Arndt-Lappe, Sabine. 2014. 46 Word-formation and analogy.
To appear 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 [Handbücher zur Sprach- und Kommunikationswissenschaft]

47. Word-formation and analogy

Abstract

The article discusses the much-debated status of analogy in contemporary theories of synchronic word-formation. It provides an overview of the key assumptions made in pertinent theoretical camps as well as of the major phenomena that have featured prominently as evidence in the debate. Theories can broadly be classified into those which assume that analogy is active only as a complementary mechanism, and those which assume that analogy is the central mechanism of productive word-formation. Among the latter, we can distinguish between general theoretical and computational analogical models. Based on a detailed definition and discussion of the analogical equation, different usages of analogy in the literature are shown to be closely tied to theory-dependent conceptualisations of productivity, predictability, and (ir-)regularity.

1. Introduction

The term *analogy* is used in many different senses and in many different contexts within morphological theory. One subdiscipline with which analogy is frequently associated is diachronic morphology, where, ever since the Neogrammarian revolution, analogical change has been seen as a central mechanism of morphological change (cf., e.g., Hock 2003 for a summary). Developments which have come to be associated with analogy are, especially, analogical extension and levelling.

In synchronic morphological theory, the type of analogy that is the subject of discussion is proportional analogy, i.e. a heuristic mechanism in which a new complex word is formed on the basis of a perceived similarity with existing base-derivative pairs (cf. section 2. below for a more detailed definition). The central question that has been debated in the contemporary literature is whether analogy is an active mechanism in synchronic morphology, and, if it is, what its relation is to other mechanisms in synchronic morphology, such as rules or constraints or schemas. Much of this debate, which has its origins far back in the history of morphological research (cf. e.g. Becker 1990, Anttila 2003 for a summary), has taken place in inflectional morphology. In this debate, the term analogy has come to be used in different senses, which are often heavily dependent on the underlying theory. In particular,

senses of analogy are often tied to particular assumptions about central theoretical notions such as regularity, productivity, variability, and the nature of lexical representations.

Thus, we find approaches that claim that analogy is the basis of any rule-based, productive behaviour in morphology (cf. e.g. Blevins and Blevins 2009a for an overview), whereas at the same time we find, especially generative, approaches that appeal to analogy exactly in those cases in which linguistic behaviour is not rule-governed, but exceptional, unproductive, unpredictable, or irregular (cf. e.g. Prasada and Pinker 1993, Pinker and Prince 1994). Also within word-formation theory, analogy has come to be used as a term opposite to the concept of the linguistic rule (cf. Bauer 1983, Bauer 2001). In usage-based and constructionist approaches, by contrast, it is argued that analogy forms the underlying principle of exemplar-based reasoning or the beginnings of low-level schematisation (Booij 2010: 88-93). Crucially, the implication in this latter group of approaches is that analogies in word-formation are regularly based on subsymbolic aspects of lexical representations. However, this assumption is not inherent in the definition of analogy per se, as we will see.

This article is concerned with concepts of analogy in synchronic word-formation. The focus is on providing an overview of the different notions of analogy as they are used in different theories of word-formation, and of the different phenomena that have featured prominently in references to analogy in the word-formation literature. Reference to parallel developments in theories of inflection will be made be made occasionally, where necessary.

The structure of the article is as follows: We will begin with a general definition of proportional analogy as a heuristic device (section 2.). Section 3. will then provide an overview of the status of analogy in different theories of word-formation. We will see how different theories operationalise different aspects of the structure of proportional analogy in different ways, resulting in radically different views about the regularity, productivity, and predictability of analogical formations. Based on this overview, theories will be grouped broadly into (a) theories that consider analogy to be an irregular or exceptional process and (b) those that consider analogy to be the basic process underlying word-formation. Sections 4. and 5. will then be devoted to these two classes of theories, respectively, discussing pertinent word-formation phenomena for which analogy has been invoked. The discussion in section 5. will specifically focus on computational analogical models, which will be shown to provide interesting solutions to some of the criticism that has traditionally been mounted against analogy-based models, but will also be shown to be limited in terms of the range of processes covered in such approaches to date. The article ends with a conclusion (section 6.).

2. Definition and terminology

Analogy as used in the word-formation literature is usually described in terms of a proportion (proportional analogy), as in (1).

(1) a:b = c:x

In this equation, 'x' is the new form, i.e. a morphologically complex word that is about to be coined. 'a', 'b', and 'c' are forms that already exist in the lexicon. 'a' and 'c' are (potential) base forms, whereas 'b' is an existing complex form. What happens in an analogical formation, then, is that the relationship between 'a' and 'b' is used as a model example for the formation of 'x'.

As an example, consider the English compound *chairperson*, which we may plausibly assume to have been formed on the basis of analogy with the existing compound *chairman*. In this case, we can fill the variables in the equation above as in (2).

(2) *chair* : *chairman* = *chair* : *chairperson*

Interestingly, there is no established terminology in the morphological literature to refer to most parts of the analogical equation. The only established term seems to be 'analogue', which is usually used to refer to the complex form on which a newly coined word is modelled ('b' in (1), *chairman* in (2)). In order to facilitate further discussion, however, it is useful to have labels to refer to the other parts of the equation as well. These labels are presented in Figure 1, again using *chairperson* as an example.

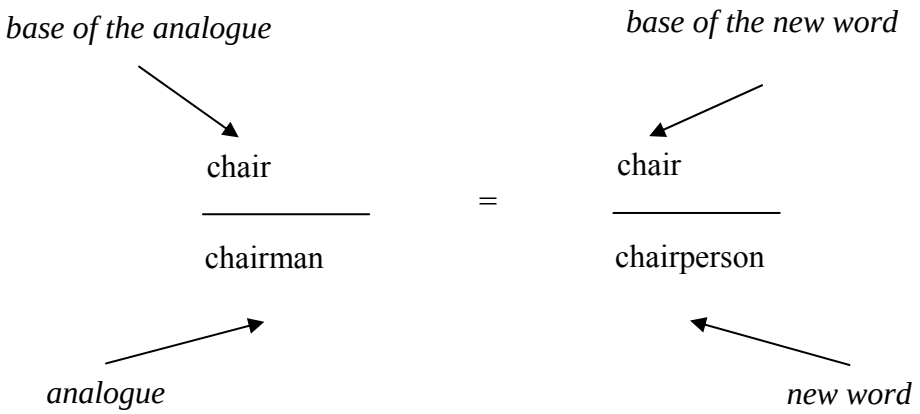

Fig. 47.1: elements of the analogical equation

In accordance with most of the literature, we will use the term 'analogue' to refer to the complex word that serves as a model for the coining of a new complex word. The new word that is about to be coined will be referred to as the 'new word'. Finally, there are two bases involved in an analogy, for which we will use the terms 'base of the analogue' and 'base of the new word', respectively.

As is clear from the example, a key role in the process of analogy is played by the (perceived) similarity between the elements of the equation. Figure 2 provides an overview of (and labels for) the similarity relations we find within the analogical equation.

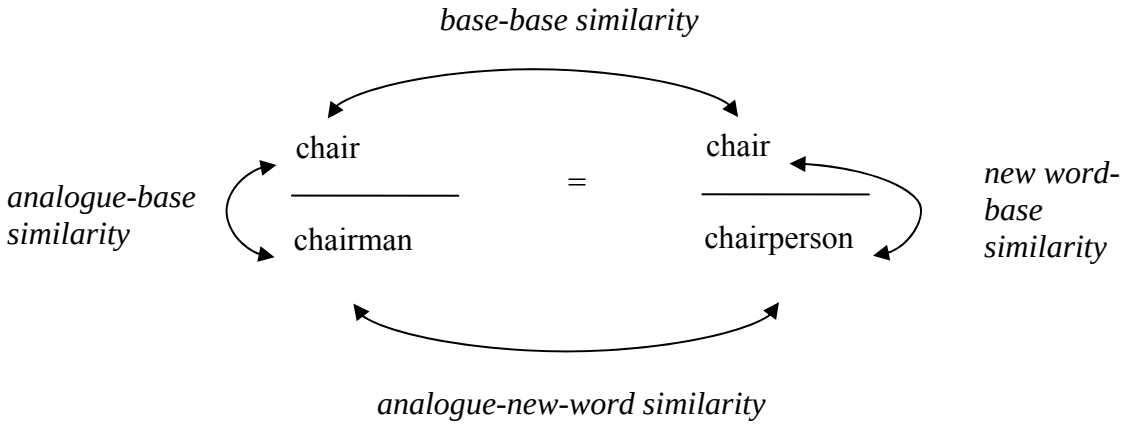

Fig. 47.2: similarity relations in the analogical equation

In the example *chairperson* the base of the analogue and the base of the new word are identical (both *chair*). Hence, it is easy to see that they may be perceived as being similar to

each other by the hypothetical speaker(s) who coined *chairperson*. The second aspect where similarity plays a role is the relation between the base of the analogue (*chair*) and the analogue (*chairman*), which must be perceived as being similar to the relation between the base of the new word (*chair*) and the new word (*chairperson*). In *chairperson*, the relation between the analogue and its base is a morphological relation, pertaining to both the form and the meaning of the two lexemes. The form of the base, *chair*, appears as the first constituent of the compound *chairman*. Semantically, we could broadly say that the relation between the analogue and its base is that between a role (*chair*) and the occupant of that role, who is human and male (*chairman*). Like the relation between *chair* and *chairman*, the relation between *chair* and *chairperson* (i.e. that between the new form and its base) also pertains to both the form and the meaning of the elements involved. Again, *chair* appears as the first constituent of a compound. Semantically, the relation between the new word and its base is that between the role (*chair*) and the occupant of that role, who is human but, crucially, not necessarily male.

Our example already indicates that similarity is not only a key determinant of a morphological analogy, but that it is also one of the key problems in defining and explaining analogies, and, on a theoretical level, a key challenge for any morphological theory that is based on analogies. The reason is that the basis for the computation of similarity is not part of the equation.

Whereas the word-formation literature generally agrees that in cases of analogy similarity must be given both in terms of form and in terms of meaning, there is almost no restriction on precisely which formal and semantic properties can make an analogue and a potentially corresponding new word similar. In addition, there is no agreement about how analogical similarity relations (those schematised in Fig. 2) map onto morphological complexity relations. For example, when discussing *chairperson* as a product of analogical reasoning on the basis of *chairman*, we assumed that the relation between *chair* and *chairman* was the base for the analogy. At the same time, however, it is also clear that *chairman* is related to *man* as much as it is related to *chair*, and that the relation between *man* and *chairman* is similar to the relation between *person* and *chairperson*. Thus, Fig. 3 below is another plausible representation of an analogical relation between *chairman* and *chairperson*.

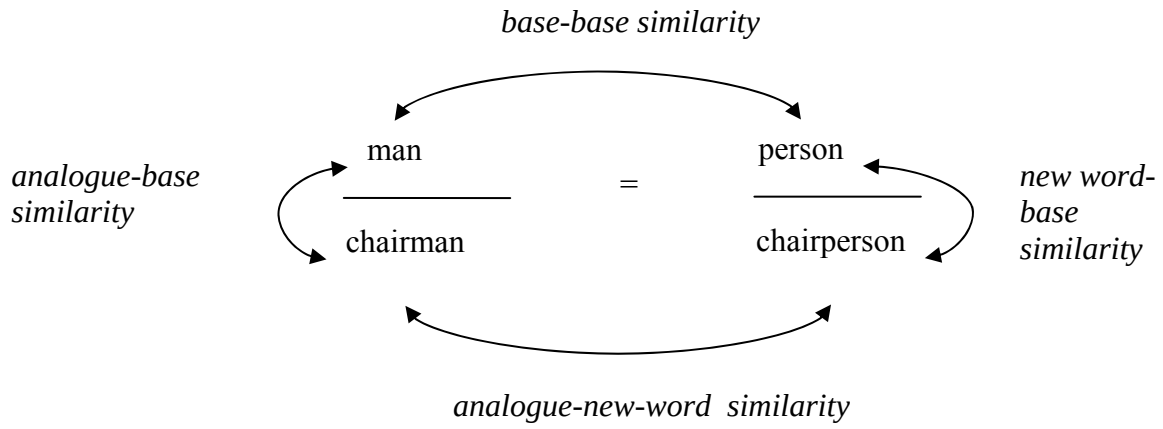

Fig. 47.3: *chairman – chairperson* – an alternative

What this tells us is that it is not always clear what exactly the base of an analogy is. There is, from a theoretical point of view, no restriction on which of the multiple similarity relations that exist between words in the mental lexicon may form the basis of an analogical formation.

In sum, we see that *analogy* as a heuristic formalism does not say much about many of the issues that morphological theory needs to be explicit about. Specifically, it does not say anything about (a) which features (formal, semantic, syntactic, etc.) establish similarity relations on which analogies may be based, and (b) which of the existing similarity relations may or will form the basis of a new analogical formation. Furthermore, as we will see later on in this article, it does not say anything about (c) how many lexemes are involved in an analogy. This explains, in part, the great diversity of usages of the term analogy in the literature, to which we will now turn.

3. Senses of *analogy* in word-formation theory

The focus in this section will be on showing the scope of senses in which analogy has been used in the literature on synchronic word-formation. Rather than attempting to be exhaustive, the discussion will be restricted to a sample of representative theories.

Analogy is often discussed in the context of the theoretical divide between word-based, paradigmatic, and syntagmatic approaches to morphology (cf., e.g., Becker 1990, Becker 1993a and references therein for discussion). It is clear that approaches to word-formation which attribute a systematic role to analogy are all word-based approaches. However, not every word-based approach assumes that analogy is an active mechanism in

productive word-formation. This is true in spite of claims often made in the literature that the mechanisms underlying word-based formalisms can be described in terms of a proportional analogy. Word-based approaches are divided in terms of whether they consider rules or analogy to be the central mechanism in productive word-formation. In rule-based approaches of this type, analogy is often invoked to explain irregular, or unproductive behaviour. In analogy-based approaches, analogy is invoked to explain regular, productive behaviour. The term 'paradigmatic' approaches to morphology is difficult to apply here, because it is used in different senses in the literature. Whereas it is assumed to be synonymous with 'word-based' approaches by some authors, others use it rather in the sense that has been labelled 'analogy-based' above. In what follows we will discuss senses of *analogy* in word-based theories.

One type of approach that takes an extreme position with respect to the rule-analogy divide is comprised of, mostly generative, paradigmatic frameworks which make a radical distinction between analogical formations and regular processes of word-formation. Regular processes are the product of an abstract formalism that operates independently from individual lexemes, on symbolic features that are shared by pertinent lexemes (cf. e.g. Aronoff 1976). Analogy, in this view, is always local in the sense that it affects only few and very specific lexical items. Productivity is rule-application, with the consequence that a low degree of variation is predicted for morphological rules. Unlike regular and productive word-formation, then, processes of analogy are unpredictable and unproductive. A clear expression of this view, which is found frequently in the generative literature, is found, for example, in Bauer's (1983) textbook on English word-formation (but cf. also Bauer (2001: 75-97), where this view has been relativised considerably):

> If instances of word-formation arise by analogy then there is in principle no regularity involved, and each new word is produced without reference to generalizations provided by sets of other words with similar bases or the same affixes: a single existing word can provide a pattern, but there is no generalization. [...] If it is true that there are in principle no generalizations, then a generative account of word-formation is at best a convenient fiction and at worst an irrelevancy.
>
> (Bauer 1983: 294)

At the other end of the divide we find approaches that assume that analogy with existing lexemes is used regularly and productively in the formation of new words. In what follows I

will refer to such approaches as 'analogical' approaches. Conceptually, they are rooted mainly in two traditions: One is the Neogrammarian and the American descriptivist tradition (cf. esp. Becker 1990 for details and discussion), the other comprises functionalist, exemplar-based, and usage-based theories of grammar, i.e. theories that are grounded in the assumption that analogy constitutes a central cognitive mechanism that is active in human cognition in general, and in language in particular (cf. esp. Bybee 2001, Bybee 2010, Gahl and Yu 2006, Blevins and Blevins 2009b). Work in both types of theories has traditionally focussed on inflection and on diachronic language development. Recent times, however, have seen a growing number of publications devoted to word-formation phenomena.

For example, Becker (1990, 1993a, 1993b) proposes for both inflection and derivation that all morphological operations are analogical, in the sense that they describe relations between existing words, on the basis of which speakers productively coin new words by means of proportional analogy (e.g. Becker 1990: 187). His proposal builds on and extends those of especially van Marle (1985, 1990), who also claims a synchronic relevance for analogical formation in word-formation, but distinguishes between (rule-based) productivity and (analogy-based) creativity, the former producing regularity, and the latter allowing for some degree of unpredictability.

One obvious characteristic of approaches that do not embrace the distinction between analogical and rule-based word-formation is that they consider variability and gradience to be a key property of morphological operations. This variability has often been identified with unpredictability, and the failure of many analogical approaches to be predictive constitutes one of the key points of criticism against these approaches (cf. above, cf. Bauer 2001: 75-97 for a summary of pertinent arguments – and counterarguments). However, there is also a growing body of literature springing from mainly quantitative work in morphology that challenges the view that analogical models must necessarily be nonpredictive (cf. e.g. Baayen 2003, Hay and Baayen 2005). The main argument is that variability in general and different degrees of variability in particular can be predicted in a probabilistic approach to linguistic categorisation. One major class of such probabilistic approaches explicitly draws on analogy as the fundamental underlying principle of morphology and has produced a growing number of studies that model word-formation phenomena with the help of computational implementations of analogical models of grammar (specifically: AM(L), Skousen 1989, 1992, Skousen, Lonsdale and Parkinson 2002; TiMBL, Daelemans, Zavrel, van der Sloot and van den Bosch 1999 et seq.). What these models have in common is that they apply similarity-based, analogical reasoning, creating new forms on the basis of the similarity of the base of

the new form with existing forms in the lexicon. Analogies are therefore very rarely local; the idea advanced in many rule-based approaches that analogies must be based on a single lexical item is conceived to be only one of many possibilities. Much more frequently analogues in such approaches are sets of words in the lexicon. Behaviour that is described as rule-based in other approaches emerges exactly in situations in which analogues for a given new form comprise a large set. Details of such models will be discussed in section 5.2.

Whereas the said analogical models fundamentally differ from rule-based models in the way in which they view variability, however, not all of them differ from rule-based models in terms of the way in which they conceptualise lexical representations. Thus, in some analogical approaches it is assumed that similarity relations between words are established on the basis of symbolic features (e.g. Becker 1990: 63-71). In some exemplar-based models, by contrast, it is assumed that the exemplars that serve as potential analogues also comprise information that is more detailed and specific than the abstract features traditionally associated with rule-based grammatical models (cf. e.g. work on compounding, esp. Krott, Baayen and Schreuder 2001, 2002, Plag, Kunter and Lappe 2007, Arndt-Lappe 2011).

An intermediate position between views that consider analogy to be exceptional and those that consider it to be the basis of regular word-formation is found in constructionist theories. The interesting question here is how analogy is related to schemas or schematisation, which are considered to be the central mechanism in word-formation. A clear view on this is found in Booij's recent proposal (Booij 2010). Here it is claimed that schemas and subschemas may operate on symbolic features, and that the crucial difference between analogical formations and schema-based formations lies in their making reference to different degrees of abstraction. Analogy in this model is defined as strictly local analogy, which is complementary to schemas and may constitute an initial stage of the development of a schema (cf. esp. Booij 2010: 88-93, cf. section 4.1. for examples). Thus, Booij's constructionist approach is different from both analogical and non-analogical approaches outlined above. There are differences between what happens in productive word-formation and what happens in an analogical formation in his sense, but this difference is a gradual difference, and not, as in the generative tradition, a difference that concerns the fundamental nature of the system.

In the remaining two sections of this article I will provide an overview of the type of word-formation phenomena for which analogy has been invoked in the literature. Section 4. will deal with analogical formation in approaches which attribute a complementary role to analogy, i.e. rule- or constraint- or schema-based approaches in the sense outlined above.

Section 5. will be devoted to phenomena discussed in analogical approaches, both non-computational and computational.

4. Complementary analogy

We find pertinent appeals to analogy mainly in three domains. The first comprises relatively local analogies, explaining the emergence of small-scale patterns (section 4.1.). Such small-scale patterns are often seen as the precursors to morphological processes proper. The second domain comprises apparent cases of reanalysis (section 4.2.). The third domain is allomorph selection (section 4.3.).

4.1. Local analogies

An analogy is usually considered to be local if (a) the analogue is restricted to one particular lexeme, (b) a very high degree of similarity is involved, and (c) the productivity of the process is very limited, in the extreme case producing only one new word. Needless to say, 'locality' of an analogy is a gradual notion.

Traditionally, many such cases are of the type *chairman* :: *chairperson* discussed in section 2. above, where both formally and semantically there is complete identity of one of the bases involved. A representative recent analysis is Booij's (2010) constructionist approach, where a distinction is made between 'analogical word-formation in the strict sense' (p. 89) on the one hand and constructional schemas on the other hand. Booij's examples of analogical word-formation are given in (3). They are all from Dutch.

(3) Examples of analogical word-formation (from Booij 2010: 89, glosses are taken from the original)

new word	*analogue*
paniek-haas, 'lit. panic-hare, panicky person'	angst-haas, 'lit. fear-hare, terrified person'
vader-taal, 'lit. father-language, native language of father'	moeder-taal, 'lit. mother language, native language'
muis-vaardig, 'lit. mouse-able, with mouse-handling skills	hand-vaardig, 'lit. hand-able, with manual skills'
oud-komer, 'lit. old-comer, immigrant who arrived a long time ago'	nieuw-komer, 'lit. new-comer, recent immigrant'

For the distinction between analogical word-formation and constructional schemas, Booij considers it to be crucial that

> [f]or these words [i.e. those in (3), S. A.-L.] we can indeed point to one particular existing compound as the model for the formation of the new compound, and the meaning of this new compound is not retrievable without knowing the (idiomatic) meaning of the model compound.
>
> (Booij 2010: 90)

Even in approaches using the term analogy in such a restrictive sense, it is often noted that analogies of the type exemplified in (3) can give rise to new word-formation patterns (cf. Booij (2010: 90-91), Szymanek 2005 for examples and pertinent references). The claim in many theories is, however, that once a new pattern has been created, its description in terms of analogy is no longer appropriate.

There are two conceptual problems involved for an analogical description of the new pattern. One is that, once a pattern has arisen, it is impossible to trace the analogy back to one single analogue lexeme. Another problem is that many of the examples that are usually quoted in the literature involve reanalysis of either the analogue or the base of the analogue as morphologically complex. The problem of reanalysis will be discussed in section 4.2. below. For the problem of analogue selection, reconsider our example *chairperson* from section 1., where we have assumed that its analogue is *chairman*. However, there also exists *chairwoman*, which, like *chairman*, predates *chairperson* (the *Oxford English Dictionary* records 1699 as the date of the earliest attestation of *chair-woman* and 1971 as the corresponding date for *chairperson*). It is, therefore, unclear whether *chairperson* was modelled on *chairman* or on *chair-woman*. Indeed, the definition of the OED as 'a chairman or chairwoman' (OED, s.v.) suggests that it may have been modelled on both.

On a theoretical level, the example shows that the assumption that the distinction between 'analogy in the strict sense' and a pattern is not without problems. Thus, if we take Booij's approach, we would assume the existence of a schema to account for the triplet *chairman – chairwoman – chairperson*. Apart from the apparent stipulation that bases of analogy must be single lexemes, it is, however, unclear why *chairperson* should be attributed a different status in the word-formation system than cases like *paniek-haas* (cf. (3) above). An alternative approach would assume that analogues may be sets of lexemes, which is, for

example, inherent in many traditional, analogy-based accounts of cases of reanalysis, to which we now turn.

4.2. Reanalysis based on analogy

The reason why analogy is often invoked in reanalysis cases lies in the fact that reanalysis obviously happens on the basis of similar lexemes that are stored in the lexicon. There are two pertinent classes of processes: certain cases of backformation and affix secretion (Marchand 1969: 210-214). Note that for both classes also rule-based accounts have been proposed. Examples of backformation are given in (4).

(4) Cases of backformation

new word	*base of new word*	*analogue/base*	*reference*
burgle	burglar	write/writer, sing/singer, etc.	(Marchand 1969: 391)
televise	television	act/action, revise/revision, etc.	(Marchand 1969: 395)
self-destruct	self-destruction	cases of noun-verb pairs where *–ion* is added to the verbal base	(Bauer 2001: 83)

Cases like *to burgle* (derived from: *burglar*) are commonly analysed as involving reanalysis of the base word as morphologically complex, on the basis of analogy with existing pairs of lexemes. In this case, the bases of the analogue and the new word share a form, *-er*, but this form does not have the same meaning in the two bases. Note, however, that still there is a semantic similarity between the two bases: In the case of *burglar* and its analogues, for example, they all denote agents.

One major theoretical issue in approaches to backformation cases is whether back-formation is a diachronic process (cf. Becker 1993a for discussion and a review). Examples like those in (4) have been used convincingly to demonstrate the synchronic relevance of the process. The basis of this argument is semantic complexity. Whereas, for example, *to burgle* means 'to act the burglar', the noun *burglar* does not mean 'one who burgles' (Marchand 1963, as discussed in Becker (1993a: 4-8)).

Another case of apparent backformation which has been analysed by appealing to analogy is the case of bracketing paradoxes. A well-known representative is Spencer's analysis of certain English person-denoting adjective-noun phrases (Spencer 1988). Pertinent data are given in (5).

(5) Bracketing paradoxes

new phrase	*analogue*
transformational grammarian	transformational grammar
atomic scientist	atomic science
moral philosopher	moral philosophy

The phrases in (5) are part of a group of phrases that, for some morphological theories, form bracketing paradoxes because it is unclear to which base the person-noun forming suffix (in the examples in (5): *-ian*, *-ist*, *-er*) is attached. Thus, the morphological base for *–ian* suffixation is the noun *grammar*, while the semantic base is the phrase *transformational grammar*.

According to Spencer, formations such as those in (5), which are clearly productive in English, pose a challenge to rule-based morphological theories because the derived person noun cannot be convincingly related to their bases via a syntagmatic morphological rule (which would, for example, involve suffixation in *grammarian* and suffix substitution in *philosopher*). He therefore argues that the relation between bases and derivatives is an analogical relation, pertaining between lexicalised phrases in the mental lexicon. Unlike in the 'traditional' backformation cases discussed further above, then, Spencer's claim is that analogical processes may be productive.

Another group of reanalysis cases where analogy is often assumed to play a role involves cases where new morphological patterns emerge (affix secretion, in Marchand's terminology, cf. Szymanek 2005: 431, 435-436 for English). Similarly to the backformation cases, in these cases the semantics of the base for the analogical formation arises through reanalysis of the analogue-base relation as morphologically complex. Unlike the cases in (4) and (5), however, analogue forms are not morphologically complex, at least not before the advent of secretion. Some of the pertinent cases have also been described as blends. Examples are given in (6).

(6) cases of affix secretion

new word	*analogue*	*meaning*	*reference*
candyteria	cafeteria	'shop, store, or establishment selling food x'	(Marchand 1969: 211)
Monicagate	Watergate	'political scandal involving x'	(Szymanek 2005: 436)

There is a formal overlap between the new word and the analogue (*-teria*, *-gate*). This form, however, is not a unit of meaning in the analogue base, but becomes a unit of meaning in the new word, i.e. the moment it is extended to other words. This is precisely the situation for which Booij (2010: 88-93) argues that schematisation takes place, which in his view replaces analogy as the underlying mechanism.

4.3. Allomorph selection on the basis of analogy

A third group of processes where analogy has been invoked in the literature is allomorph selection. Pertinent examples are given in (7).

(7) Irregular allomorph selection

new word	*analogue*	*reference*
orienteer	volunteer	(Bauer 1983: 290)
womanity	humanity	(Baeskow 2011: 4-5)

For examples like those in (7), analogy is often invoked to explain affix selection. The underlying assumption is that this selection is irregular or unproductive. Thus, in both cases in (7), the affix selected to derive an agent noun (*-eer*) and an abstract nominalisation (*-ity*) is allegedly unproductive for the bases *orient* and *woman*, respectively. Bauer (1983: 285-291) assumes that *–eer* is generally unproductive in Modern English, and appeals to analogy to provide an explanation for the form *orienteer*, which is an apparent counterexample. The basis for the analogy here is phonological similarity between *volunt-* and *orient*, i.e. the base of the analogue and the base of the new word.

The form *womanity* (7b) is an apparent counterexample to the generalisation that English *–ity* attaches to Latinate bases (cf., e.g., Baeskow (2011: 4-5) for discussion and further counterexamples). Like in the case of (7a), phonological and perhaps also semantic similarity between the bases of analogue and new word, *woman* and *human*, plays a large part in motivating the analogy. Unlike in (7a), the two bases are also semantically similar.

Apart from allomorph selection, also other types of selection between grammatical alternatives have traditionally been explained with the help of analogy. A case in point is stress in English nominal compounds, where two types of stress are available: left stress and right stress. Whereas left stress has traditionally been assumed to be the default pattern, cases of right stress have often been explained to be the product of analogy. Oft-cited textbook examples are compounds denoting street names. The examples in (8) are taken from Plag (2003). Stress is marked by an acute accent on the pertinent vowel.

(8) Stress in English noun-noun compounds (Plag 2003: 139)

Óxford Street	Madison Ávenue
Fóurth Street	Fifth Ávenue

The analogical effects exemplified in (8) differ in two important ways from the allomorph selection cases in (7). Thus, in (8) analogue bases are sets of words, not isolated words (i.e. all compounds whose second constituent is *street* or *avenue*, respectively). Also, within their domain, analogies in (8) are productive.

In sum, we have seen in this section that appeals to analogy in rule-, constraint-, or schema-based theories do by no means form a homogeneous group. What they have in common, is that analogy is conceived to be relatively local, usually affecting lexical items which are highly similar both phonologically and semantically. However, approaches differ in terms of whether they define elements involved in the analogy as single lexemes or sets, and in terms of which similarity relation exactly they view as being crucial to trigger an analogical formation. Also, they vastly differ in terms of whether they view analogy as a (potentially) productive process.

In the next section we turn to a group of theoretical approaches which not only assumes that analogy may be productive in morphological grammar, but that it in fact forms its underlying principle.

5. Analogy in analogy-based approaches to word-formation

Two groups of phenomena feature prominently in analogical approaches to word-formation: productive replacive formations, and cases of variability that affect the formal properties of outputs of word-formation. The former has been the object of discussion in much of the general theoretical literature, whereas the latter has been in the focus of the literature working with computational analogical models.

5.1. Productive replacive formations

The term 'replacive formation' refers to a pattern where new words are coined from existing complex words via affix replacement (Becker 1993a: 9-12). The phenomena discussed in the literature are in part the same as the phenomena discussed in sections 4.1. and 4.2. The issue under debate between analogical and non-analogical approaches here is, however, the productivity of these phenomena. Whereas in much of the generative literature it is claimed

that analogical patterns of the type discussed in section 4. are not productive, it is claimed in the analogical literature that they are productive. In addition, it is claimed that the distinction between analogy and rule-based behaviour cannot be upheld on formal grounds.

An explicit discussion of the theoretical implications that the existence of productive replacive formation has in terms of an analogical approach to word-formation is found in Thomas Becker's work (Becker 1990, 1993a, 1993b). The key argument has two parts: (a) There are productive replacive word-formation patterns (Becker 1993a, 1993b), (b) They are not different from patterns which have traditionally been described as rule-governed.

An example of a productive replacive formation pattern is the pattern producing *in-* and *ex-* prefixed words in German (discussed in Becker 1993b: 194). Examples are given in (9).

(9) German pairs of *in-* and *ex-*prefixed words (Becker 1993b: 194)

Immatrikulation	'immatriculation'	Exmatrikulation	'exmatriculation'
Inkardination	'incardination'	Exkardination	'excardination'
Internat	'boarding school'	Externat	'a school that accepts day students'

In- and *ex-*prefixation is replacive in the sense that one prefixed word is derived from the other prefixed word. In all examples in (9), the *ex-* derivative has been coined on the basis of the *in-* derivative, but there are also examples in which the reverse is the case. The pattern constitutes evidence in favour of a paradigmatic approach to word-formation because, as Becker convincingly shows, many pertinent cases cannot be described in terms of a concatenation of *in-* or *ex-* and a base. For example, *ternat* is attested only in the pair *Internat* and *Externat* in German, which makes it difficult to analyse it as a base for prefixation.

The second part of Becker's argument, i.e. that replacive formations are not different from other, allegedly rule-governed patterns, is more difficult. In Becker's analogical approach, both types of pattern are described in terms of a rule format that Becker shows to be formally equivalent to a proportional analogy, especially since the format crucially employs traditional symbolic representations of lexemes. For Becker, the only difference between word-formation patterns that have been described as rule-governed and those that have been described as analogical in the generative literature lies in their different degrees of productivity. However, like many strands of traditional generative theory, which claim

exactly the opposite, also Becker's theory is lacking a testable means to predict which patterns will be more productive and which ones will be less productive.

5.2. Cases of formal variability

Testability of degrees of variability is one of the key issues that has been addressed in simulation studies employing computational analogical models. The discussion here will focus on work based on the two analogical algorithms which are most widely used to model word-formation phenomena: the Tilburg Memory Based Learner (TiMBL, Daelemans, Zavrel, van der Sloot and van den Bosch 1999 et seq. and Skousen's Analogical Model of Language (Skousen 1989, Skousen 1992, Skousen, Lonsdale and Parkinson 2002). Another algorithm that has been used in much work on inflection is the Generalized Context Model (Nosofsky 1986, 1990). Furthermore, there is also work investigating the role of analogical factors using statistical modelling, without the implementation of a formal analogical model (cf., e.g., Plag 2006, 2010 on English compound stress).

An obvious question is what the exact nature of the analogical theory is that is implemented by algorithms like TiMBL and AM. We will address this issue after we have discussed relevant studies.

Like many analogical models, the key focus in the initial stages of pertinent morphological research was on problems of inflection (cf. esp. Daelemans and van Bosch 2005, Skousen, Lonsdale and Parkinson 2002 for a summary of central issues). Issues of word-formation addressed in the literature always concern cases in which outputs of word-formation exhibit some sort of semi-regular variability. They are 'semi-regular' in the sense that deterministic rule-based models fail to predict the attested variability. In contrast to work on inflection where there is a wealth of literature exploring the predictive power of computational analogical models (cf. relevant references in Skousen, Lonsdale and Parkinson 2002, the AM bibliography at http://humanities.byu.edu/am/am-biblio.html, Daelemans and van Bosch 2005, and, in particular, work on the past tense in various languages in Skousen 1989, Eddington 2000, Keuleers 2008), pertinent research on word-formation phenomena is still in its infancy. Existing research has focussed mainly on two word-formation phenomena: compounding and allomorphy in derivation. Pertinent studies are most often based on corpus data and, in some cases, on data in which novel complex words have been generated by experimental subjects.

Simulation studies devoted to variability in compounding have investigated linking morphemes in Dutch and German (esp. Krott, Baayen and Schreuder 2001, 2002, 2007, cf.

Krott 2009 for a summary) and stress assignment in English noun-noun compounds (esp. Plag, Kunter and Lappe 2007, Arndt-Lappe 2011, cf. also Plag 2006, 2010). Both TiMBL and AM were used as algorithms.

The three options that are available as linking morphemes in Dutch are: *-s-*, *-en-*, and *-Ø-* ('zero, no linking morpheme'). Examples are given in (10). For easier reading but contrary to orthographic conventions, the relevant morphological components in the Dutch words are separated by spaces.

(10) Variability in Dutch linking morphemes (from Krott, Schreuder and Baayen 2002: 55f.)

thee – bus	'teabox'
papier – handel	'paper trade'
plaatje s boek	'picture book'
tabak s rook	'tobacco smoke'
krent en brood	'currant bread'
boek en kast	'book case'

In English noun-noun compounds, two options are available: stress on the first constituent ('left stress') or stress on the second constituent ('right stress', sometimes also referred to as 'level stress'). Examples are given in (11). Stress is marked by an acute accent.

(11) Variability in stress assignment in English noun-noun compounds

ópera glasses	steel brídge
wátch-maker	morning páper
clássroom	silk tíe
Óxford Street	Madison Ávenue

In spite of the fact that they are concerned with different phenomena, the two groups of simulation studies show surprising agreement in terms of their findings. Thus, predictive power of the computational analogical models employed was greater than that of traditional rule-based models. Furthermore, the most important determinant of the variation was the constituent family. This means that linking morphemes or stress assignment of a given novel compound can be predicted best on the basis of the pertinent behaviour of existing compounds that share either the first or the second constituent with the novel compound. In addition, it was found in most studies that, apart from constituent family, also semantic factors, if included in the simulation, served to enhance predictive power.

Another group of phenomena that has been studied is allomorphy in derivation. Pertinent studies include, for example, work on diminutives (Daelemans, Berck and Gillis 1997 on Dutch, Eddington 2002, 2004 on Spanish), negative prefixation (Chapman and Skousen 2005, with a diachronic perspective) and comparative formation (Elzinga 2006) in English. A representative study is Eddington's study of diminutives (Eddington 2002, Eddington 2004). Examples are given in (12).

(12) Variable allomorph selection in Spanish diminutives (Eddington 2002: 402)

minut-ito	←	minuto	'minute'
gallet-ita	←	galleta	'cookie'
vidri-ecito	←	vidrio	'glass'
yerb-ecita	←	yerba	'grass'
pastor-cito	←	pastor	'shepherd'
joven-cita	←	joven	'young girl'
normal-ito	←	normal	'normal'
naric-ita	←	nariz	'nose'
pec-ecito	←	pez	'fish'
flor-ecita	←	flor	'flower'
lej-itos	←	lejos	'far away'
Luqu-itas	←	Lucas	'Luke'
patron-cita	←	patrona	'patron saint'

Variability affects at least two dimensions: the form of the diminutive suffix (mainly *–ito*/*a*, *-cito*/*a*, *-ecito*/*a*), and the form of the stem allomorph (truncated, not truncated).

In Eddington's simulation experiment AM is able to predict correctly some 92% of the data. In addition, it is shown that the variability predicted by AM is plausibly similar to the variability that exists in real life. Thus, uncertainty in the model's predictions occurs exactly where uncertainty in allomorph selection in real life manifests itself, for example, by the existence of doublets.

Eddington's study of Spanish diminutive allomorphy is also representative of this type of analogical approach in terms of the features that were given to the algorithm as its information source. These typically involve mainly the phonological (i.e. segmental and prosodic) shape, but also the relevant grammatical categories of the base words in the dataset. In Eddington's study, the latter comprised gender information.

In general, the studies that have been introduced here are representative of work employing computational analogical algorithms to model variability in word-formation. An obvious difference to much previous work in analogy is that analogical algorithms are predictive mechanisms. In what follows we will briefly address the question how this is achieved. Major differences between AM and TiMBL will be mentioned, but will not be in the focus of the discussion.

Like in all approaches discussed in this paper, analogical word-formation is assumed to be the product of a perceived formal and semantic similarity between a form that is about to be coined and its analogue (cf. section 2. above). Unlike in other approaches, however, the scenario that the analogue is only a single form or a small set of forms that is maximally similar to the new form is only one of several potential scenarios. Instead, analogies are based on those exemplars in the lexicon that are informative with respect to the given task. This group of exemplars is often called the 'analogical set' of a new form (esp. in the AM-based literature) or the 'nearest neighbour set' (in the TiMBL-based literature). Often, exemplars in the analogical set will differ in terms of their similarity with the given item. Classification of a new form will therefore always incorporate an effect of (type-) frequency because all members of the analogical set will influence classification. Thus, one element that makes algorithms like AM and TiMBL predictive is the fact that, unlike other analogical approaches, they have a principled way of determining which exemplars in the lexicon will serve as analogues, i.e. will be part of the analogical set.

The second element that makes the models predictive is that they have a principled method at their disposal to determine which types of similarity are relevant for a given classification. The basis of all computation of similarity is formed by those elements of lexical representations which are provided by the researcher as a set of coded features for each exemplar in the database. The nature of these features is, in principle, a matter of choice, and it is still an unresolved question, which types of features lead to the best predictive power of a model. Existing studies of compounding have successfully used features encoding aspects of the compositional semantics of the compound, as well as features encoding the particular identity of the compound constituents. Existing studies of derivational allomorphy have typically used syllable-based phonological representations of base words (e.g. phonemes of onset, nucleus, coda of the ultima, penult, etc.) as well as, in some cases, grammatical information such as gender, word class, etc. Crucially, the question of how abstract or symbolic representational features are is still a matter of debate.

The problem of determining which exemplars end up in the analogical set for a given new form is resolved in different ways by TiMBL and AM. What they have in common, though, is that the analogical set / nearest-neighbour set comprises those items which are similar to the new form in terms of exactly those features that are most useful for the given task. To do this, most varieties of TiMBL weigh the coded features of all exemplars in their lexicon in terms of how informative they are with respect to the given task, and treat items that share more informative features as more similar to a given new form than items that share less informative features. This means, then, that the importance of a given feature for the computation of similarity is the same for the whole lexicon. This is different in AM, where the decision of which coded features are relevant for a given exemplar to end up in the analogical set is made for each new form on an individual basis. For each new form the algorithm determines which combinations of features shared or not shared with that new form behave in a homogeneous way with respect to the given task (cf., e.g., Skousen 2002a, 2002b for a discussion of 'homogeneity').

A crucial property of computational analogical models that is particularly relevant for grammatical theory in general and morphological theory in particular is that they have been claimed to be able to account for both types of effect: one that has traditionally been described as local analogy and one that has traditionally been described as rule-governed behaviour. This point has explicitly been made mainly for inflection (cf. esp. Derwing and Skousen 1994, Eddington 2000, Keuleers 2008), but is also often alluded to in work on derivation and compounding. Recall from section 4. above that in non-analogical, rule-based approaches it has often been claimed that the distinguishing feature between regular and analogical processes is productivity. In an analogical model, however, there exists only a gradual distinction between local analogies and less local analogies. A local analogy arises if only exemplars which share many features with the new form are incorporated in the analogical set. Typically, then, analogical sets will be very small, and members of these sets will be highly similar in terms of both their phonological structure and their meaning. By contrast, behaviour that looks like rule-governed behaviour in the traditional sense will arise if the analogical set is large, with exemplars in that set sharing fewer features. Thus, members of the analogical set will be less similar to each other both phonologically and semantically.

6. Summary and conclusion

This article has presented an overview of approaches to analogy in word-formation theory. It has become clear that we have to distinguish *analogy* as a heuristic device from *analogy* as a

construct in word-formation theory. The former is a mechanism that is very open and, in principle, underspecified in terms of many issues that need to be addressed in morphological theory. This openness of analogy is reduced in the specific usages of analogy as a construct in word-formation theory. These usages are closely tied to the theories' basic assumptions about productivity, regularity, and variability. Strikingly, analogy has often been used in a narrow sense to denote local analogies, where both formally and semantically a very high degree of similarity is involved. The implication of this usage is that it is assumed that less local generalisations are non-analogical, because they involve higher degrees of abstraction (in terms of representations, rules, constraints, or schemas).

This view is not shared by analogical approaches, where local analogies are considered to be only a special case of analogy. Here emphasis is put on the generality of analogy as a mechanism, with the implication that predictability and regularity of morphological operations is gradient. Constraints on analogy are often seen as a consequence of the nature of lexical representations and usage-based factors. In computational analogical theories, we furthermore observe that analogy is conceptualised as a predictive mechanism, where predictability emerges from the fact that analogues are not selected by chance, but by algorithms that have a principled, information-theoretic method at their disposal to distinguish informative and non-informative features and select sets of analogues accordingly.

7. References

Anttila, Raimo 2003 Analogy: the warp and woof of cognition. In: Brian D. Joseph and 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. *Journal of Linguistics* 47: 549-585.

Aronoff, Mark 1976 *Word Formation in Generative Grammar*. Cambridge, MA: MIT Press.

Baayen, Harald R 2003 Probabilistic approaches to morphology. In: Rens Bod, Jennifer Hay and Stefanie Jannedy (eds.), *Probabilistic Linguistics*, 229–287. Cambridge, MA: MIT Press.

Baeskow, Heike 2011 -Ness and -ity: Phonological exponents of n or meaningful nominalizers of different adjectival domains? *Journal of English Linguistics* DOI: 10.1177/0075424211405156.

Bauer, Laurie 1983 *English Word-Formation*. Cambridge: Cambridge University Press.

Bauer, Laurie 2001 *Morphological Productivity*. Cambridge: Cambridge University Press.

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

Becker, Thomas 1993a Back-formation, cross-formation and 'bracketing paradoxes' in paradigmatic morphology. In: Geert Booij and Jaap van Marle (eds.), *Yearbook of Morphology* 1993, 1–25. Dordrecht: Kluwer.

Becker, Thomas 1993b Morphologische Ersetzungsbildungen im Deutschen. *Zeitschrift für Sprachwissenschaft* 12(2): 185–217.

Blevins, James P. and Juliette Blevins (eds.) 2009 *Analogy in Grammar*. Oxford: Oxford University Press.

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

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

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

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

Chapman, Don and 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.

Daelemans, Walter, Peter Berck and Steven Gillis 1997 Data mining as a method for linguistic analysis: Dutch diminutives. *Folia Linguistica* 31: 57–75.

Daelemans, Walter and Antal d. van Bosch 2005 *Memory-Based Language Processing*. Cambridge: Cambridge University Press.

Daelemans, Walter, Jakub Zavrel, Ko van der Sloot and Antal van den Bosch 1999 et seq. *TiMBL: Tilburg Memory Based Learner* : Available from http://ilk.uvt.nl/timbl/.

Derwing, Bruce I. and Royal Skousen 1994 Productivity and the English past tense: Testing Skousen's analogical model. In: Susan D. Lima, Roberta Corrigan and Gregory K. Iverson (eds.), *The Reality of Linguistic Rules*, 193–218. Amsterdam/Philadelphia: 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: Benjamins.

Elzinga, Dirk 2006 English adjective comparison and analogy. *Lingua* 116(6): 757–770.

Gahl, Susanne and Alan C. L. Yu (eds.) 2006 *Special Issue on Exemplar-based Models in Linguistics*. *The Linguistic Review* (23). Berlin: Mouton de Gruyter.

Hay, Jennifer and Harald R. Baayen 2005 Shifting paradigms: gradient structure in morphology. *Trends in Cognitive Science* 9: 342–348.

Hock, Hans H 2003 Analogical change. In: Brian D. Joseph and Richard D. Janda (eds.), *The Handbook of Historical Linguistics*, 441–460. Malden, MA: Blackwell.

Keuleers, Emmanuel 2008 Memory-Based Learning of Inflectional Morphology. PhD. dissertation, Antwerpen: University of Antwerp.

Krott, Andrea 2009 The role of analogy for compound words. In: James P. Blevins and Juliette Blevins (eds.), *Analogy in Grammar*, 118–136. Oxford: Oxford Universtiy Press.

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

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

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

Marchand, Hans 1963 On content as a criterion of derivational relationship with backderived words. *Indogermanische Forschungen* 68: 170–175.

Marchand, Hans 1969 *Categories and Types of Present-Day English Word-Formation*. München: C.H. Beck.

Nosofsky, Robert M 1986 Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General* 115: 39–57.

Nosofsky, Robert M. 1990 Relations between exemplar similarity and likelihood models of classification. *Journal of Mathematical Psychology* 34: 393–418.

Pinker, Steven and Alan Prince 1994 Regular and irregular morphology and the status of psychological rules in grammar. In: Susan D. Lima, Roberta Corrigan and Gregory K. Iverson (eds.), *The Reality of Linguistic Rules*, 321–351. Amsterdam/Philadelphia: Benjamins.

Plag, Ingo 2003 *Word-Formation in English*. Cambridge: CUP.

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, Gero Kunter and Sabine Lappe 2007 Testing hypotheses about compound stress assignment in English: a corpus-based investigation. *Corpus Linguistics and Linguistic Theory* 3(2): 199–233.

Prasada, Sandeep and Stephen Pinker 1993 Generalization of regular and irregular morphological patterns. *Language and Cognitive Processes* 8: 1–56.

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

Skousen, Royal 1992 *Analogy and Structure*. Dordrecht: Kluwer.

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

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

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

Spencer, Andrew 1988 Bracketing paradoxes in the English lexicon. *Language* 64(4): 663–682.

Szymanek, Bogdan 2005 The latest trends in English word-formation. In: Pavol Štekauer and Rochelle Lieber (eds.), *Handbook of Word-Formation*, 429–448. Dordrecht: Springer.

van Marle, Jaap 1985 *On the Paradigmatic Dimension of Morphological Creativity*. Dordrecht: Foris.

van Marle, Jaap 1990 Rule-creating creativity - analogy as a synchronic morphological process. In: Wolfgang U. Dressler, Hans C. Luschützky, Oskar E. Pfeiffer and John R. Rennison (eds.), *Contemporary Morphology*, 267–273. Dordrecht: Foris.

Sabine Arndt-Lappe, Düsseldorf (Germany)