further six contexts along that dimension of similarity. Furthermore, less general contexts will occur in more than one dimension; for example, shared feature 2 is a subcontext both of shared features 1&2 and of shared features 2&3.

In Figure 2, we see that the only exemplars in the model's lexicon that share three features with *fish pie* are those that share the N2 lemma and both semantic features. These three exemplars (*lemon píe*, *steak píe*, and *mince píe*), which are to be found in the innermost zone in Figure 2, are selected for the analogical set because they are maximally similar to *fish pie* along this dimension. All three are right stressed, and so they predict right stress for *fish pie* with 100% probability.

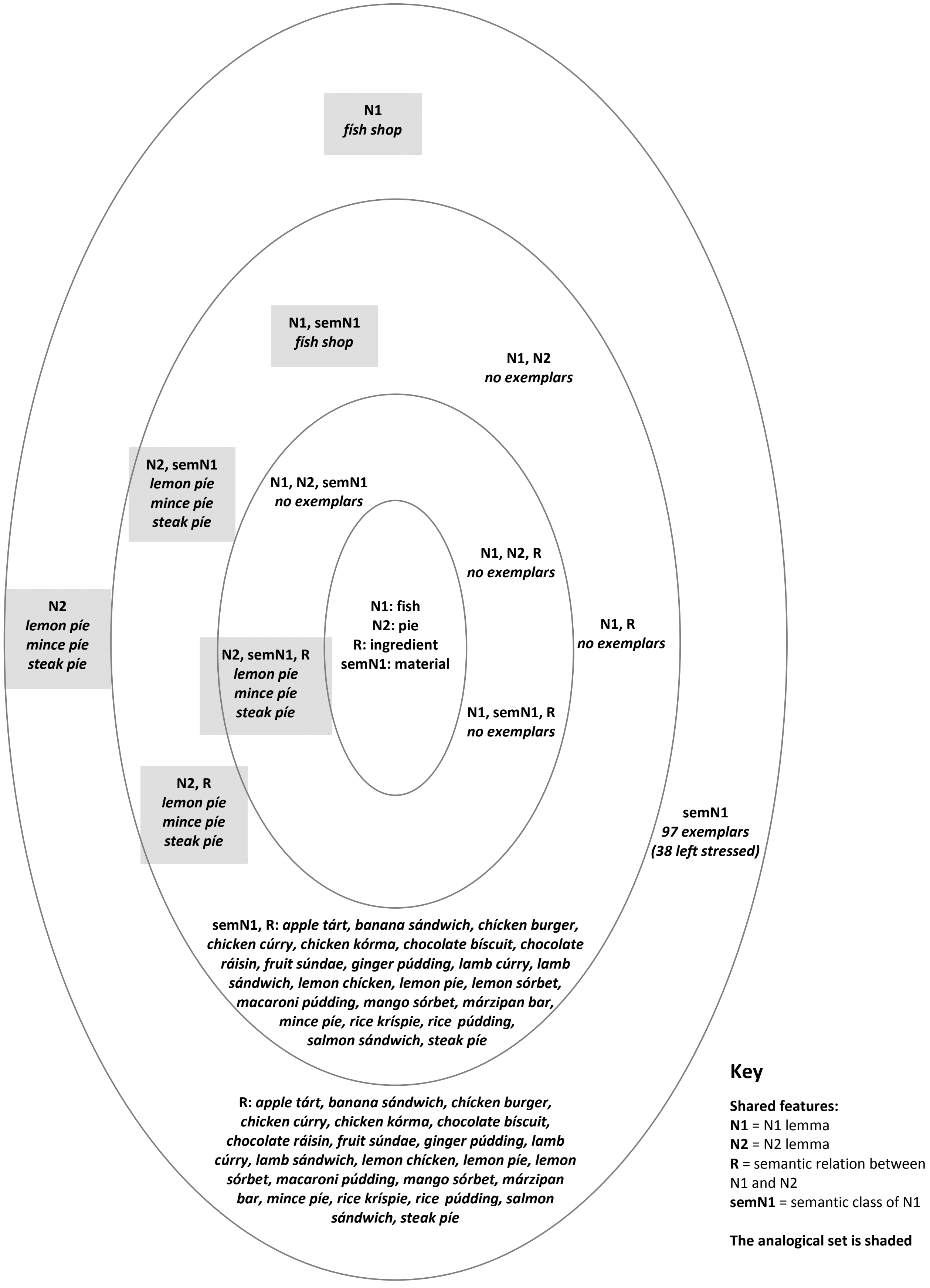

Figure 2: The analogical set for *fish pie* (36 rightward votes; 2 leftward votes)

The algorithm next checks to see whether it can include into the analogical set exemplars with only two of these features, without reducing the certainty of its prediction. In the case of ‘N2, R', as well as ‘N2, semN1', this is possible. Only the same three exemplars (*lemon píe*, *steak píe*, and *mince píe*) belong to each of these contexts, and so including these contexts does not reduce the probability of *fish pie* being right stressed. However, the context ‘R, semN1’ includes two exemplars that are left stressed (*chícken burger* and *márzipan bar*). Including this context would therefore reduce the certainty of the model’s stress prediction, and so AM excludes the context from the analogical set. Having excluded exemplars in this context, the model will not proceed to more general contexts along this dimension, and so two of the contexts with a single shared feature, namely ‘N1’ and ‘semN1’, are not considered. On the other hand, the third single-feature context, ‘N2’, is included because it contains only right-stressed exemplars.

The alert reader might be wondering at this stage why the compound *físh shop* is included in the analogical set for *fish pie*, since it is left stressed and therefore reduces the overall certainty of the model’s prediction compared with the outcome based only on the three *pie* compounds. In fact, *físh shop* is included for exactly the same reason that the *pie* compounds are included: it is maximally similar to *fish pie* along its dimension of similarity. In other words, there is no exemplar in our sample lexicon that shares both 'N1' and 'semN1' with *fish pie*, and is more similar to it than *físh shop* is. In both of the contexts ‘N1, semN1' and ‘N1', *físh shop* is the only exemplar in the lexicon. Both of these contexts are therefore included in the analogical set, since inclusion of the less similar context does not change the predicted outcome along this dimension.

From this description it can be seen that, in selecting which exemplars to include in the analogical set, AM gives equal weight to all coded features. In contrast to rule-based frameworks, the analogical mechanism gives equal weight to very general factors such as semantic relation as to very specific factors such as N1 lemma, with no a priori ordering. In other words, all determinant factors are treated as equally important although they differ in terms of the number of compounds to which they apply.

AM's classification is probabilistic: the algorithm gives the probability of each possible value being assigned, based on the distribution of these values amongst exemplars in the analogical set. And in calculating these probabilities, AM takes into account the degree of similarity between an exemplar and the new item, as well as the number of exemplars with a particular set of features. The more similar an exemplar is to the new item, the more weight it receives, and the more exemplars that share a particular set of features, the greater the weight assigned to each of them. In our case, the target category is 'stress', with possible values 'left' and 'right', and the basis of classification is the stress pattern of exemplars in the analogical set. Each exemplar in the analogical set receives a number of votes for its own stress pattern, depending on how many coded features it shares with the new compound, and how many other compounds in the lexicon share that particular set of features.

In Figure 2, the analogical set contains one left-stressed and three right-stressed exemplars. However, not all of these exemplars are equally similar to the new word, *fish pie*. Specifically, left-stressed *físh shop* shares two features with the target, namely 'N1' and 'semN1', whereas right-stressed *lemon píe*, *mince píe*, and *steak píe* each share three features with the target, namely 'N2', 'semN1', and 'R'. Furthermore, *lemon píe*, *mince píe*, and *steak píe* form a group belonging to exactly the same set of contexts, known in the AM literature as a 'gang', whereas *físh shop* is the only exemplar with N1 = *fish*. On both these counts, the right-stressed *pie* gang will each receive more votes than the left-stressed *físh shop*.[3]

To facilitate assessment of the predictive power of concrete simulations, studies using AM have often transformed the output probabilities into categorical choices. Such categorical choices are based on the majority vote of exemplars in the analogical set; in the case of *fish pie*, discussed above, this would mean categorically right stress. However, the fact that AM's stress predictions are actually probabilistic allows us to model within-type variability: a successful model will be one where the probability of a particular value as given by the model closely reflects the attested frequency of that value. For example, if our model predicts right stress for *fish pie* with a probability of 95%, then we would expect nearly all tokens of *fish pie* to be right stressed, but would not be surprised to find a few tokens (5%) with left stress.

In what follows we will present two simulation studies with AM which show that the challenges posed by English compound stress assignment to grammatical theory can be successfully overcome by an analogical view of linguistic generalisation. In the first study we will focus on different degrees of generality and show how the observed patterns of between-type variability emerge from the different types of analogical set in the model. In the second study we will look at within-type variability. The discussion of the two studies will be preceded by a short presentation of our data.

---

[3] AM achieves this weighting by assigning to each exemplar in the analogical set one vote for every member of its gang for each context where the gang occurs in the analogical set: remember that the more similar an exemplar is to the target, the more contexts in which it will occur. In the example given, *lemon píe*, *mince píe*, and *steak píe* each receive 12 votes because they occur as a gang of three in four different contexts. *Físh shop*, on the other hand, receives only two votes because it occurs as a gang of one in only two included contexts. Overall, the analogical set therefore produces 36 votes for right stress but only 2 votes for left stress.

4. Data and coding

The set of compounds used for our models, together with their attested stress patterns, come from Bell's large-scale empirical study of the English noun-noun construct (Bell 2013); this data has kindly been made available to us by the author. The full dataset comprises 1,000 compound types extracted from the demographic section of the *British National Corpus* (BNC); this section of the BNC contains 4.23 million words of spontaneous conversation recorded by informants selected to represent a demographically representative sample of the population. All the types sampled were tagged in the corpus as noun-noun combinations, and all occurred at least once as spaced, i.e. two-orthographic-word, tokens. Stress information was gathered in a reading experiment where each compound was read in the same carrier sentence by four different native speakers of British English. Stress was then coded by two independent, phonetically trained listeners using two different methodologies, and tokens for which the stress judgements were not unanimous were excluded. Subsequent removal of types with fewer than four tokens with agreed stress, yielded a final dataset of 3,405 compound tokens distributed over 802 different compound types. This dataset contains some types where stress in all four experimental tokens was invariably left or right (left: 377 types = 47% of all types; right: 257 types = 32% of all types), and others where stress was variable, i.e. left in some tokens and right in other tokens (168 types = 21% of all types).

From Bell's final dataset (ibid.), we extracted two subsets to be used in our AM simulations. A constraint on both subsets was that at least one of the compound constituents must have a family in the dataset, so that the algorithm would have access to both semantic and constituent-family information for each type. In other words, a compound was only included if at least one of its constituents appeared in the same position in at least one other compound in the dataset, e.g. *fish pie* and *fish shop*. Furthermore, compounds were excluded if they displayed ambiguity in terms of the semantic features coded or if either constituent was itself a compound. The subset thus produced contained 1,913 tokens representing 486 different types. Of these, 293 (49%) were consistently left-stressed, 147 (30%) were consistently right-stressed, and 100 (21%) were produced with both stress patterns. We used this dataset to investigate within type variability in stress placement, since it included both compounds that showed within-type variability in the reading experiment and others that did not. In our first simulation, however, we wanted to investigate in more detail how AM integrates predictors with different levels of generality. We therefore selected only those compound types produced with the same stress pattern by all speakers in the reading experiment. Thus, within-type variability was deliberately excluded from the lexicon for this simulation, in order to facilitate the analysis of analogical sets in the model. Excluding variable types also meant that some additional non-variable types had to be excluded, since they no longer had a family member in the dataset, i.e. another compound with a shared constituent. The final dataset for this first study comprised 368 compound types, of which 232 (63%) were left stressed and 136 (37%) were right stressed.

The coding of the predictor variables for our data also comes from Bell (2013), where all compounds were coded for their first and second constituent lemma (i.e. the singular form of each constituent) and their semantics. The semantic coding included both semantic relations and classes of N1 that have in previous studies been shown to trigger right stress (esp. Plag et al. 2007, Plag et al. 2008, Bell 2013). Figure 2 lists the relations and classes coded, with examples from the dataset.

Semantic relation between constituents	
N2 is made of N1	*leather chair*
N2 is N1	*baby boy*
N2 occurs during N1	*winter month*
N2 is located at/on/in N1	*country lane*
N1 has N2	*plant root*
N1 is an ingredient in foodstuff NN	*apple tart*
Semantic class of N1	
N1 is a name	*Tyson case*
N1 is a material	*cotton sheet*
N1 is a time period	*morning peak*
N1 is a location	*farm yard*
N1 is a social group	*family argument*

Table 1: Semantic categories coded

The six semantic relations and the category 'N1 is a name' were coded by two independent raters, one a trained linguist and the other a professional lexicographer. Cases of initial disagreement were discussed until a consensus was reached. A semantic relation was coded to apply only if the meaning of the relevant compound entails that relation. This meant, for example, that the locative relation 'N2 is located at/on/in N1' was coded to apply in the case of *garage wall*, but not in the case of *door handle*. Although in both cases N1 describes a location where N2 can be found, the two compounds differ in terms of the status of N1 as a location. In the case of *garage wall*, the usual meaning of the compound entails that the wall is located in a garage. This is not true for *door handle*, where an object can still be a door handle without actually being fixed to a door. This entailment criterion means that the coding of semantic relations in Bell (2013) and in the present paper is more restrictive than in some previous studies (e.g. Plag et al. 2007, Plag et al. 2008). Bell (ibid.) argues that compounds for which the entailment criterion applies are more semantically transparent than those where it does not apply, and that these ‘rightward leaning’ semantic relations have an effect on compound stress only when they apply with a maximal degree of transparency (cf. Bell 2013 for discussion). If this is correct, then we assume that the tendency for compounds with these relations to favour right stress is the inverse of the tendency noted in previous studies for semantically lexicalised (and therefore opaque) compounds to favour left stress (e.g. Giegerich 2004, Plag et al. 2007, Plag et al. 2008).

The remaining classes of N1, namely ‘N1 is a material’, ‘N1 is a time period’, ‘N1 is a location’, and ‘N1 is a social group’, were coded using the WordNet lexical database (Miller et al. 1990). The database was checked for hypernyms of each N1, and an N1 was coded as belonging to a given class only if listed as a hyponym of a relevant concept. The hypernyms used are shown in Table 2.

N1 semantic class	**Hypernym in WordNet**	**WordNet definition**
material	substance (sense 1)	that which has mass and occupies space
	fabric (sense 1)	cloth, material, textile
	building material	material used for constructing buildings
time period	time period	period of time, period
location	location (sense 1)	a point or extent in space
social group	social group	people sharing some social relation
	social class	socio-economic class

Table 2: Hypernyms used to define classes of N1

In a few cases, more than one relation or class of N1 applied to the same compound. In these cases, the compound was coded for both features. For example, the semantic relation for *school piano* was coded as N1hasN2/N2 is in N1, the class of N1 for *Christmas trimming* was coded as N1is a name/N1is a time period, and the class of N1 for *school lunch* was coded as N1is time period/N1 is a social group.

Before we proceed to the discussion of the simulations, a principled note is in order about the theoretical implications of the coding of the variables. First and foremost, it is clear that the algorithm is given information about which features are relevant for its particular task, i.e. stress assignment. In this respect, the present model does not differ from many other grammatical models, where it is standardly assumed that knowledge about which categories are relevant for a particular type of generalisation is part of the system. In the exemplar-based literature, however, we also find the idea that feature relevance is emergent rather than preconceived (cf., e.g., Skousen 2009, Skousen 2005 for some discussion). In the present study, we remain agnostic to this issue. A second note concerns the issue that in our coding there is considerable overlap between the categories that were coded. Thus, for example, we can expect that in compounds where 'semN1' is classified as 'material', the semantic relation 'R' between N1 and N2 is likely to be a material one. What is more, the constituent identity of N1 or N2 will also overlap to a considerable degree with the semantic categories. This overlap is a feature that has been found to be characteristic of naturally occurring compounds (e.g. Maguire et al. 2010). Furthermore, having run different simulation experiments using different feature combinations, we also find that excluding one of the semantic features does not lead to fundamentally different outcomes in our models (cf. section 5 below).

5. Study 1: Different levels of generality in compound stress

The dataset used in this study is the set of 368 compound types that do not display within-type variability. The algorithm was given this same dataset both as a lexicon and as a test set, but the AM parameters were set in such way as to make sure that analogical sets never included the item to be classified. The algorithm therefore treated all compounds tested as novel compounds.[4]

In what follows, we first provide an overview of our model's success in predicting stress placement, and then investigate how factors with different levels of generality interact as determinants of stress in the model. To do this, we perform an in-depth analysis of how the algorithm composes analogical sets. We will see that generalisations with different degrees of generality emerge in an analogical model as an epiphenomenon of different set sizes. Local patterns emerge from small analogical sets whose members are highly similar to one another and to the test item; more general effects emerge from larger analogical sets whose members are still similar, but have fewer features in common. Both types of set result from the distribution of pertinent features amongst items in the lexicon.

5.1 Overall performance of Model 1

To assess the overall performance of the model, we had AM transform its probabilistic stress predictions into categorical predictions for each compound. This means that if the percentage

[4] In all experiments reported in sections 5 and 6, we had the algorithm compute analogical sets using pointers, not occurrences (see Parkinson 2002 for a general outline of the options provided by AM). Furthermore, best results were achieved if AM treated both the presence and the absence of a semantic category as an informative value (the –nulls parameter in the program was set to 'include').

of votes for either left or right stress was greater than 50%, the test compound was predicted to be categorically left- or right-stressed, respectively. This procedure is well-established in the literature applying exemplar-based models to linguistic tasks (cf. e.g. papers in Daelemans & van Bosch 2005, Skousen et al. 2002 for examples and discussion). Also, this procedure allows us to compare the performance of our model with those tested in Arndt-Lappe (2011), where AM was set the task of assigning stress in compounds from a different dataset (cf. section 2.2 for discussion), and where the same assessment procedure was used.

Table 3 provides an overview of the model's predictive power. As measures of success, we provide the percentage of correct predictions and F-scores for left and right stresses as well as two types of averaged F-scores for the complete dataset (cf. e.g. Daelemans & van Bosch 2005, Arndt-Lappe 2011 for explanation and discussion of the measures). Unlike measures that only provide the percentage of correct predictions for a given task, F-scores are sensitive to both underprediction and overprediction of a given outcome. Micro-averaged F-scores constitute an average F-score for the whole dataset that takes into account skewings in the distribution of observed outcomes: in this case, left stress is more frequent than right stress in the dataset, hence predictive power for left stress is given more weight than that for right stress. In contrast, macro-averaged F-scores treat both possible outcomes as equally important.

F-score, micro-averaged:	0.93
F-score, macro-averaged:	0.92
% correct predictions (overall):	93%
F-score for left stress:	0.94
% correct left predictions:	94%
F-score for right stress:	0.91
% correct right predictions:	91%

Table 3: Overall performance in Study 1 using both semantics and constituent information

Using a combination of constituent family information and semantic information, AM successfully predicts more than 90% of stresses in the data, irrespective of the measure used for assessment. This overall accuracy is very similar to the overall levels of accuracy found by Arndt-Lappe (2011); where the two studies differ, however, is in the predictive accuracy for right stress. Whereas in our study, predictive accuracy for right stress is only slightly below that for left stress, predictive accuracy for right stress in Arndt-Lappe's study (ibid.) never exceeded 67%. We hypothesise that this increase in accuracy for right stressed types is associated with the more restrictive criteria by which our semantic features were coded.

In terms of the information sources used, the present simulation also differs remarkably from those presented in Arndt-Lappe (2011), since semantic information plays a crucial role in enhancing the predictive accuracy of our model. This is illustrated in Table 4, where we compare F-scores for our model using both constituent and semantic information (shaded, same model as in Table 3) with those for models that use only semantic information or only constituent family information, for the same dataset. We see that the best predictive results are achieved if the model is allowed to use both types of information source. This is contrary to the findings in Arndt-Lappe (2011), where semantic information did not improve predictive power compared with a model using constituent information alone.

Information source: constituents	Information source: semantics	F_{left}	F_{right}	$F_{microav.}$	$F_{macroav.}$
✓	✓	0.94	0.91	0.93	0.92
×	✓	0.93	0.88	0.91	0.90
✓	×	0.90	0.82	0.87	0.86

Table 4: Performance in Study 1 using different information sources

However, note that, unlike the dataset used by Arndt-Lappe, the present dataset cannot be used to compare the relevance of constituent family information and semantic information in absolute terms. This is because, in the present dataset, constituent family information is not available for all compound constituents: most compounds in the dataset share only one of their constituents with another exemplar or exemplars in the data. This means that for most compounds, constituent family information is only available either for N1 or for N2, not for both N1 and N2. Whereas this may well explain the discrepancy between Arndt-Lappe (2011) and the present study in terms of the predictive power of the constituent family, note that the fact that the dataset does not contain all potentially available constituent family information does not bear on the point that the present paper is making. There is independent evidence that both semantics and constituent family information play an important role in compound stress assignment: the point of interest here is to show how these different effects can successfully be accommodated within an analogical model of grammar.

5.2 Local and general analogies – a matter of degree

In order to explore how effects with different degrees of generality emerge in an analogical model, we conduct an in-depth analysis of the composition of analogical sets in the model presented in Table 3. Recall from Section 3 that AM will include less similar exemplars in an analogical set under two conditions: one is where less similar exemplars either maintain or increase certainty about the value of the output variable compared with more similar exemplars along the same dimension; the other is where there are no exemplars in the lexicon that share the same features and are more similar to the target item. Under either of these conditions, the analogical set will include exemplars that have relatively few features in common with the target item. On the other hand, if neither condition applies, the analogical set will contain only exemplars that are highly similar to the target item. We refer to the members of such highly similar analogical sets as 'local' analogues. On the whole, more general analogical sets will contain more exemplars than more local sets, and we can take advantage of this fact to identify these different types. Figure 3 provides an overview of the sizes of the analogical sets in our model.

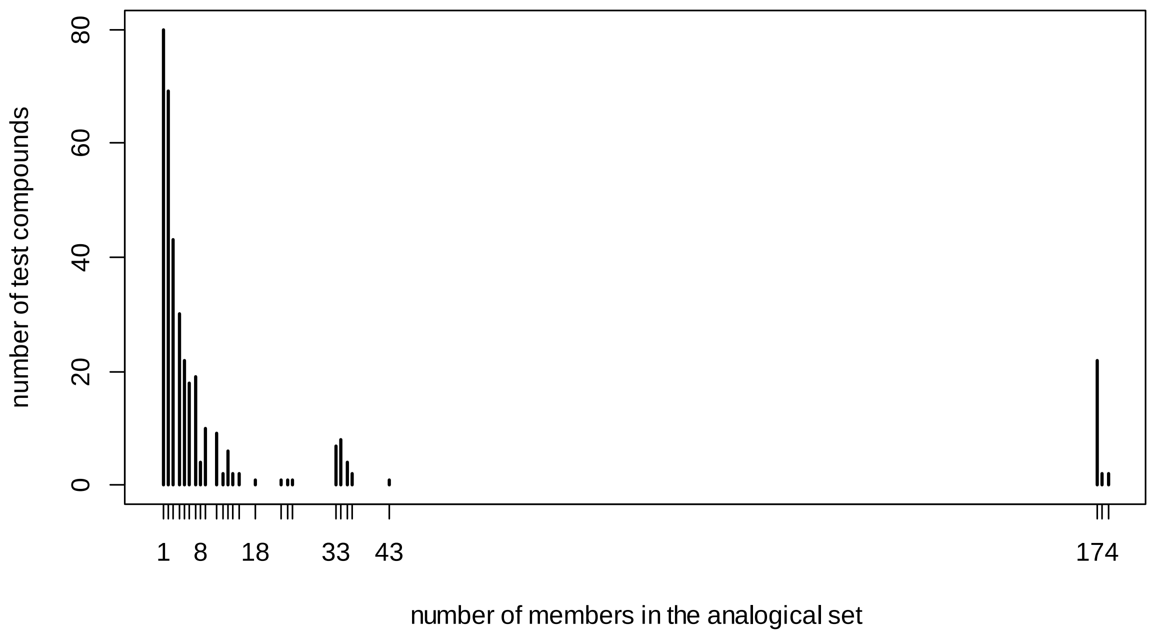

Figure 3: Sizes of analogical sets in Study 1

In Figure 3, the sizes of the analogical sets in our model, in terms of the number of members, are plotted against the number of compounds which have an analogical set of that particular size. We see clearly that analogical set sizes are not evenly distributed among all test compounds, but that, instead, there are clusters of different set sizes. For the vast majority of compounds in our dataset, the analogical set is very small. For example, for 149 of all 368 test compounds (i.e. for 40.49%), the set contains no more than two exemplars. The leftmost cluster of bars in Figure 3 comprises 306 test compounds (i.e. 83.15% of the dataset), with analogical sets of between 1 and 12 exemplars. For these compounds, it is likely that the model predicts stress assignment on the basis of highly local analogies. For the remaining 62 compounds, analogical set sizes fall into two groups. One group, the rightmost group in Figure 3, comprises 26 compounds (i.e. 7.07% of the dataset) with a very large analogical set (174 - 176 members). These compounds are therefore predicted to be stressed on the basis of very general analogies. The other group, consisting of 36 compounds (i.e. 9.78% of the dataset) is heterogeneously distributed over several smaller clusters. The analogical set size here varies between 13 and 43 members; analogies for these compounds are thus more general than in the larger cluster, but less general than in the rightmost cluster.

In what follows, we will investigate why analogical sets fall into the distinct size clusters shown in Figure 3. We will see that these clusters reflect not only the different degrees of similarity involved in analogical set formation but also, in the case of mid- and large-sized sets, analogical gang behavior: specifically, a situation where the same compounds reappear again and again in a number of analogical sets. In the case of the small-sized cluster, different sets are composed not of the same compounds, but of compounds that share the same type of features with the target. The size clusters identified in Figure 3 reflect situations which in other frameworks have been ascribed to different mechanisms: in ascending order of set size, these are local analogy, rule-governed behaviour, and a default situation. We will look at each type of analogical set in turn: the small sets (1 - 12 exemplars), the mid-size sets (13 - 43 exemplars), and the large sets (174 - 176 exemplars).

5.2.1 Small analogical sets

Among the 306 compounds with small analogical sets, 188 compounds (i.e. 61.44%) have leftward stress and 118 compounds (i.e. 38.56%) have rightward stress. Stress was predicted correctly for 180 of the left-stressed compounds and for 108 of the right-stressed compounds.

Closer inspection of the small analogical sets reveals that they all contain compounds sharing a constituent with the target word. In fact, for 251 of the 306 compounds in this group (i.e. 82.03%), the analogical set consists exclusively of items belonging to the constituent family or families. Examples of pertinent analogical sets containing two exemplars are given in Table 5.

Target word	**Analogical set**	
cat food	*convénience food*	*cát muck*
banana sandwich	*lamb sándwiches*	*salmon sándwiches*
health care	*cár care*	*héalth farm*

Table 5: Examples of small analogical sets in Study 1

The fact that most analogical sets in our model are small means that the model bases most of its predictions on very local analogies. 'Local' means that members of the analogical set are highly similar to the target word and, as a consequence, small in number. In AM, having such small analogical sets also means that the model has not incorporated more, less similar exemplars into the analogical set because this move would have led to a reduction in certainty about stress behaviour.

It is important to note, however, that, although all small analogical sets in our model contain members of the constituent family, this does not mean that constituent family information is more important for stress assignment than semantic information. The example of *banana sandwich* is representative here. The members of the analogical set, *lamb sandwiches* and *salmon sandwiches*, share not only the second constituent, but also both semantic features with the target: N1 is a material, and the relation between N1 and N2 is that N1 is an ingredient in foodstuff NN. What is relevant for the formation of analogical sets, then, is a high degree of similarity, which is measured in terms of the number of features shared between an exemplar and its analogue(s), rather than privileging one type of feature over another. Furthermore, we see that local analogies in compound stress assignment arise in conditions where there is a strong correlation between semantic and constituent family information, in the sense that the presence of a certain constituent is likely to co-occur with a specific type of compound semantics. This is not an artefact of our data but a characteristic of naturally occurring compounds (cf. e.g. Brown 2006, Bat-El 2006). What the presence of many local analogies in our analogical model shows, then, is that constituent family information adds to semantic information by restricting the size of analogical sets.

5.2.2 Mid-size analogical sets

Closer inspection reveals that the analogical sets in our model that have between 13 and 43 members are characterised by a specific type of gang behaviour, in which the same gangs of exemplars consistently reappear as a group in a number of analogical sets, in which they constitute the dominant group of exemplars to determine stress classification. A gang (cf. especially Skousen 1995, Skousen 2002a, Skousen 2002b, Skousen 2005) is a group of exemplars within an analogical set which all have a number of features in common. One of the most interesting features of an analogical model is that there are gangs that tend to appear and reappear in a substantial number of analogical sets [5]. Such behaviour leads to effects that, on the surface, appear rule-governed. However, such ‘rule behaviour’ is epiphenomenal in AM.

There are four main gangs that are active in the mid-size analogical sets in our model:

1. a gang of compounds sharing the feature ‘N2 is N1’
2. a gang of compounds in which the relation is ‘N1 is an ingredient in foodstuff NN’
3. a gang of compounds in which N1 denotes a material and the relation between N1 and N2 is ‘N2 is made of N1’
4. a gang of compounds in which N1 denotes a material but in which no rightward-leaning relation between N1 and N2 applies.

Each of these gangs is based on one or more semantic features. The gang characterised by the feature ‘N2 is N1’ (gang no. 1 above) comprises a group of 13 compounds, 10 of which are right-stressed and 3 of which are left-stressed. These 13 compounds reappear together as a gang in 10 analogical sets. In (1) we provide a list of the gang for illustration.

(1)	arrángements bit	miracle báby
	baby bóy	óldies team
	bastard éngine	surprise présent
	bastard téacher	sýcamore tree
	bitch téacher	toy cúps
	imitation trées	washer dríers
	minimum áge	

The gang of compounds in which N1 is a food ingredient (gang no. 2 above) comprises a group of 21 right-stressed and 2 left-stressed compounds which reappear in three analogical sets. Examples of gang members are *chocolate buíscit*, *rice kríspies*, *chícken burger*, and *mango sórbet*. The three test compounds for which these food compounds act as a gang are *apple tart*, *fruit sundae*, and *marzipan bar*. In the gang in which N1 is a material and the relation between N1 and N2 is ‘N2 is made of N1’ (gang no. 3 above) we find 33 right-stressed compounds and one left-stressed compound. Examples are *iron lég*, *gold jéwellery*, *cotton shéet*, and *metal ríng*. They act as a gang for six test compounds, namely for *bread roll*, *brick wall*, *denim jackets*, *junk earrings*, *silicone chips*, and *cotton buds*. Finally, the gang in which N1 is a material but no rightward-leaning relation between N1 and N2 applies (gang

[5] Note that, strictly speaking, the gangs do not reappear as fully identical sets because in our experimental setup the test compound is never part of the analogical set and, hence, never part of the gang. Thus, there are small variations in the way gangs are composed.

no. 4 above) comprises 31 left-stressed compounds and 2 right-stressed compounds. Examples of gang members are *pízza oven*, *ápple cart*, *téa rooms*, *gás canister*, and *jacket potáto*. They act as a gang for 15 test compounds, including *bread queue, apple cart, egg yard*, and *sandwich box*: many of the compounds for which this gang is active are compounds where N1 is a food item, but the relation between N1 and N2 is some sort of functional relation rather than an ingredient relation.

It is obvious how the behaviour of the four gangs described above can lead to the impression that the behaviour observed in the test compounds is rule-governed. For the first three gangs, we could write rules that the relations ‘N2 is N1’, ‘N1 is an ingredient in foodstuff NN’ and ‘N2 is made of N1’, respectively, lead to right stress. In the case of the fourth gang, a rule could be postulated, for example, that the relation ‘N2 is for N1’ leads to left stress. The 1-3 compounds in each gang that display the ‘illegal’ stress patterns would then be listed as exceptions. Such rules have indeed been suggested in the literature, together with lists of exceptions (cf. e.g. Fudge 1984). In an analogical model, however, rules do not exist, and their apparent existence is epiphenomenal of the fact that pertinent compounds in the lexicon reappear as a gang in the analogical sets of a number of test compounds. Also, note that even if the model's behaviour may resemble that of a rule in some cases, there are significant differences between the two generalisation mechanisms that have important implications for the prediction of variation.

As an example, consider the test compound *brick wall*. Its analogical set, graphically represented in Figure 4, contains 36 compounds, of which 33 compounds belong to the material gang described above (gang no. 3). We see that, in addition to the material gang, the analogical set for *brick wall* also contains three exemplars which share its N2 lemma, but differ from it in terms of both semantic features. Even though they share only one feature with the target compound, these exemplars are included in the analogical set because there are no other exemplars that share this feature and are more similar to the target, *brick wall*. It happens that these three exemplars are right-stressed, so that their inclusion in the analogical set leads to an even higher predicted probability of right stress for *brick wall* than that based on the material gang alone. In other compounds, similar constellations lead to slightly different probabilities of right stress. A case in point is *silicone chip*, whose analogical set includes left-stressed *sóund chip* in addition to the material gang. In this case, *silicone chip* is still predicted to be right-stressed, but with a lower probability than *brick wall*. It can be seen that, although the constituent family gang for *brick wall* includes only right-stressed exemplars, whereas the constituent family gang for *silicone chip* includes only a left-stressed exemplar, this makes very little difference to the model’s overall predictions for these compounds. This is because of the inclusion of the large material gang in the relevant analogical sets. Whenever a large gang appears in an analogical set, the weight given by AM to gang size means that the large gang will tend to dominate the model’s predictions. This is what leads to the illusion of rule-governed behaviour.

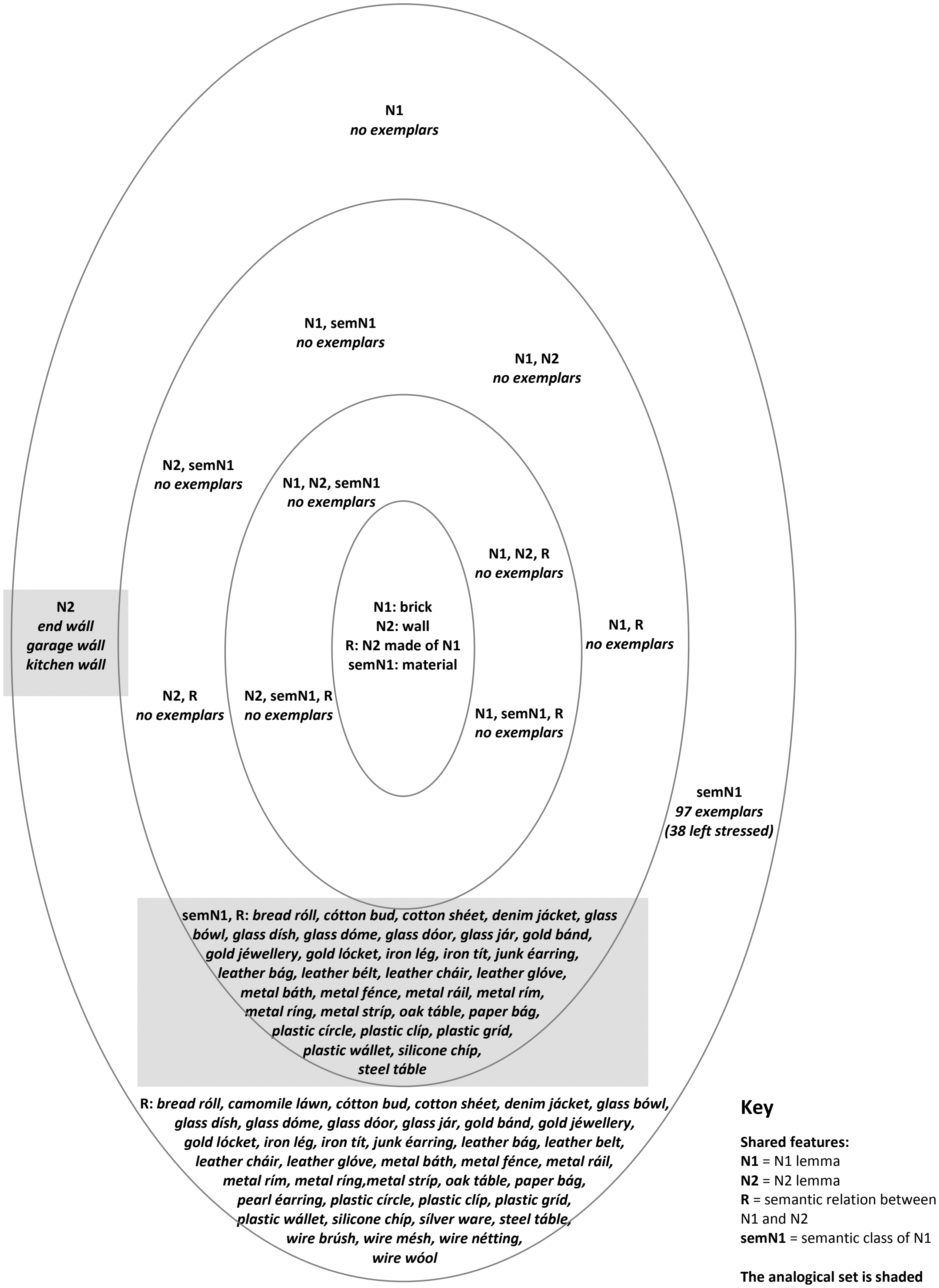

Figure 4: The analogical set for *brick wall* (1065 rightward votes; 33 leftward votes)

The example of *brick wall* and *sound chip* thus shows that the emergence of generalisations in the 'mid-size' analogical sets differs significantly from the behaviour of traditional linguistic rules in a number of ways. One difference between analogical and rule-based mechanisms concerns the question of how deterministic linguistic generalisations are. Linguistic rules are usually conceived to be fully deterministic, in the sense that they make a deterministic prediction for contexts in which they apply. In an analogical model, on the other hand, deterministic behaviour is the exception rather than the rule. In our model, for example, deterministic behaviour can only emerge if the lexicon contains a group of highly similar compounds that exhibit fully consistent stress behaviour, and that act as a gang for test compounds whose analogical sets show no evidence of the competing stress pattern. In our simulation, this situation never actually occurs among compounds with mid-size analogical sets. None of the gangs that lead to mid-size analogical sets is fully deterministic. The gangs that show the most nearly-deterministic stress behaviour are those that involve material compounds (gangs 3 and 4 in the list above). But even in these gangs, we find individual compounds that do not conform to the majority stress pattern. In sum, both between-type and within-type variation is inherently expected in an analogical system, where it emerges as a result of the analogical mechanism acting on items in the lexicon. This constitutes a radical departure from key assumptions made in especially generative frameworks, where the lexicon, but crucially not the rule system, is a repository of such variation.

In Section 5.1, we saw that AM can be set to make categorical predictions on the basis of a majority vote in the analogical set. However, we have also seen that classification in AM is underlyingly probabilistic, with the probability of a particular outcome being equal to the proportion of the vote that favours that outcome. This probabilistic classification can be interpreted in two ways: it can be interpreted as the level of certainty with which the 'winning' outcome is predicted, or it can be interpreted as the predicted degree of variation within a new type: in our case, the likelihood of any given token receiving one stress pattern or the other. A way of modelling this second possibility is to set AM's parameters so that the target feature is based on an exemplar chosen at random from the analogical set. Provided the probability of selecting a given exemplar reflects the weighting of votes produced by the algorithm, then over time, the distribution of stresses produced in tokens of any given type will reflect the probabilities in the model. This is how AM produces within-type variation. Between-type variation then arises as an inevitable consequence of the same mechanism: since the predictions for different types are based on different analogical sets, the probabilities of right or left stress will differ between types and so, therefore, will the distributions of stress patterns in the output. For some types, the probability of right stress will be almost 100%, in which case nearly all tokens will be produced with right stress. For other types, the probability of right stress will be close to zero, in which case nearly all tokens will be produced with left stress.

Another difference between a rule-based model and the generalisations that emerge from the mid-size analogical sets in our model is that, unlike linguistic rule application, the activity of pertinent gangs in an analogical system is not obligatory once a certain context is met. It is a striking characteristic of all four gangs listed above that they are not active in all test compounds with the relevant semantic feature(s). For example, gang no. 1 (based on the semantics ‘N2 is N1’) comprises 13 compounds but applies to only 10 test compounds, in spite of the fact that in our experimental setup the lexicon and the group of test compounds are identical. Even more radically, gang no. 3 (based on the semantics ‘N1 is a material’ and ‘N2 is made of N1’) comprises 34 compounds, but applies only to 6 test compounds. The reasons for these discrepancies lie in the distribution of salient features amongst items in the lexicon. As an example, we consider here the activity of gang no. 3. The 34 compounds that constitute the gang are listed in (2). The 6 test compounds to which the gang applies are shaded.

(2)	bread roll	glass domes	leather belt	metal strip
	brick wall	glass door	leather chair	oak tables
	cotton buds	glass jar	leather glove	paper bags
	denim jackets	gold band	metal baths	plastic circles
	junk earrings	gold jewellery	metal fence	plastic clips
	silicone chips	gold locket	metal rails	plastic grid
	cotton sheets	iron legs	metal rim	plastic wallet
	glass bowls	iron tits	metal ring	steel table
	glass dish	leather bag		

In the analogical sets of 28 of its members, gang no. 3 is not active. This is not due to the semantics of these 28 compounds because they all share both coded semantic features with the 6 compounds for which the gang is active. Instead, it is due to the fact that, for these 28 cases, the lexicon contains exemplars which are more similar to the target compound than the gang compounds are, and which, crucially, provide the system with more certainty about the stress prediction for the target compound than the gang compounds would.

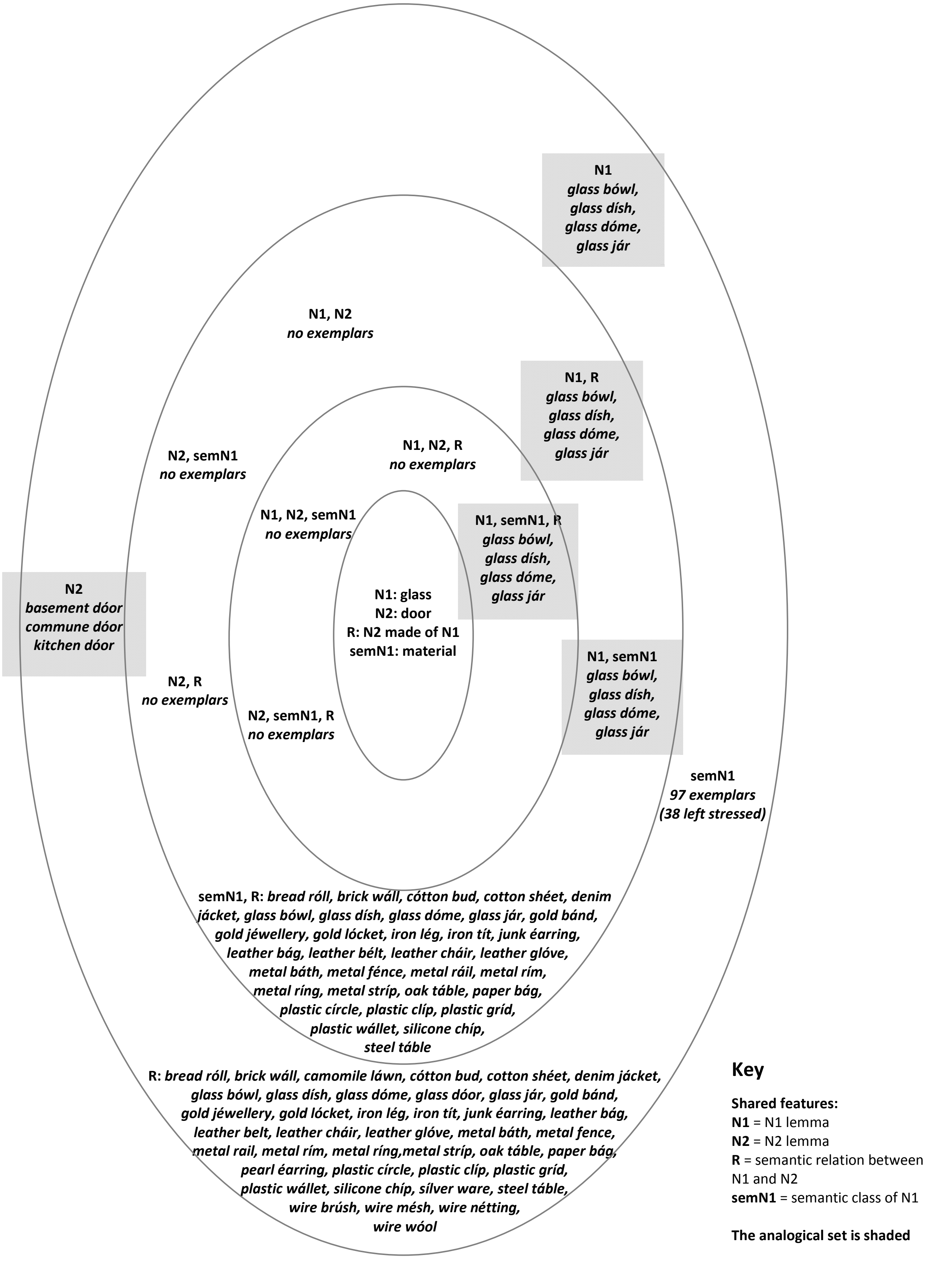

Figure 5: The analogical set for *glass door* (73 rightward votes; 0 leftward votes)

As an example, we consider the test compound *glass door*, for which the analogical set is illustrated in Figure 5. The situation for *glass door* is different than for the 6 compounds where gang no. 3 is active (compare *brick wall* in Figure 4 above), in that the lexicon contains exemplars which share both the semantics and a constituent with the target, and which, unlike the gang compounds, have deterministic stress. All exemplars that share both N1 lemma and two semantic features with *glass door* are right stressed. Including those exemplars that share only the semantics would therefore reduce the certainty of the stress prediction along this dimension, since the 'N1 is a material and N2 is made of N1' gang contains one exemplar that is left stressed. As a consequence, the system does not incorporate the gang compounds into the analogical set. The implication, then, is the following: behaviour that appears to be the effect of a general rule emerges only if it is compatible with more local analogies. Again, this is contrary to many conceptions in rule-based frameworks, where local analogies are conceptualised as exceptions whose motivation is often unclear.

5.2.3 Large analogical sets

For 26 of our test compounds the analogical set comprises between 174 and 176 items. Similar to what we found for the mid-size analogical sets, the large analogical sets in our model are the result of gang behaviour: a core group of 174 exemplars appears as a gang in all 26 of the large analogical sets, determining stress for the relevant test compounds.

In (3) we provide some examples of the 174 exemplars that make up the large-set gang. Compounds for which these 174 exemplars appear in the analogical set are shaded. Within this gang, 166 (95.40%) of the 174 exemplars are left-stressed, and 8 exemplars (4.60%) are right-stressed.

(3)

árm bands	alárm business	bóot bag	cóffee jars
asséssment piece	antíques day	bráin work	colléction card
bégging bowl	árt centre	búbble bath	convénience foods
bíngo money	áttache case	canóe bag	decísion time
cámping holiday	báby stuff	car rádios	scíence grade
cónker trees	bánking job	caréer woman	spórts centre
frúit stalls	bíke things	cárrot maps	télephone woman
héad phones	bín day	cát foods	tráffic light
square róot	bóarding schools	chícken pox	tráin station

All compounds in the large gang are characterised by the fact that neither the semantic relation nor the semantic class of N1 belongs to any of the rightward-leaning categories coded. Thus, the large analogical sets differ from the mid-size analogical sets in that the exemplars in the large sets are semantically more heterogeneous. What they have in common is that they do not fall into any of the categories that trigger right stress. Thus, on the surface, the generalisation that emerges from these sets looks very much like the situation that is often described as a default in rule-based frameworks. In an analogical model, this situation emerges from cases where the analogy drawn is, in a sense, the most non-local type of analogy conceivable (cf. Derwing & Skousen 1994, Eddington 2000 on inflection).

Another striking feature of the large analogical sets is that, as with the mid-size sets, the relevant gang of exemplars applies to far fewer compounds than it contains. As discussed in section 5.2.2, the reason lies in the fact that for those compounds for which the gang does not play a role, a more local analogical set makes more certain predictions with respect to stress assignment. In the case of the large analogical sets, the more local set always involves members of a constituent family: since both semantic features are already matched in the large gang, the only way another exemplar can be more similar to the target (i.e. more local) is by also sharing a constituent with it.

As an example, we will compare AM's predictions for the test compounds *begging bowl* and *weather man*. Both compounds are in the large analogical gang. However, the large gang is relevant for stress assignment for *begging bowl*, but not for *weather man*. The reason is that for *weather man* there exists a more local set of exemplars which are a subset of the gang of 174, but share a constituent with the target in addition to both the semantic features (or rather lack of them). The members of this smaller gang are all left stressed: *bándman*, *chéese man*, *cón man*, *fíreman*, *jám man* and *shówman*, whereas the large gang includes exemplars with both stress patterns. The stress prediction based on just the small gang is therefore more certain than it would be if the larger gang were included. Since the small gang is also more similar to the target, along the same dimension as the large gang, the latter is excluded. The situation for the test compound *begging bowl* is different. Within our model's lexicon, only the exemplars *glass bówl* and *púdding bowl* share a constituent with *begging bowl*. Neither of these are members of the gang of 174, since they each have at least one rightward-leaning semantic feature: for *púdding bowl* 'N1 is a material' and for *glass bówl* both 'N1 is a material' and 'N2 is made of N1'. There are therefore no exemplars in the lexicon that share three features with *begging bowl*; in other words, no exemplars are more similar to *begging bowl* than the large gang, with which it shares the two semantic features. AM therefore includes the large gang in the analogical set for *begging bowl*.

The existence of the large gang in our model is interesting because its members are united not by the presence of particular features but rather by their absence: specifically, by the absence of any rightward-leaning semantic category. This shows that information about both the presence of features and their absence can be used by an analogical mechanism for the purposes of analogical set formation.[6]

6. Study 2: Predicting within-type variation

In section 5, we saw how AM predicts both between-type and within-type variability in compound stress, but for ease of exposition we restricted our lexicon to types that showed no variation in Bell's (2013) reading experiment. As discussed in section 2.2, however, compound stress in general does display within-type variability, and a more realistic model would therefore include such variation in the lexicon. Furthermore, in order to assess the success of our model, we transformed AM's probabilistic output into a categorical decision, assigning either left or right stress to the item to be classified on the basis of a majority vote in the analogical set. This meant that we could assess AM's predictive power for between-type variability but not for within-type variability. In this section, we therefore present a simulation in which we use AM to predict within-type variation in English compound stress on the basis of a more realistic representation of the lexicon as including within-type variability. We use this model to demonstrate that AM is highly successful in predicting not only between-type variation but also variation within types.

6.1 Within-type variation in the lexicon

Within-type variability is present in our data because each compound was experimentally elicited from four different speakers, and in many cases the four speakers did not all produce a given compound with the same stress pattern. We can therefore represent within-type variability in the lexicon of AM by storing individual tokens, rather than types. The dataset to be used is again a subset from Bell (2013), comprising all compounds where either the first or the second constituent has a constituent family in the dataset, but neither constituent is itself a compound and the coded semantic features are not obviously ambiguous. This dataset contains 1,913 tokens representing 486 different types. Of these, 293 (49%) were left-stressed by all four readers, 147 (30%) were right-stressed by all four readers, and 100 (21%) were produced with both stress patterns. Amongst the variable types, we include compounds for which Bell's dataset contains only two or three tokens with agreed stress, provided these tokens include at least one with each stress pattern.

[6] The fact that the absence of pertinent semantic features is indeed useful information for the model is further corroborated by the fact that excluding this information leads to a loss of predictive power. In a series of experiments we manipulated the settings of the algorithm in such a way that absence of a semantic feature could not be used to compute similarity. In these experiments, AM predictions were consistently worse than in experiments where absence of a feature could be used as an information source (cf. Parkinson (2002) for an explanation of the relevant parameter settings).

In order to simulate within-type variability, we use the 1,913 compound tokens as the lexicon and predict stress for one token of each of the 486 types. Using the set of tokens as the database means that most compound types are represented in the lexicon four times. Compounds that showed no within-type variation in the reading experiment are represented by four tokens with the same stress pattern, whereas compounds that did display within-type variation are represented by at least one token with each pattern. An example is *police control*, which was realised with left stress by two speakers and with right stress by the other two speakers. For testing, AM parameters were set in such way as to make sure that for a given test item, all tokens of the same compound type were excluded from consideration. The rationale behind this was that, for better assessment of the model's predictive power, we simulated a situation where each test compound was a novel compound.[7] The information sources used in the experiment were identical to those in the experiments discussed in section 5, comprising the constituent family or families and right-predicting semantic classes and relations.

To understand how the model works, we consider the analogical set for the compound *police helmets*, graphically represented in Figure 7. Both between-type and within-type variation are included in the lexicon. As an example of between-type variation, *polı́ce car* is represented by four left-stressed tokens, whereas the compound *government pólicy*, which shares both its semantic features with *police helmets*, is represented by four right-stressed tokens. An example of within-type variation is *police control*, which is represented by two left-stressed and two right-stressed tokens. Both kinds of variability will be involved in determining which exemplars enter the analogical set for the target word *police helmets*, as well as the distribution of stresses predicted by this set.

The token-based lexicon contains 12 exemplars that are maximally similar to the target, each sharing the N1 lemma as well as both semantic features. Of these exemplars, 10 are left-stressed while 2 (two tokens of *police control*) are right-stressed. No further exemplars, that could affect the probability of the outcome, share the N1 lemma, and so this gang occurs in the analogical set in four contexts. Although there are many other exemplars in the lexicon that share one or both semantic features with the target, these other gangs include a higher proportion of right-stressed tokens and their inclusion in the analogical set would

[7] In terms of usage-based models of grammar, note that a common assumption is that stored tokens of the same type will have an influence on the classification of novel types (cf. e.g. Bybee (2001), Bybee (2010)). In terms of an assessment of predictive power in AM, however, including tokens of the test type in the lexicon renders the experiment less interesting and convincing, for two reasons. The first is that, because an exemplar's weight in the vote within the analogical set is correlated with its degree of similarity with the new item, tokens of the same type will often come to dominate analogical sets. The second is that our dataset, which contains a maximum of four tokens for each type, cannot be used to realistically represent differences in token frequencies between lexical types in the lexicon.

therefore reduce the certainty of the model's prediction: because of this, they are excluded. The vote is therefore based just on the 12 tokens with N1 = police, which each receive 48 votes. Since only one two of these tokens are right-stressed, there are 480 votes for left stress (10 x 48) and 96 votes for right stress (2 x 48). The model therefore predicts 83.33% left stress for *police helmets* (480/576). This prediction closely reflects the empirical facts, since 3 of the 4 tokens of *police helmets* in the data, i.e. 75%, were produced with left stress.

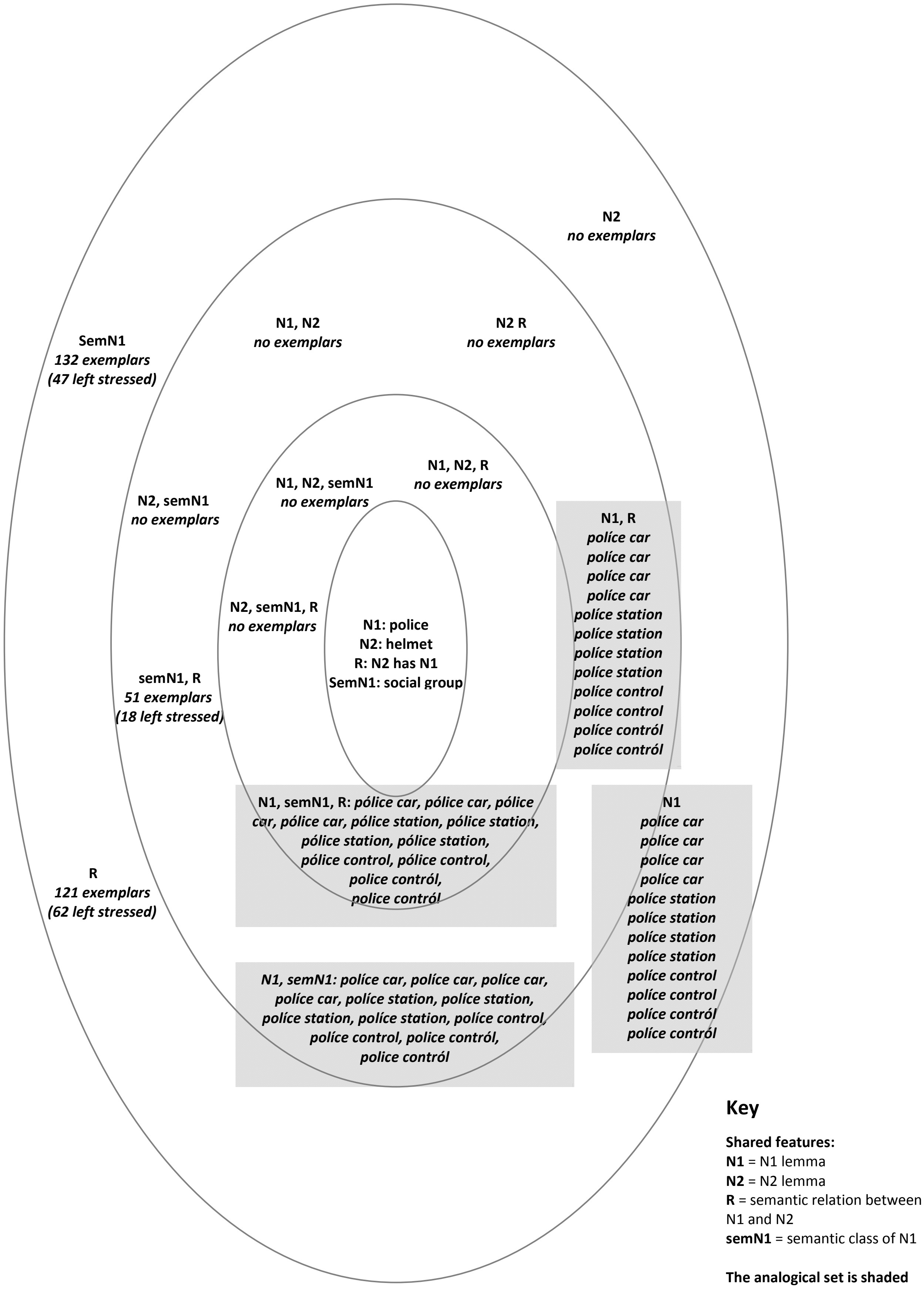

Figure 7: The analogical set for *police helmets* (480 leftward votes; 96 rightward votes)

6.1 Overall performance of Model 2

To investigate how accurately AM predicts within-type variability on the basis of this token data, we divided our compound types into three groups: those that were invariably left-stressed by Bell's (ibid.) subjects (N = 239), those that were invariably right-stressed (N = 147), and those that were produced with both patterns (N = 100). For each of these groups, we plotted the variation in the percentage of right-stressed tokens, as predicted by AM. If the model is successful at predicting within-type variation, then the variable group should show the greatest variation in predicted percentage of right stress, whereas the left and right-stressed groups should ideally show respectively 0% and 100% predicted right stress, with little variation. The results are shown in the box and whisker plot in Figure 6.

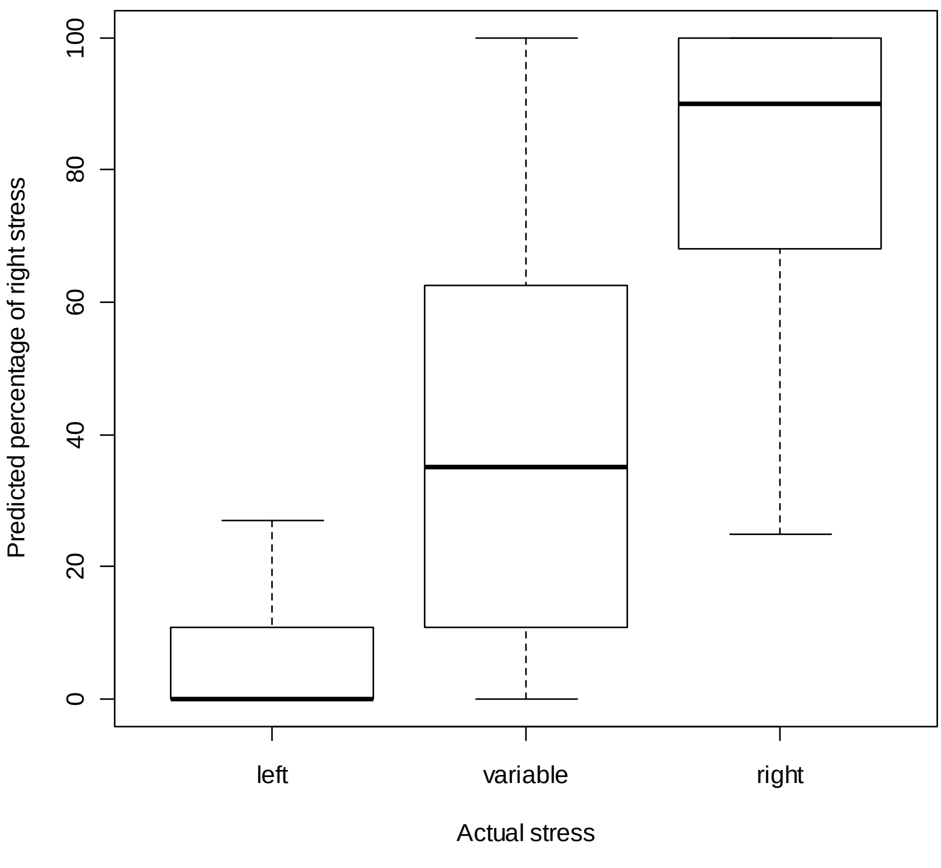

Figure 6. Predicted vs. observed within-type variability in Study 2 (N = 486)

The figure shows very clear differences between the three stress groups, in ways expected if the model is successful. For compounds that are consistently left-stressed in our corpus, the median of the predicted percentage of right stress is 0, whereas for compounds with variable stress it is 35% and for those that are consistently right-stressed it is 90%. Predicted within-type variability for the left-stressed items is rather limited: the length of the whiskers in the boxplot for this group of compounds indicates that for some 75% of the data the probability of right stress is predicted to be less than 30%. Predicted within-type variability for the right-stressed group is somewhat larger, but there is nevertheless little overlap with the left-stressed group in terms of predicted percentage of right stress. In contrast, the variable stress group has a much greater range of predicted probabilities than either of the other two groups. A Kruskal-Wallis rank sum test shows that the differences between the medians of the three groups are highly significant (chi-squared = 249.34, df = 2, p-value < 2.2e-16), and a Fligner-Killeen test of homogeneity of variances shows that the three groups also differ significantly in their degree of variability (chi-squared = 66.89, df = 2, p-value = 2.983e-15).

Overall, these plots and figures show that within-type variability is not only predicted by the model, but is predicted to occur significantly more often with compound types which are indeed variable in our database than with compounds which are not variable. It also appears from these results that the prediction of variability in an analogical model constitutes not only a virtue, but also a downside of such a model. Specifically, within-type variability is overpredicted, i.e. it is predicted to occur for compound types which are not in fact variable in this dataset. However, this may well be a result of the fact that our database only contained four compound tokens of the same type (and in some cases even fewer than that). This small number may have been insufficient to capture the full range of variability so that what appears to be overprediction of variability may, in fact, be a reflection of the full range of variability.

7. Summary and conclusion

Any mechanism that can produce accurate linguistic output is, in effect, a grammar – and we would argue that a constrained analogical algorithm such as AM represents such a grammar. The models presented in this paper have demonstrated that a single analogical mechanism can account for a variety of observed linguistic phenomena: for both rule-like almost categorical effects and for more variable ones, as well as for the relevance of different degrees of generality in linguistic generalisation. Furthermore, being constrained by a clearly defined mechanism for calculating similarity, AM has a restricted set of potential analogs for any given test item; yet at the same time, the set of potential bases for any test item is individually computed, so that even apparently similar items may have different analogical sets. The criticism of analogical models that they are unconstrained is therefore not true in a case like this, although the diversity of possible analogs may give this impression.

In section 5 we saw how three different types of behaviour emerge in the analogical model of English compound stress: behaviour that is based on local analogies involving the constituent family, behaviour that is based on larger sets of analogical gangs which are formed by semantically similar compounds, and behaviour that is based on an analogical set that comprises a considerable part of the lexicon, whose members have in common that none of the right-favouring semantic features applies. In much of the rule-based literature, these three types of behaviour have been assigned to different mechanisms: exceptional analogical formations, rule-based behaviour, and a default case. In this paper we have shown that they can all emerge from a single mechanism. Our findings thus add support to the growing body of literature that argues that apparent differences in types of linguistic patterns can be interpreted as epiphenomena of differences in lexical distributions and accounted for with a single underlying mechanism (cf. esp. Derwing & Skousen 1994, Eddington 2000, Keuleers 2008 on inflection).

Variability in the output, both between and within test types, is an expected consequence of an analogical mechanism operating on a lexicon that itself includes variation. In Section 5, we saw that AM is highly successful at replicating the between-type variation produced by human subjects, and in Section 6, we saw that the extent of within-type variation produced by the algorithm also reflects the actual variability found in the human data. The question therefore arises as to whether human grammars might also be analogical in nature. As discussed in the introduction, analogical reasoning is extremely pervasive in human cognition, and the hypothesis that this pervasiveness encompasses language is worthy of serious consideration.

One question that arises in relation to this hypothesis is how linguistic variability might be represented in the mental lexicon. In the lexicon of an AM model, the only technically viable option is to represent tokens, but for the mental lexicon it is possible to conceive of this variation in different ways: as an unstructured representation of individual tokens, as an exemplar cloud, or as a type-based representation where stress is probabilistic. We work from the basic assumption that individual exemplars, i.e. tokens, are initially stored, but that speakers may subsequently abstract away from these tokens to produce a type-based representation. We further assume that this more abstract representation is associated with information about the degree of variability amongst the underlying tokens: in our case, the degree of variability in stress placement. Thus, for example, the coding of the compound types in the lexicon of exemplars in Figures 1, 2, 4 and 5 as left- or right-stressed actually constitutes an abstraction of the fact that our model speaker has evidence that stress for these compounds is non-variable.

A second issue is how an analogical human grammar would produce within-type variation. We have seen that AM does this by selecting exemplars at random from the analogical set; in the case of human language production, we envisage that any such mechanism would be subject to the influence of contextual and other pragmatic factors.

An analogical grammar is fundamentally different from a rule-based grammar in being 'bottom up' rather than 'top down': in an analogical model, general patterns emerge from the specific examples of language encountered; in rule-based models the direction is reversed, with specific examples of language being based on abstract rules. Rules tend to map inputs onto outputs. To give an example involving English compound stress: if the compound has the semantic relation 'N2 is made of N1' (input), then produce it with right stress (output). In contrast, an analogical grammar always submits the input to the same mechanism, in the context of a particular lexicon. The outcome is not pre-determined, but depends on the composition of the lexicon. Apart from its empirical success, one of the great appeals of an analogical grammar is therefore its consistency: although outputs vary, as they do in natural language, the central mechanism is invariant.

References

Albright, Adam. 2002. Islands of reliability for regular morphology: evidence from Italian. *Language* 78(4). 684–709.

Albright, Adam. 2009a. Feature-based generalisation as a source of gradient acceptability. *Phonology* 26. 9–41.

Albright, Adam. 2009b. Modeling analogy as probabilistic grammar. In James P. Blevins & Juliette Blevins (eds.), *Analogy in Grammar*, 185–213. Oxford: OUP.

Albright, Adam & Bruce Hayes. 2003. Rules vs. Analogy in English Past Tenses: A Computational/Experimental Study. *Cognition* 90. 119–161.

Anttila, Raimo. 2003. Analogy: the warp and woof of cognition. In Brian D. Joseph & Richard D. Janda (eds.), *The Handbook of Historical Linguistics*, 425–440. Malden, MA: Blackwell.

Arndt-Lappe, Sabine. 2011. Towards an exemplar-based model of stress in English noun–noun compounds. *Jounal of Linguistics* 47(11). 549–585.

Arndt-Lappe, Sabine. in press. 47. Word-formation and analogy. In Peter O. Müller, Ingeborg Ohnheiser, Susan Olsen & Franz Rainer (eds.), *Word-Formation - An International Handbook of the Languages of Europe*. Berlin: de Gruyter Mouton.

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

Bauer, Laurie. 1983a. *English Word-Formation*. Cambridge: CUP.

Bauer, Laurie. 1983b. Stress in compounds - a rejoinder. *English Studies* 64. 47–53.

Bauer, Laurie. 2001. *Morphological Productivity*. Cambridge: CUP.

Becker, Thomas. 1990. *Analogie und morphologische Theorie* (Studien zur theoretischen Linguistik). München: Fink.

Bell, Melanie. 2013. *The English Noun Noun Construct: Its Prosody and Structure*. Cambridge: Ph.D. dissertation, University of Cambridge.

Bell, Melanie & Ingo Plag. 2012. Informativeness is a determinant of compound stress in English. *Jounal of Linguistics* doi:10.1017/S0022226712000199.

Bell, Melanie & Ingo Plag. 2013. Compound stress, informativity and analogy. *Word Structure* in press.

Blevins, James P. & Juliette Blevins. 2009. Introduction: Analogy in grammar. In James P. Blevins & Juliette Blevins (eds.), *Analogy in Grammar*, 1–12. Oxford: OUP.

Bloomfield, Leonard. 1933. *Language*. New York: Holt, Rinehart & Winston.

Bod, Rens, Jennifer Hay & Stefanie Jannedy (eds.) (2003). *Probabilistic Linguistics*. Cambridge, MA: MIT Press.

Booij, Geert. 2010. *Construction Morphology*. Oxford: Oxford University Press.

Bresnan, Joan. 2007. Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In Sam Featherston & Wolfgang Sternefeld (eds.), *Roots: Linguistics in Search of Its Evidential Base*, 77–96. Berlin: Mouton de Gruyter.

Bybee, Joan. 2001. *Phonology and Language Use*. Cambridge: CUP.

Bybee, Joan. 2010. *Language, Usage and Cognition*. Cambridge: CUP.

Bybee, Joan & Carol L. Moder. 1983. Morphological classes as natural categories. *Language*(59). 251–270.

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. 1975. *The Logical Structure of Linguistic Theory*. Chicago: Chicago University Press.

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

Coetzee, Andries W. 2009. Grammar is both categorical and gradient. In Steven Parker (ed.), *Phonological Argumentation: Essays on Evidence and Motivation*, 9–42. London: Equinox.
Coetzee, Andries W. & Joe Pater. 2011. The place of variation in phonological theory. In John Goldsmith, Jason Riggle & Alan C. L. Yu (eds.), *The Handbook of Phonological Theory*, 401–434. Malden, MA: Wiley-Blackwell.
Daelemans, Walter & Antal d. van Bosch. 2005. *Memory-Based Language Processing*. Cambridge: CUP.
Daelemans, Walter, Antal van den Bosch & Jakub Zavrel. 1999. Forgetting exceptions is harmful in language learning. *Machine Learning* 34. 11–41.
Daelemans, Walter, Jakub Zavrel, Ko van der Sloot & Antal van den Bosch. 1999 et seq. *TiMBL: Tilburg Memory Based Learner* : Available from http://ilk.uvt.nl/timbl/.
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.
Day, Simon & Dedre Gentner. 2007. Nonintentional analogical inference in text comprehension. *Memory and Cognition* 35. 39–49.
Derwing, Bruce I. & Royal Skousen. 1989. Morphology in the mental lexicon: a new look at analogy. In Geert Booij & Jaap van Marle (eds.), *Yearbook of Morphology 1989*, 55–71. Dordrecht: Foris.
Derwing, Bruce I. & Royal Skousen. 1994. Productivity and the English past tense: Testing Skousen's analogical model. In Susan D. Lima, Roberta Corrigan & Gregory K. Iverson (eds.), *The Reality of Linguistic Rules*, 193–218. Amsterdam / Philadelphia: John Benjamins.
Eddington, David. 2000. Analogy and the dual-route model of morphology. *Lingua* 110. 281–298.
Eddington, David. 2002. Spanish diminutive formation without rules or constraints. *Linguistics* 40(2). 395–419.
Eddington, David. 2004. *Spanish Phonology and Morphology: Experimental and Quantitative Perspectives*. Amsterdam / Philadelphia: John Benjamins.
Ernestus, Mirjam. 2011. Gradience and categoricality in phonological theory. In Marc van Oostendorp, Colin Ewen, Elizabeth Hume & Curt Rice (eds.), *The Blackwell Companion to Phonology*, 2115–2136. Oxford: Wiley-Blackwell.
Faiß, Klaus. 1981. Compound, pseudo-compound, and syntactic group especially in English. In Peter Kunsmann & Ortwin Kuhn (eds.), *Weltsprache Englisch in Forschung und Lehre: Festschrift für Kurt Wächtler*, 132–150. Berlin: Schmidt.
Fudge, Erik C. 1984. *English Word-Stress*. London: George Allen & Unwin.
Gahl, Susanne & Alan C. L. Yu. 2006. Introduction to the special issue on exemplar-based models in linguistics. In Susanne Gahl & Alan C. L. Yu (eds.), *Special Issue on Exemplar-based Models in Linguistics* (23), 213–216.
Gentner, Dedre & L. Smith. 2012. Analogical reasoning. In Vilanayur S. Ramachandran (ed.), *Encyclopedia of Human Behavior (2nd Ed.)*, 130–136. Oxford: Elsevier.
Giegerich, Heinz. 2004. Compound or phrase? English noun-plus-noun constructions and the stress criterion. *English Language and Linguistics* 8(1). 1–24.
Giegerich, Heinz. 2009. The compound stress myth. *Word Structure*(2). 1–17.
Goldberg, Adele. 2006. *Constructions at Work - the Nature of Generalization in Language*. Oxford: OUP.
Goldberg, Adele & Ray Jackendoff. 2005. The end result(ative). *Language* 81(2). 474–477.
Hay, Jennifer & Harald R. Baayen. 2005. Shifting paradigms: gradient structure in morphology. *Trends in Cognitive Science* 9. 342–348.
Itkonen, Esa. 2005. *Analogy as Structure and Process*. Amsterdam: John Benjamins.

Keuleers, Emmanuel. 2008. *Memory-Based Learning of Inflectional Morphology*. PhD. dissertation, Antwerpen: University of Antwerp.
Kim, John, Stephen Pinker, Alan Prince & Sandeep Prasada. 1991. Why no mere mortal has ever flown out to center field. *Cognitive science* 15(2). 173–218.
Krott, Andrea. 2009. The role of analogy for compound words. In James P. Blevins & Juliette Blevins (eds.), *Analogy in Grammar*, 118–136. Oxford: OUP.
Krott, Andrea, Rob Schreuder, Harald R. Baayen & Wolfgang U. Dressler. 2007. Analogical effects on linking elements in German compounds. *Language and Cognitive Processes* 22. 25–57.
Kunter, Gero. 2007. *Within-speaker and Between-speaker Variation in Compound Stress Assignment.* Paper presented at the Second International Conference on the Linguistics of Contemporary English ICLCE2, July 2-4, Université de Toulouse.
Kunter, Gero. 2009. *The Phonetics and Phonology of English Compound Stress*. Ph.D. dissertation, Universität Siegen.
Kunter, Gero. 2011. *Compound Stress in English - The Phonetics and Phonology of Prosodic Prominence* (Linguistische Arbeiten 539). Berlin: de Gruyter Mouton.
Kunter, Gero & Ingo Plag. 2009. *Constituent families and compound stress assignment in English*. Paper presented at the 3rd International Conference on the Linguistics of Contemporary English (ICLCE 3), London, July 15-17 2009.
Ladd, D. R. 1984. English compound stress. In Dafydd Gibbon & Helmut Richter (eds.), *Intonation, Accent and Rhythm*, 253–266. Berlin: Mouton de Gruyter.
Lappe, Sabine & Ingo Plag. 2007..The Variability of Compound Stress in English: Towards an Exemplar-Based Alternative of the Compound Stress Rule. Proceedings of the ESSLLI workshop on exemplar-based models of language acquisition and use. Dublin, Ireland, August 2007, 8-23.
Lees, Robert B. 1963. *The Grammar of English Nominalizations*. The Hague: Mouton.
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.
Maguire, Phil, Edward J. Wisniewski & Gert Storms. 2010. A corpus study of semantic patterns in compounding. *Corpus Linguistics and Linguistic Theory* 6(1). 49–73.
Marchand, Hans. 1969. *Categories and Types of Present-Day English Word-Formation*. München: C.H. Beck.
Marcus, Gary F. Stephen Pinker, Michael Ullman, Michelle Hollander, T. J. Rosen, Fei Xu & Harald Clahsen. 1992. *Overregularization in Language Acquisition* (Monographs of the Society for Research in Child Development). Hoboken, New Jersey: Wiley.
Miller, George, Richard Beckwith, Christiane Fellbaum, Derek Gross & Katherine Miller. 1990. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*(3). 235–244.
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.

Payne, John & Rodney Huddleson. 2002. Nouns and noun phrases. In Rodney Huddleson & Geoffrey K. Pullum (eds.), *The Cambridge Grammar of the English Language,* 323–524. Cambridge: CUP.
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.
Pinker, Stephen. 1991. Rules of language. *Science* 253. 530–535.
Pinker, Stephen & Alan Prince. 1988. On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition* 28. 73–193.
Pinker, Stephen & Alan Prince. 1991. Regular and irregular morphology and the psychological status of rules of grammar. In Laurel A. Sutton (ed.), *Proceedings of the 17th Annual Meeting of the Berkeley Linguistics Society*, 230–251. Berkeley: Berkeley Linguistics Society.
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). 243–282.
Plag, Ingo & Melanie J. Bell. 2012. Compound stress, informativity and analogy: submitted for publication.
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.
Prasada, Sandeep & Stephen Pinker. 1993. Generalization of regular and irregular morphological patterns. *Language and Cognitive Processes* 8. 1–56.
Sampson, Rodney. 1980. Stress in English N+N phrases: a further complicating factor. *English Studies* 61. 264–270.
Schmerling, Susan F. 1971. A stress mess. *Studies in the Linguistic Sciences* 1. 52–65.
Skousen, Royal. 1989. *Analogical Modeling of Language*. Dordrecht: Kluwer.
Skousen, Royal. 1992. *Analogy and Structure*. Dordrecht: Kluwer.
Skousen, Royal. 1995. Analogy: A non-rule alternative to neural networks. *Rivista di Linguistica* 7(2). 213–231.
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. 2005. Analogical Modeling (49). In Reinhard Köhler, Gabriel Altmann & Rajmund G. Piotrowski (eds.), *Quantitative Linguistik*: *Ein internationales Handbuch* = *Quantitative linguistics* (Handbücher zur Sprach- und Kommunikationswissenschaft = Handbooks of linguistics and communication science = Manuels de linguistique et des sciences de communication / mitbegr. von Gerold Ungeheuer. Hrsg. von Armin Burkhardt … ; Bd. 27), 705–716. Berlin: de Gruyter.
Skousen, Royal. 2009. Expanding Analogical Modeling into a general theory of language prediction. In James P. Blevins & Juliette Blevins (eds.), *Analogy in Grammar,* 164–184. Oxford: OUP.
Skousen, Royal & Thereon Stanford. 2007. *AM: Parallel* : available from http://humanities.byu.edu/am/.

Skousen, Royal, Deryle Lonsdale & Dilworth B. Parkinson (eds.) (2002). *Analogical Modeling*. Amsterdam / Philadelphia: John Benjamins.
Soskuthy, Marton. 2013. Analogy in the emergence of intrusive-r in English. *English Language and Linguistics* 17(1). 55–84.
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.
Sproat, Richard. 1994. English noun-phrase accent prediction for text-to-speech. *Computer Speech and Language* 8. 79–94.
Tomasello, Michael. 2003. *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Cambridge, MA: Harvard University Press.
van den Bosch, Antal & Walter Daelemans. 2013. Implicit schemata and categories in Memory-based Language Processing. *Language and Speech*. DOI: 10.1177/0023830913484902.