Moving spaces: Spelling alternation in English noun-noun compounds

Victor Kuperman

McMaster University, Canada

Raymond Bertram

University of Turku, Finland

Address all correspondence to:

Victor Kuperman

Dept. of Linguistics and Language

McMaster University

1280 Main Street West

Hamilton, Ontario, Canada, L8S 4M2

Email: vickup@mcmaster.ca

Phone: +905-5259140

Fax: +905-5776930

Abstract

The present study explores linguistic predictors and behavioral implications of the orthographic alternation between a spaced (*bell tower*), hyphenated (*bell-tower*) and concatenated (*belltower*) format observed in English compound words. On the basis of two English corpora, we model the evolution of spelling for compounds undergoing lexicalization, as well as define the set of orthographic, distributional, and semantic properties of the compound's constituents that co-determine the preference for one of the available realizations. We explore iconicity and economy as competing motivations for both the diachronic change and synchronous preferences in spelling. Observed patterns of written production closely mirror the demands and strategies of recognition of compound words in reading. Orthographic choices that go against the reader's economy of effort come with a high recognition cost, as evidenced in inflated lexical decision and naming latencies to concatenated compounds that occur in other spelling formats.

Keywords: compounding, production, comprehension, spelling, corpus linguistics, competing motivations

Introduction

A substantial body of linguistic literature has explored the question of what motivates the language user's choice of expression for equivalent meanings. For instance, the notion of a person giving an object to another person can be expressed via the ditransitive dative "Tom gave the boy the book" or its formally more complex (longer) prepositional counterpart "Tom gave the book to the boy". The preference for one or the other alternant in natural speech production can be predicted with an accuracy of 94% from the semantic (e.g., animacy, definiteness, pronominality or concreteness) or formal (length in words) properties of the objects (*the boy* and *the book*) and the verb (*give*) (Bresnan, Cueni, Nikitina, & Baayen, 2007; Bresnan & Ford, 2010). Likewise, the non-arbitrary relationships between the complexity of chosen syntactic alternants and semantic complexity, accessibility and informativeness of their constituents is reported in a number of other meaning-equivalent alternations: see for instance recent work on *that*-mentioning in English complement clauses (Ferreira, 2003; Jaeger, 2010; Roland, Elman, & Ferreira, 2006); particle placement in phrasal verbs (Gries, 2003; Lohse, Hawkins & Wasow, 2004; Wasow, 2002) and morpho-syntactic contraction (Bybee & Scheibman, 1999; Frank & Jaeger, 2008).

The present study adds to this research by exploring a less studied type of writers' choices at the *word level*, i.e. the spelling of English noun-noun compound words. In English, three spelling variants are possible for a compound word. A compound may appear in concatenated (*boyfriend*), spaced (*blood pressure*) or hyphenated format (*word-play*). The spelling of compounds is a source of frustration for many writers of English, as the rules that prescribe the correct spelling of compounds are neither accurate nor exhaustive. Thus, the style manual of the U.S. Government printing office devotes 13 pages to listing the rules of compound spelling, another 85 pages to examples and exceptions, and also notes that '[I]n applying the rules in this chapter... the fluid nature of our language should be kept in mind. Word forms constantly undergo modification.' (US Government Printing Office, 2008, page 62). In fact, many compounds occur in more than one alternative spelling variant (*carwash, car-wash, car wash; roleplay, role-play, role play*). Moreover, the spelling often varies from one source to another. To use a classic example from Bauer (1988), the compound *girl+friend*[1] is to be spelled as *girlfriend* according to Hamlyn's Encyclopaedic Word Dictionary; as *girl-friend* according to the Concise Oxford Dictionary (and the Oxford English Dictionary); and as *girl friend* according to Webster's Third New International Dictionary. Unsurprisingly, it has often been argued (Bauer, 1988; Jespersen, 1977) that compound

spelling in English is to a large extent inconsistent and arbitrary. We expand on earlier research (Mondorf, 2009a, 2009b; Rakic, 2009; Sepp, 2006) to show that the choice of one orthographic variant over others is not arbitrary, but is co-determined by multiple factors. Based on behavioral data, we will further argue that – to a large extent – spelling preferences in written production are motivated by the cognitive demands of online word recognition.

Explanatory factors for compound spelling

What motivates the choice of one formal alternant over others? Prior work of Sepp (2006) identified multiple distributional variables that are predictive of compound spelling, including compound frequency, constituent frequencies and biases towards one of the spelling variants within compound families (sets of compounds that share one of the constituents, e.g., *post+man, post+card, post+office*). In addition, Sepp (2006) considered a range of phonological factors, such as the length of a compound measured in letters, phonemes or syllables, stress position (see also Plag, Kunter, & Lappe, 2007) and the presence of identical phones straddling the major constituent boundary (*lamp+post*). Sepp found a strong tendency for long compounds (with length measured in letters, phonemes or syllables) to appear in a non-concatenated (spaced or hyphenated) format. This finding converges with Rakic's (2009) observation that compounds are rarely concatenated when their constituents, especially the left constituents, have a complex morphological structure (*apartment block*). Overall, however, Sepp reports phonological factors to be less influential than distributional measures in predicting the orthographic alternation.

In the present study we expand on the study of Sepp (2006) by considering a larger number of factors that may be predictive of the orthographic choices of writers in English compound spelling. Thus we add semantic factors (i.e., the semantic distance between constituents) as a potential predictor of compound spelling preference. We also make use of a larger corpus of English compounds (about 1.5 million tokens vs about 260,000 tokens in Sepp, 2006) and employ statistical techniques that estimate the magnitude and direction of effect for each predictor rather than estimating the amount of variance explained by one group of predictors over and above the other as in Sepp (2006). Finally, we only consider compounds attested in multiple formats to tap directly into language variation, whereas Sepp (2006) considered both alternating and non-alternating compounds in her study.

On a theoretical level, the question of what affects the form choice, or broader – how language use affects language structure – is often couched in terms of typological principles of iconicity and

economy (see Haiman, 1983; Hawkins, 2004). Iconicity can be defined as a notion that more complex, more independent and less cohesive meanings tend to correlate with an increased complexity or distance in the linguistic form, e.g., non-concatenated spelling (cf. Haiman, 2008; Mondorf, 2009a, 2009b; Newmeyer, 1992). An alternative motivation that does not require the speaker to directly assess or accommodate semantic properties in language behavior is economy, or the tendency to associate frequently expressed meanings with forms less effortful for production. The effect of economy is achieved either by reducing the form of expression for the meanings that are in frequent use (see Zipf, 1935) or as a by-product of an easier, routinized activation and retrieval of lexical meanings that are simpler and well entrenched in the mental lexicon (e.g., Bybee, 2003). For a theoretical debate on, and an overview of empirical evidence for, the roles of iconicity and economy in language change, the reader is referred to Croft (2003; 2008), Haiman (2008), and Haspelmath (2008). Our data will then contribute to the theoretical debate on the respective roles of iconicity and economy as underlying competing motivations that are argued to be at work in both syntactic alternation, lexical choice and, the phenomenon at hand, i.e. the orthographic alternation in compounds (Mondorf, 2009a,b; Sepp, 2006).

Evolution of compound spelling

As our second goal, we concentrate on the role of compound frequency on the choice of compound spelling and in the change of compound spelling over time. A common notion is that higher-frequency compounds are often concatenated (e.g., *boyfriend*), whereas low-frequency and newly formed compounds are often not (e.g. *youth team*). A corpus study by Sepp (2006) confirms the tendency in English as she observes a negative correlation between log compound frequency and the degree of formal separateness, evidenced by the presence of a space or a hyphen. It thus seems that the evolution of compound spelling proceeds from more effortful productions with an extra symbol (space, hyphen) to less effortful, concatenated ones. English grammars propose an even more specific three-stage orthographic evolution of a compound as a function of its increasing frequency of use: from the spaced representation as two separate words, to a hyphenated representation as one word but with a clear separating cue, and to a concatenated representation as one word (cf. Borjars & Burrige, 2001; Crystal, 2001; Huddleston, 1984; Quirk, Greenbaum, Leech, & Svartvik, 1985; all cited from Shie, 2002).

In this study we will investigate the evidence that lexicalization of a compound receives formal expression in the evolution from spaced to hyphenated to concatenated format over the years. We will pursue this question in three ways, by considering compounds with alternating spelling from a

synchronous corpus of Wikipedia collected in 2008, and from a diachronic corpus consisting of 20 years of the New York Times newspaper texts (1986-2007). First, we will assess whether there is a correlation between the frequency of use of a compound and its preferred orthographic format. On the three-stage model, one would expect that most of the high-frequent compounds are concatenated, that medium-frequent compounds are typically hyphenated and that most of the low-frequent compounds are spaced. Second, we will trace over time spellings of those compound words that gained substantially in their frequency of use and we will examine whether, for one and the same word, lexicalization comes with a change in word spelling. Lexicalization and compound frequency as its distributional proxy have direct bearing on the expected economy of effort: thus, the observed temporal and frequency-driven changes in the compound spelling will also shed light on the motivations of meaning-equivalent choices.

The impact of spelling alternation on compound processing

Another question we pursue is whether spelling alternation of compound words has implications for online processing behavior. More specifically, we ask whether the processing of a compound in one spelling (*belltower*) is affected by the existence or frequency of that compound in alternative spellings (*bell-tower, bell tower*). An eye-tracking study of Cherng (2008) suggest that the effect is in place, as she found hyphenated compounds to be processed slower than concatenated counterparts, when the concatenated variant was more frequent. Yet both variants were processed equally fast when the concatenated variant was less frequent than the hyphenated one. The relevance of the spelling choice becomes even more evident in studies that present normally concatenated compounds with a space or a hyphen inserted between constituents. Inhoff and Radach (2002) showed that initial landing position on the first constituent shifts to the left when inserting hyphens into German compounds. Other studies report that a space or a hyphen serve as excellent cues for a fast segmentation of a compound into constituents and swift *extraction* of constituent meanings, as reflected in early eye movement measures such as first fixation duration. However, the same studies show that the space hampers the *integration* of constituent information, leading to longer processing times at later stages of word processing (see Inhoff, Radach, & Heller, 2000; Juhasz, Inhoff, & Rayner, 2005). Bertram et al. (2011) and Bertram and Hyönä (in submission) further demonstrate that hyphenation does not hamper integration of constituent information, at least not in long triconstituent or biconstituent Finnish compounds.

Even if processing differences are expected for available spelling formats, it can be hypothesized that readers are able to optimize their processing strategies for that compound over time when a

compound consistently appears in the same format (e.g., *boyfriend* always written without a space, or *amusement park* always with a space). In the first case, readers would learn that *yf* marks a constituent boundary, and in the second they would be aware of the possibility that the left constituent is followed by another noun to be integrated into a compound. Conversely, if a reader encounters one and the same compound alternatively in concatenated, spaced or hyphenated format, the frequency of any given format is lower, leading to weaker learning effects. It also becomes more complicated to fine-tune the processing system for how to perform segmentation or how/when to integrate constituent meanings. In the current study, we extracted lexical decision and naming latencies for compounds with alternating spellings from the English Lexicon Project (Balota et al., 2007) to explore whether the presence of alternatives influences the processing of concatenated compounds.

To sum up, the goals of the study are threefold. First, we identify orthographic, phonological, semantic and distributional explanatory factors affecting the preference writers show for one of the three spelling variants in English compounds. Second, we test whether there is a route to lexicalization and if there is, whether this route goes via the hyphenated format. Third, we examine the processing consequences of presenting compounds that can occur in more than one spelling variants. We address the first two questions by analyzing alternating compounds attested in Wikipedia and the New York Times corpora. Subsequently, we turn to behavioral data from the English Lexicon Project to answer the third question.

Data sources and data trimming

English Wikipedia (2008)

We opted for using Wikipedia in the English language as one of our data sources. The English Wikipedia is several orders of magnitude larger than the corpora used for previous studies of spelling alternation in compounds (cf. the 1.2 billion tokens in Wikipedia vs the 14 million-token corpus in the combination of corpora used by Sepp, 2006, or the 230,000 word types of the Longman Dictionary of Contemporary English used in Rakic, 2009). While the thematic coverage of Wikipedia is difficult to assess, its genre of encyclopedia is likely to grant it a broadly distributed thematic representation: this helps to avoid an overrepresentation of domain-specific compounds (see Sepp, 2006, for exclusion criteria). Editing is allowed to Wikipedia contributors, yet no

uniform or shared set of spelling rules regarding compounding is prescribed (see Wikipedia Manual of Style at http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Spelling). Thus any written compound represents an idiosyncratic decision of a language user, be that person an original contributor or a voluntary editor. There is no easy way to ensure that textual contributions to the English-language Wikipedia are only supplied by the speakers of any native variety of English; the same holds true for most online resources that are open for public access and contribution. The existence of Wikipedia in languages other than English may diminish the proportion of non-native linguistic input to Wikipedia in English, hence one cannot deny the possibility of non-native influences on the spelling of compounds in our source. We note, however, that even if produced by a non-native speaker of English, a spelling alternative strengthens orthographic representations of native speakers of English as well, and as such is of potential influence for the spelling decisions that native English speakers make. Thus, we conclude that - to a first approximation - Wikipedia in English is an adequate data source for our purposes.

The data source for the collection of lexical materials was the October, 2008 repository of Wikipedia in the English language (1.2 billion tokens). All Wikipedia texts were parsed using the Stanford Parser (Klein & Manning, 2003) and supplied with the word-by-word part-of-speech tags and other syntactic information. We extracted sequences of two words both tagged as common singular nouns, where the words were separated either by a space or a hyphen: these were identified as spaced or hyphenated noun-noun compounds, respectively. We further searched Wikipedia for word forms obtained by the removal of a hyphen or a space from the two spelling formats described above. This procedure identifies all those concatenated compounds that alternate with either spaced or hyphenated compounds, or both[2].

Our focus on singular nouns as constituents stemmed from the evidence that lexical processing and potential spelling patterns differ for cases where one of the constituents is plural or in pluralia tantum (e.g., *scissors, glasses*) (e.g., Cunnings & Clahsen, 2007; Rakic, 2009). Major regional discrepancies in spelling were identified using the Wikipedia style guidelines (en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Spelling) and homogenized into the US American spelling to enable compatibility with the English Lexicon Project word list. The retrieved results were further filtered using the following criteria: a sequence of two nouns was removed if preceded or followed by another noun (e.g., *birthday party* or *party hat* in *birthday party hat*) or a

numeral (e.g., *century drama* in *twentieth century drama*). We also excluded synthetic compounds formed by adding the suffix *-ing* to verbs (Fabb, 1998: 68; Lieber 1983) (e.g., *house+warming*). As argued in Sepp (2006), the derivational character of deverbal compounding may set it apart from nominal compounding in that phonological properties of the former compounding type are weaker predictors of its spelling (see also Mondorf, 2009b). Finally, a spelling type for a compound was considered attested if it occurred 10 times or more in the corpus. The cut-off frequency helped to weed out cases of misspelling, abundant in the corpus.

After the filtering procedures and the frequency cut-off were applied, the type count (rounded to hundreds) for the three spellings was as follows: 75,000 spaced compounds, 2,900 hyphenated compounds, and 1,600 concatenated compounds (see footnote 2 for a corrected count of concatenated compounds). We further compiled a subset of compounds that alternated in spelling, i.e. were attested with the above-the-cut-off frequency in any two or all three spelling variants. Finally, we manually went through the subset to remove cases of misspellings or erroneous part-of-speech tagging (e.g., *post-modernism* tagged as a noun-noun compound). The resulting list of compounds with alternating spelling comprised 2,306 types, accounting across all spelling variants for 1,484,175 tokens.

The New York Times (1987-2006)

The New York Times (NYT) corpus contains 1.8 million documents spanning over 20 years (1987-2006). We searched for all compounds that were attested as alternating in Wikipedia (see selection criteria above) and checked whether they appeared in NYT in more than one variant with the frequency of 10 occurrences of above. We further tracked the spelling of those compounds that underwent a drastic increase in frequency [factor of 2 in log frequency] in this time period. Some of them were newly coined words (*cyber+cafe*), others were broadly attested even in the beginning of the 20-year span (*data+base*). Naturally, the NYT documents are edited, so most compounds are spelled as the guidelines suggest and offer no variation in their spelling over time. Yet, as we demonstrate below, there were a number of compounds that substantially varied in their spelling over the course of the 20 years or even elicited a change in editorial spelling policies.

A route to lexicalization

In order to answer the question whether there is a route to lexicalization and if so, whether it is mediated by hyphenation, we first considered the distributional data of alternating variants in the English Wikipedia. We explored what pairs of spelling variants in the Wikipedia corpus are more likely to alternate by collecting the statistics on the number of types for compounds that represent every kind of the two-way alternation (spaced vs concatenated; concatenated vs hyphenated; spaced vs hyphenated) as well as compounds that are found in all three spelling variants. The findings are reported in Table 1.

INSERT TABLE 1 ABOUT HERE

The descriptive statistics in Table 1 shed light on the issue of how the process of lexicalization affects spelling of compounds. Spaced compounds alternated with both concatenated and hyphenated compounds, with the former alternation being significantly more frequent (984 vs 856. $\chi^2 = 8.9$, $p = 0.002$). There was also a substantial number of compounds that took all three variants (432). The type count of compounds that were found in both the concatenated and hyphenated form (e.g. *band+pass, show+piece, side+saddle*) was extremely small (34) and was significantly smaller than expected given the prevalence of both formats in other alternations ($p < 0.0001$ in all chi-squared tests). The fact that the most frequent alternation is between spaced and concatenated compounds provides evidence that lexicalization favors the direct transition from spaced to concatenated spelling. Also, given that alternation between concatenated and hyphenated compounds hardly occurs, one may conclude that there is no 'free-flowing' alternation between hyphenated and concatenated forms. They only co-occur if a spaced variant is also attested for that compound, which is likely a sign of the simultaneous alternation between spaced compounds and each of the two other variants. Taken together, the patterns suggest that the route towards lexicalization does *not* implicate the mediation of a hyphen.

This conclusion is supported when considering the full model of explanatory factors for the preference writers show for one of the three spelling variants in English compounds in Wikipedia. We will discuss the full models in detail below, but will present here the results relevant to the question of lexicalization. We found that frequency of a compound computed across spelling variants (JointFreq) affected the choice of a spelling in the way suggested by the distributional analyses. For the compounds that alternated between the concatenated and spaced variant, higher-frequency compounds were more likely to be spelled as one word [log compound frequency: $\beta = 0.72$; $SE = 0.05$; $p < 0.01$], see Table 2. For the compounds that alternated between the hyphenated

and spaced variant, a higher log frequency of the compound (JointFreq) came with a bias towards spacing [log compound frequency: $\beta = 1.19$; SE = 0.04; $p < 0.01$], see Table 3. That is, compounds were more likely to become spaced rather than hyphenated as their frequency increased. We will further discuss the latter – slightly unexpected result – in the General Discussion. For now, we note that the results clearly go against the three-stage spelling evolution proposed in prior literature (Shie, 2002 and references therein).

In order to zoom in on the evolution of compounds from spaced to concatenated format as a function of frequency, we extracted the total of 24 compounds from the NYT that alternated in spelling and underwent a drastic increase in frequency in a period of 20 years (see above for selection criteria)[3]. We remind the reader that even a small number of compounds that alternate multiple times is surprising to find in this data source, as any case of spelling alternation violates the principle of uniformity imposed by stringent editorial guidelines of NYT. In convergence with the results obtained from Wikipedia, the most frequent alternation was between spaced and concatenated compounds (18 compounds, see Figure 1). Only two compounds alternated between the concatenated and hyphenated formats (*care+giver, show+biz*), but these compounds were also documented in the spaced format with an above-the-cutoff frequency. Finally, 6 compounds ("jump+start", "race+day", "salary+cap", "die+cast", "fund+raiser", "call+center") alternated between spaced and hyphenated formats. Again, this grants no support to the view that lexicalization proceeds from spaced to concatenated compounds via the hyphenated variant.

In a more direct test, we computed correlations between the log frequency of the compound per year in all formats, and the bias towards concatenated format (computed as the compound's frequency of concatenated (unspaced) use, BiasU, divided by compound frequency in all formats per year). Sixteen of 18 compounds started off as typically spaced and ended up with an increased bias towards concatenated spelling: the Pearson's correlation coefficients ranged from 0.35 to 0.94 (all ps < 0.05). The exceptions were *shoe+box* and *help+line* which did not show a reliable tendency ($r = -0.07$, $p = 0.7$, and $r = -0.37$, $p = 0.10$, respectively). Figure 1 (panel A) plots the change in compound frequency and Figure 1 (panel B) plots the bias towards concatenation for these compounds (an increase, with the exception of *helpline* and *shoebox*).

INSERT FIGURE 1 ABOUT HERE

These correlations offer direct evidence for the spelling change in English compounds over time. Taken together with the data from the English Wikipedia, these observations bear witness to the preferences in spelling choices from spacing to concatenation as a function of compound frequency. Contra a long descriptive tradition, we do not find evidence in contemporary English for hyphenation as an obligatory, or even common, stage in the spelling change. The NYT data also illustrates how frequency of use affects editorial policies and styles: similar analyses can be used to investigate the time lag between the increase in frequency and the change in the spelling conventions.

Explanatory factors for compound spelling: The full model

One of our goals was to specify the linguistic variables that co-determine the likelihood of the choice between three available spelling variants: spaced, hyphenated and concatenated. As will be explained below, this ternary choice can be construed as a set of binary choices without reducing the accuracy of predictions. Thus, we estimated the likelihood of the binary choices by fitting logistic regression models to subsets of data with compounds that represent two of three spelling variants (and are not attested with the third variant).

Dependent variables

The dependent variable required for logistic regression models is the number of successes (i.e. the frequency of a compound's occurrences in one spelling variant) and the number of failures i.e., the frequency of the compound's occurrences in the opposite spelling variant). For instance, the compound *apple+sauce* was attested 27 times as concatenated (*applesauce*) and 27 times as spaced (*apple sauce*), and 0 times as hyphenated. The tuple (27, 27) would define the dependent variable for this compound in the model fitted to 984 compound types demonstrating the spaced-concatenated alternation. We also reported below a model fitted to 856 spaced-hyphenated compounds. We did not pursue characterization of predictors for the hyphenated-concatenated alternation that was only attested in 34 instances, as any modeling effort would severely overfit the data.

Independent variables

Distributional variables

As reported in Sepp (2006), frequency and related measures are influential predictors of the orthographic representation of compounds. We considered the frequency of a compound (*ball+point*, 167), computed across all spelling variants (*ball point*, 31; *ball-point*, 10; *ballpoint*, 126), labeled JointFreq:. Frequencies and length of constituents (*ball* and *point*) were considered too, labeled as LeftFreq, LeftLength, RightFreq and RightLength. We also calculated the family sizes (and frequencies) of the compounds' left constituents as the number of types of compounds (or tokens of compounds) in the alternating subset of Wikipedia that share the left constituent, across all spelling variants. For instance, the left constituent family of *bird* contained *bird+cage*, *bird+house*, *bird+life*, *bird+seed*, and *bird+song* and so had the family size of 5 and the family frequency (the summed frequency of compounds in the family) of 590: the labels we used for these variables were LeftFamSize and LeftFamFreq. The same calculations were made for right constituent family sizes and frequencies (RightFamSize and RightFamFreq).

We calculated the amount of paradigmatic support for, or the family bias towards, each of spelling variants for a given left and right constituent. For instance, there were 5 types in the left constituent family of *bird* that were attested with the concatenated spelling, 5 with the spaced spelling, and 0 with the hyphenated spelling: the respective numbers of tokens were 343, 247, and 0. Token-based measures of the family bias proved to be more predictive of the spelling alternation: in what follows we only report these. Appendix 1 provides details on how the family bias towards each of spelling variants was computed. All frequency-based measures were calculated on the basis of the Wikipedia corpus (1.2 billion tokens). To reduce the skewness in the distribution of compound and constituent frequencies, as well as family sizes and frequencies, we used log-transformed values.

Orthographic and phonological variables

The orthographic length of the compound is known to be a major determinant of the spelling choice in English (Sepp, 2006; Rakic, 2009). We considered orthographic lengths of constituents, measured in characters. Zooming in on constituents, rather than the compound as a whole, additionally shed light on whether it was the length of the modifier (typically, the left constituent) or that of the head (typically, the right constituent) that influenced the choice of spelling. Constituent lengths were also measured in phonemes and syllables: as these measures did not explain any variance over and above orthographic length, we only report the effects of the latter.

We also coded every compound for whether there were two or more identical characters straddling the constituent boundary (e.g., *boat+tail*). A similar coding was made to mark the cases of identical sounds straddling the constituent boundary (e.g., *bomb+maker*). Furthermore, as suggested in Sepp (2006), we introduced a set of binary indices that reflected whether there was a consonant cluster at the end of the left constituent, a consonant cluster at the beginning of the right constituent, or a consonant cluster was formed across the constituent boundary: none of these factors proved influential for spelling choice.

The choice of spelling in compounds correlates with the stress pattern that the compound adopts, as demonstrated in Plag et al. (2007). Since the direction of causality is unclear in this pair of variables, we did not consider compound stress or primary stresses of constituent words as predictors of compound spelling.

Semantic variables

One of predictions that the principle of iconicity makes is that a larger conceptual distance between meanings of compounds and their constituents would be reflected in a larger linguistic or formal distance (e.g., Haiman, 1983; Mondorf, 2009a). Thus, constituents that are semantically more congruent with each other (e.g., *window* and *pane* in *window+pane*) would tend to concatenate, whereas conceptually distant constituents (*dead* or *line* in *dead+line*) would more often elicit spacing or hyphenation. We note that no clear prediction can be made for the alternation between two non-concatenated formats, as either a hyphen or a space signal formal separation of constituents, and which of them signifies a larger linguistic distance is open for discussion. We opted for Latent Semantic Analysis (LSA), a computational technique for estimating similarity between words, as a numeric index of conceptual distance (Landauer & Dumais, 1997). LSA scores range from -1 to 1, with a higher score standing for a stronger semantic similarity and shorter semantic/conceptual distance between two words. For an overview of evidence that LSA scores are adequate indices of conceptual similarity of words, collocations or texts, see Landauer, Foltz and Laham, 1998; for evidence on morphologically complex words, see e.g., Gagné & Spalding, 2009; Milin, Kuperman, Kostic, and Baayen, 2009b; Moscoso del Prado Martín et al., 2005; Rastle, Davis, & New, 2004. LSA scores for the similarity of the left and right constituents (modifier vs head, tooth vs paste) were obtained for 1844 out of 2306 compounds in our data set. The source for scores was the LSA pairwise comparison utility available at http://LSA.colorado.edu and used with the default option of General Reading up to the 1st year of college and 300 factors (i.e., each word is associated with a vector in the 300-dimensional semantic space, cf. Landauer & Dumais, 1997).

Statistical considerations

While there are three spelling variants in English compounds, we demonstrated above that the observed spelling alternation is reducible to two binary choices (spaced vs concatenated, and spaced vs hyphenated), as there is no evidence for the concatenated-hyphenated alternation. We used library lme4 of the statistical package R to fit logistic regression mixed-effects models with the binomial link function to estimate the influence of multiple predictors on each of those binary choices: for a detailed treatment of the logistic regression methodology as applied to linguistic data see Baayen (2008) and Jaeger (2008). The models defined compound as a random effect, which enabled the estimation of systematic effects of linguistic predictors over and above the compound-specific variability in the preferred spelling variants. No other random effects or interactions were supported by the log-likelihood model comparison test. We standardized all numerical predictors by subtracting the mean value of the predictor from the original value and dividing the difference by one standard deviation of the predictor: standardization allows for comparing the relative strengths of effects of predictors on the dependent variable. We further applied the step-wise backward elimination procedure, which started off with the full set of predictors and removed one-by-one those predictors whose removal did not significantly decrease the model's performance. The relative model's performance was estimated via the log-likelihood model comparison test that compared at each step a model with a given predictor to a model without that predictor. Below we report the outcomes of the elimination procedure, i.e. the models with the predictors that survived the step-wise model comparison test.

Results and Discussion

Spaced vs Concatenated Alternation

Table 2 summarizes the regression analysis of the 984 compounds alternating between the spaced and concatenated format (but not found in the hyphenated format). The best-performing model of the spaced/concatenated alternation solely included distributional and orthographic properties of compounds and their constituents, and none of the many phonological properties we considered. This runs counter to the results of Sepp (2006) who reported strong predictivity of phonology over and above the influence of lexical or orthographic properties.

The orthographic length of the compound has long been known to predict its spelling: unlike other Germanic languages, compounds longer than 3 syllables are rarely concatenated in English (Rakic, 2008). Our results suggest that spacing is predicted stronger by the length of the left constituent (the compound's modifier), rather than by the length of the right constituent [left constituent length: β = -0.25; SE = 0.05; p < 0.01].

INSERT TABLE 2 ABOUT HERE

As discussed above, frequency of a compound computed across spelling variants, JointFreq, affected the alternation in the way predicted by a number of earlier reports (e.g., Sepp, 2006): The more frequently the compound was used, the more likely it was to be spelled in concatenated format [log compound frequency: β = 0.72; SE = 0.05; p < 0.01]. Also, the compound's frequency of occurrence was the strongest predictor of this alternation, as was evident from the largest absolute value of its regression coefficient among linguistic predictors in the model.

Higher values of both the left and the right constituent frequency came with a higher likelihood of producing a spaced compound [log left constituent frequency: β = -0.15; SE = 0.05; p = 0.01, log right constituent frequency: β = -0.11; SE = 0.05; p = 0.04].

The preference for one spelling variant in the compound's left/right constituent family strongly biased that compound towards the same variant. In our model, a stronger bias towards spacing in the family came with a bias towards spacing in the compound [left constituent family frequency-based bias: β = -0.16; SE = 0.05; p < 0.01, right constituent family frequency-based bias: β = -0.40; SE = 0.05; p < 0.01]; likewise, a stronger bias towards concatenation in the family translated into a more likely choice of concatenation in the compound (model not shown).

Finally, we tested whether conceptual distance between constituents of the compound would influence the orthographic distance between those constituents. LSA scores were available for pairs of constituents in 1844 of our compounds: of these, 791 compounds were attested with the above-the-cutoff frequency in the spaced and concatenated, but not in the hyphenated, variant. The logistic regression mixed-effects model fitted to the 791 compounds showed a reliable effect of LSA constituent-constituent scores on the spelling choice: a higher score (corresponding to a stronger semantic similarity) increased the likelihood of spaced spelling [LSA score: β = -0.15; SE = 0.06; p = 0.01]. Our analysis of the correlations between LSA scores and distributional or ortho-

phonological properties of compounds and their constituents ruled out the possibility that the effect of semantic similarity is a frequency or length effect in disguise: all absolute values of Pearson's correlations were below < 0.10. This effect went in the *opposite* direction from the one predicted by the principle of iconicity: we return to this finding in the General Discussion.

We tested whether our findings would hold if we fitted the model to all cases of spaced-concatenated alternation, regardless of whether or not compounds could also be hyphenated. The resulting data set contained 1416 compounds, and the logistic regression mixed-effects model we fitted to the data set was nearly identical to the one in Table 2 in terms of the magnitude, direction and statistical significance of regression coefficients. The correlation between regression coefficients in the two models was nearly perfect ($r = 0.99$; $p < 0.0001$). We conclude that our conclusions regarding the spaced/concatenated alternation are generalizable over a larger set of compounds. The stability of results in the larger data set where hyphenation was a viable option for production lends further support to our distributional data above, showing that the phenomenon of alternation between spaced and concatenated compounds does not interact with, but is merely parallel to, alternation between spaced and hyphenated spelling variants.

Spaced vs Hyphenated Alternation

Table 3 summarizes the output of the logistic regression model fitted to the subset of 856 compounds which are found in spaced and hyphenated but not in concatenated format. The model in Table 3 confirmed the prevalence of orthographic and distributional factors, as compared to phonological ones, as co-determiners of the spelling variation. Longer constituents came with a preference for spacing, rather than hyphenation. While constituent lengths were only marginally significant in this data set [left constituent length: $\beta = -0.07$; SE = 0.04; $p = 0.08$, right constituent length: $\beta = -0.06$; SE = 0.04; $p = 0.11$], they are well below the conventional threshold of significance in a larger data set based on all 1288 compounds that alternate between spacing and hyphenation, regardless of whether the compounds are also attested in a concatenated variant (both $ps < 0.0001$): the lower p-value in the larger dataset was probably due to increased statistical power.

INSERT TABLE 3 ABOUT HERE

As noted earlier, a higher log frequency of the compound, calculated across all spelling variants, came with a bias towards spacing [log compound frequency: $\beta = -1.19$; SE = 0.04; $p < 0.01$].

We observed the analogical effects of the morphological family of both the left and the right constituent, similar to the ones found in the spaced/concatenated alternation [left constituent family frequency-based bias: $\beta = 0.29$; SE = 0.04; $p < 0.01$, right constituent family frequency-based bias: $\beta = 0.12$; SE = 0.04; $p < 0.01$]. A stronger token-based bias towards hyphenation in the paradigm (family) of compounds that shared the left/right constituent predicted a stronger bias towards hyphenation in a compound that belonged to that family: same held for the bias towards spacing.

Finally, semantic similarity of the left and the right constituents, measured via their LSA scores and indicative of the conceptual distance between constituents, was not predictive of the spaced/hyphenated alternation. As discussed above, the predictions of the iconicity principle are unclear here, as constituents are arguably as distant from each other orthographically in spaced and hyphenated compounds.

Expansion of the data set from 856 compounds that alternate only between spacing and hyphenation to 1288 compounds that show same alternation but were also attested in the concatenated form rendered statistically significant (at the 0.01 level) the effects of constituent lengths, see above. Effects of other predictors on alternation in the larger set were nearly identical in magnitude, direction and statistical significance to those reported in Table 3. The correlation between regression coefficients in the models fitted to a smaller and a larger data set with the spaced/hyphenated alternation was nearly perfect, $r = 0.98$; $p < 0.0001$].

To sum up, the models for both spaced/concatenated and spaced/hyphenated alternations show that the orthographic choice in written production of meaning-equivalent words is statistically co-determined by a number of simultaneous operating orthographic, distributional and semantic factors.

Behavioral Data

The fact that readers are confronted with alternating variants of a compound may have an impact on their processing efficiency. Fine-tuning the system for processing a compound in a certain format is more problematic if that compound often appears in other formats. Alternation may also increase uncertainty in tasks like lexical decision, where a memory of encountering a compound in multiple formats may lead to longer or less accurate judgments about the lexicality of any given format. Finally, exposure to a compound in multiple formats will leave weaker traces in the orthographic memory for any given format, as compared to the trace left by the same amount of exposure to an orthographically invariable compound. In a task like word naming the inferior familiarity with the given spelling might also lead to inflated naming latencies. In what follows we report lexical decision and naming data from the English Lexicon Project indicating that the processing of concatenated English compounds is affected by the number and frequency of the alternating variants lingering in the reader's lexical memory.

Lexical Decision

We used the English Lexicon Project database (ELP; available at http://elexicon.wustl.edu/) to extract mean lexical decision and naming latencies for noun-noun compounds that have alternating spellings. We obtained mean lexical decision response times for 503 concatenated compounds (for details of lexical decision data collection and experimental procedure, see Balota et al., 2007). The critical predictor for the RTs was the bias towards concatenated spelling (BiasC) established using the Wikipedia statistics. Some of the compounds documented in the ELP as concatenated were not attested as such even once in the English Wikipedia (e.g., ticker+tape, shell+shock), most concatenated ELP compounds were however more likely to appear as concatenated in Wikipedia too. The bias ranged from 0 to 0.98 in the ELP dataset (mean = 0.74; SD = 0.25). Other predictors included: compound frequency (JointFreq), frequencies and lengths of the left and the right constituents (LeftFreq, LeftLength, RightFreq, RightLength), and family frequencies of the left and the right constituents (LeftFamFreq and RightFamFreq). We (natural-)log transformed lexical decision latencies, as well as all frequency-based measures in this dataset to attenuate the influence of outliers on the predictions of statistical models: the prefix "l" was added to labels of transformed variables. As shown above, the bias towards concatenation strongly correlates with many of the distributional factors that we include as covariates in our regression model. Such collinearity is known to affect the estimates of standard errors in regression models and thus may decrease the accuracy of inferential estimates for the predictor of critical interest (see Baayen, 2008 for detailed discussion). To avoid collinearity with bias, we partialed out the influence of all control predictors

from our critical variable, the bias towards concatenation. This was achieved by fitting a linear regression model to BiasC with predictors as described above and considering residuals of this model (rBiasC) as a good approximation of original estimates of concatenation bias (Pearson's correlation $r = 0.87$, $p < 0.01$), minus potential confounds.

INSERT TABLE 4 ABOUT HERE

Residuals of the multiple regression model for durations were almost always skewed. To reduce skewness and to eliminate overly influential outliers, we removed outliers from the respective datasets, i.e., points that fell outside the range of -2.5 to 2.5 units of SD of the residual error of the model. Once outliers were removed, the model was refitted to the remaining 495 datapoints. Table 4 summarizes the final model and Figure 2 depicts the partial effects of several variables.

The model in Table 4 and the plots in Figure 2 reveal a very reliable effect of (residualized) bias towards concatenation on mean lexical decision latencies [$\beta = -99.21$; $SE = 19.14$; $p < 0.001$]. The contrast in lexical decision latency between concatenated compounds with the maximum and the minimum bias towards concatenation – with other predictors held constant - was estimated at 100 ms. In magnitude, this effect was on par with the effect of compound frequency, and stronger than the effects of constituent lengths and family frequencies. The contrast between extreme values of the bias was as strong as the model-estimated contrast between 4-letter and 8-letter words. Our corpus data above suggest that there are about 2500 compounds that alternate in spelling in English. Given the observed strong effect of the concatenation bias on lexical decision latencies to alternating compounds, it is exceedingly clear that for such compounds, it is crucial to take into account in experiments how biased the compound is toward the chosen spelling (see also Cherng, 2008 for converging eye-movement data).

INSERT FIGURE 2 ABOUT HERE

Word Naming

We further zoomed in on the effect of spelling bias on the response latencies to word naming task reported in ELP for the same set of 503 concatenated compounds. Again, the critical predictor of interest was the residualized bias towards concatenation (see above). After the same trimming procedures as described above, the final multiple regression model was fitted to mean naming

response times to 495 compounds. Table 5 summarizes the output of the regression model with predictors whose removal substantially decreased the model's performance, as indicated by the likelihood ratio test.

INSERT TABLE 5 ABOUT HERE

Figure 3 also plots partial effects of influential predictors in the regression model fitted to mean naming latencies.

INSERT FIGURE 3 ABOUT HERE

The higher (residualized) bias towards concatenation came with faster mean naming times [β = -44.65; SE = 12.78; $p < 0.001$], while the estimated contrast between compounds that were least and most likely to be concatenated was 45 ms.

The fact that there were no spaced compounds in the ELP word list may have resulted in the change of reader's expectations over the course of the naming or lexical decision experiment[4]. The probabilities of compound formats that are derived from occurrences of those formats in natural texts would change to the expected zero probability of spaced or hyphenated compounds and the probability of 1 of concatenated compounds. In other words, it could have well been that concatenated compounds that frequently alternate in natural texts are processed more efficiently by the end of the experiment than in the beginning of the experiment. In this case, one would expect an interaction between the compound's bias towards concatenation and its position in the (randomized) experimental list, measured as trial number (see Bertram et al., 2011, for a similar interaction between trial number and compound presentation format in triconstituent Finnish compounds). No such interaction, however, reached significance either in lexical decision data or word naming data. Possibly, one's exposure to compounds was too diluted in the course of an experiment by other word and non-word stimuli to alter the expected probabilities of available spelling formats.

Our analyses of two most commonly employed word recognition tasks indicate that the presence and frequency of alternatives affect the processing of compounds in concatenated spelling. These findings seem to tie in with our hypothesis that readers are not able to optimize their processing strategies for a concatenated compound over time when that compound does not consistently appear in such format. More specifically, it may be more complicated to fine-tune the processing system for how to perform segmentation, as in the alternative formats more salient segmentation cues (space or hyphen) are present. In addition, alternation may increase uncertainty regarding lexicality

of a concatenated word and hamper the activation of the word's phonological representation, translating into inflated lexical decision and naming latencies[5]. Finally, if a reader encounters one and the same compound alternatively in concatenated, spaced or hyphenated format, the frequency of any given format is lower, leading to weaker learning effects.

While potential causes of the observed effects are many, it is clear that these effects challenge the work done on English compound processing up to now, as the bias towards the presented spelling is a factor that has not been taken into account in practically any of the compound studies in English (e.g., Andrews, Miller, & Rayner, 2004; Inhoff, Starr, Solomon & Placke, 2008; Juhasz, 2008; for an important counterexample see Cherng, 2008). In fact, it may be that some of the mixed results in English compound processing stem from not taking this factor into account. For instance, the first constituent frequency effect on reading times is not stable across studies of English compounds. Some studies report a first constituent frequency effect as reliable (Inhoff et al., 2008; Juhasz, 2008, for short compounds), while in others the effects is marginal at best (Andrews et al., 2004; Juhasz, 2008, for long compounds). Likewise, the effect of semantic transparency on eye-movement measures was not reliable for concatenated compounds in Frisson et al. (2008; Experiment 1), reliable for concatenated compounds in Juhasz (2007), and reliable for spaced compounds in Frisson et al. (Experiment 2). If the frequency or semantic transparency effects in English interact with the bias towards the presented format, the composition of the target word list (likely, a mixture of alternating compounds with a range of biases and non-alternating compounds) would affect the resulting pattern of findings.

GENERAL DISCUSSION

Even though compound spelling is a source of frustration for many writers of English, it is by no means the case that the orthographic choice made by writers of English is arbitrary. In what follows, we summarize the corpus-based and behavioral evidence that spelling is in fact a reflection of lexicalization in compounds, as well as the distributional, orthographic and semantic salience of its constituents. We also discuss iconicity and economy as theoretical motivations for our findings.

Evolution of compound spelling

Our first point of interest is in the orthographic correlates of *lexicalization*, a process of a word's entrenchment in the lexical memory of the speaker's community which is commonly driven by the increased frequency of the word use (e.g., Brinton & Traugott, 2005; Bybee, 2003). The dominant view in the literature advocates the three-stage orthographic reflection of lexicalization: from spaced (*girl friend*) to hyphenated (*girl-friend*) to concatenated (*girlfriend*). One proposed motivation for such a change is iconicity, or the notion that the formal distance (visually maximal in spaced compounds, medium in hyphenated ones and minimal in concatenated ones) reflects the degree of complexity, cohesion or independence of the compound as a whole or constituents within a compound (see Mondorf, 2009a). Thus, the three-stage change in spelling as compounds become increasingly more integrated in the lexical system of the language would be in line with the principle of iconicity of independence (Haiman, 1983; Lohmann, 2011).

A proposed alternative to iconicity is economy: see Sepp (2006) and Mondorff (2009a) for discussion of economy as a motivation of compound spelling. An economy of processing effort, or processing efficiency, is a notion that linguistic units that are less informative (e.g., due to their higher frequency or predictability) are less crucial for the communication success and are likely to undergo the reduction of the phonological or orthographic form (for an early proposal see Zipf, 1929). A reduction in the form of frequent units has been argued to result from (a) a higher resting activation level of such units in the long-term lexical memory of the language producer; (b) the ease of the oral or written production of units that is routinized through practice; (c) the rational strategy of increasing the utility of communication; and a number of other proposals and their combinations (for an extensive recent overview of efficiency research, see Jaeger & Tily, 2011). For the case of compounds, the economy account predicts that lexicalization comes with a change from a more complex (longer: spaced or hyphenated) format to easier-to-produce concatenated compounds. As typing a more frequently occurring symbol (and physically striking a larger key) of space is arguably less effortful than typing a hyphen, hyphenated compounds would tend to become spaced when becoming more frequent, on the economy account.

In our data, compound frequency, attested across all spelling formats, proved to be the most powerful predictor for compound spelling. Statistical models (Tables 2 and 3) reveal that the more frequent a compound is, the more likely it is concatenated. If a relatively frequent compound is not in concatenated format, it is most likely to be spaced rather than hyphenated. Moreover, distributional data from Wikipedia and the New York Times corpus revealed no evidence that

hyphenated compounds become concatenated as a result of lexicalization. Alternation was observed either between spaced and concatenated or between spaced and hyphenated compounds: for some compounds the two alternations were simultaneous. Finally, the diachronic data from the New York Times corpus pointed to concatenation as the end point of lexicalization of spaced compounds: hyphenation did not play a mediating role either. These findings enable us to answer the question whether there is an orthographic route to lexicalization in the affirmative and whether this route goes via hyphenation in the negative. Most commonly (in 43% of alternating compounds in Wikipedia) spaced compounds become concatenated. Not only does this reduce the typing effort for more frequent, lexicalized compounds, it also reduces the effort on the comprehender's side, as printed compounds are typically recognized faster in concatenated than in spaced format (see e.g., Cherng, 2008; Ji, Gagné, & Spalding, 2011; Juhasz et al., 2005). This runs counter to iconicity-motivated suggestions of the three-stage evolution (Shie, 2002). The predictions of the economy account are not fully borne out either. It is true that we observed the predicted formal reduction in the spaced-concatenated alternation. Yet on the economy account, one expects to also see a massive reduction of hyphenated compounds into concatenated ones, contrary to the fact (only 1.5% of Wikipedia compounds demonstrated this alternation).

We also note the prevalence of compounds that are attested in all three formats (about 19% in the Wikipedia set). For these compounds, lexicalization appears to take a different route: from hyphenated (least frequent) to spaced to concatenated (most frequent), see Tables 2 and 3 for statistical support. The question is then why a compound that starts out as a hyphenated compound would evolve into a spaced one once it becomes more frequent. One possibility may be linked to the fact that hyphenated noun-noun compounds tend to function as adjectives modifying nouns, e.g., *client-side scripting*, *price-drop tv*. It is possible that writers preserve the hyphen for those adjectival uses, and remove the hyphen when the compound is to be used in its nominal sense (*The store organized an unprecedented price drop*.) thus providing orthographic cues to the compound's syntactic function (for a similar discussion see Bauer & Renouf, 2001 and references therein). The resulting spaced compounds may further undergo lexicalization and become concatenated. However, whether this account for the hyphenation-spacing-concatenation route is valid clearly requires future research.

Explanatory factors for compound spelling

Table 6 summarizes the factors that are at work when spellers make their choice between spelling formats when using a compound in a Wikipedia article. Table 6 shows that the orthographic choice

between meaning-equivalent compounds is co-determined by a conflation of orthographic, distributional and semantic properties of compound's constituents and their morphological families. We argue here that the effects point at the economy of effort, rather than iconicity, as an underlying motivation of choice, where both the writer's and the reader's effort are factors of influence.

INSERT TABLE 6 ABOUT HERE

Morphemic salience

Laudanna and Burani (1995) introduced the notion of affixal salience, the likelihood of a morpheme to be recognized as a processing unit in its own right, even when embedded in a derived word. This likelihood is reportedly higher for morphemes that are longer, more frequent or productive (e.g., Bertram & Hyönä, 2003; Kuperman, Bertram, & Baayen, 2010; Laudanna & Burani, 1995). Expanding this notion to compounds, the effects of constituent length, frequency or family size can be argued to reflect morphemic salience of constituents, or their likelihood of perceptually "breaking out" from the written compound and being processed as independent units. In what follows we argue for the relevance of morphemic salience in production behavior.

Orthography: Longer constituents tend to be separated in writing, and when the choice is between a hyphen and a space, the preference is for a more definite visual cue, i.e. the space. The choices are fully in line with the well-documented behavior that *readers* exhibit when encountering compounds. Cross-linguistic studies on the visual comprehension robustly demonstrate that longer compounds (i.e. compounds with longer constituents) are more likely to be processed via morphemes than via full-form lexical representations (Bertram & Hyönä, 2003; but see Juhasz, 2008, for evidence that short English compounds may also be processed via morphemes) and that readers use a variety of segmentation cues that facilitate segmentation of the compound into morphemes (Bertram, Kuperman, Baayen, & Hyönä, 2011; Bertram, Pollatsek, & Hyönä, 2004). As known from eye-tracking studies, a longer left constituent also implies that the first fixation on the compound is farther away from the constituent boundary and thus the need for segmentation cues is the stronger (Bertram et al., 2004). When inserted into a compound with a conventional concatenated spelling, a space or a hyphen serve as strong segmentation cues and have been repeatedly shown to facilitate initial stages of compound processing (Bertram et al., 2011; Inhoff et al., 2000; Juhasz et al., 2005). In compounds with longer constituents, the preference for spacing as a stronger segmentation cue (as compared to either concatenation or hyphenation) may be construed

then as the reader-oriented strategy of introducing segmentation cues to the words the comprehension of which is cognitively demanding.

Constituent frequency: It has been repeatedly reported that constituents are involved in recognition of compounds, as evidenced by the effects of constituent frequency on behavioral latencies. Moreover, recent studies observed interactions revealing that morphological constituents with a higher frequency of occurrence (or a larger family size) may actually attenuate the effect of compound or derived word frequency (Baayen, Wurm, & Aycock, 2007; Balling & Baayen, 2008; Kuperman, Bertram, & Baayen, 2008; Kuperman et al., 2010; Kuperman, Schreuder, Bertram, & Baayen, 2009).

We propose that it is the salience of constituents in visual comprehension that underlies production choices. Consider also that the effects of constituent length and frequency on the likelihood of spacing show the same direction, while they are typically counter-directed in either production or comprehension latencies (negative for frequency, positive for length). We interpret this as evidence for the mediating role of morphemic salience to which both the constituent lengths and frequencies contribute. Constituents that are more likely targets for recognition even in a concatenated format due to their length or higher-frequency, are more likely to break out from compounds formally, and to be spelled with a space. Avoidance of concatenation may then reflect the response of the production system to the difficult comprehension of concatenated compounds with salient constituents: in other words, it is motivated by the economy of the reader's effort.

Analogy

A choice of any of the three formats for a compound was more likely if the format was also more common in either the left or the right morphological family of that compound. The influence of a morphological paradigm on the choice of alternants in a given family member is the hallmark of paradigmatic analogy (cf. Blevins, 2006) and has been empirically confirmed not only in English compound spelling (Sepp, 2006), but also in, among others, the choice of the linking element in Dutch and German compounds (Krott, Baayen, & Schreuder, 2001; Krott, Schreuder, Baayen, & Dressler 2007), and the choice of the past-tense markers in Dutch and English verb forms (Keuleers, 2008).

While the predictive role of analogy is not easy to relate to the principle of iconicity, it falls out immediately from the economy of effort. Studies in speech production, written production (typing),

and auditory and visual comprehension of compound words demonstrate that, across tasks and modalities, constituent families are activated during online compound processing (Bien, Levelt, & Baayen, 2005; Sahel et al., 2008). Producing a compound that is more predictable due to its consistency with the preferences of analogous compounds, conceivably reduces the production effort (for the effort of recognizing complex words diverging from their paradigms, see Milin et al., 2009b, and Milin, Filipovic Durdevic, & Moscoso del Prado Martin, 2009a). Likewise, maintaining consistency of a chosen compound's format with compounds that share constituents gives readers an effort-saving opportunity to fine-tune their processor for either a more difficult segmentation (as is the case with concatenated compounds) or a more difficult semantic integration (observed in visual comprehension of spaced or hyphenated compounds).

Semantic effects

The effect of semantic similarity between constituents on spelling preferences is a direct test of predictions that iconicity of independence makes: more similar (less distant) meanings of constituents are expected to come with less distant, i.e., concatenated, formal expressions. In fact, the opposite held true for the concatenated-spaced alternation: compounds with relatively similar constituent meanings were likely to be spaced. We argue again that while unintuitive given the producer's demands of the least effort, the tendency for spacing in transparent compounds can be explained by the economy of the reader's effort. The critical finding for this argument is Frisson et al.'s (2008) observation that semantic transparency (the degree of similarity between a constituent's and a compound's meaning) affected eye-movements in visual recognition of compounds when they were spaced. Transparent compounds were processed faster (Experiment 2), but no reliable effect of semantic similarity was observed in concatenated compounds (Experiment 1; but see Juhasz, 2007 for counterevidence). Similar effects were observed in Ji et al. (2011) in visual lexical decision. No difference was observed between transparent and opaque compounds in concatenated format, but clear processing advantages were found for transparent compounds relative to opaque compounds in spaced format (*moon light* eliciting shorter latencies than *honey moon*). Given that semantic integration matters most in the processing of spaced compounds, it is efficient to preserve spacing for relatively easy-to-comprehend transparent compounds. This is consistent with the preference for spacing that we observed in compounds with relatively transparent compounds (those with high LSA scores) Conversely, opaque compounds that could give rise to difficult recognition were more frequently found in concatenation, the format that semantics is reported to influences less. This testifies that the spelling system is statistically attuned to reducing the effort of the reader. Implementing the preferences that the iconicity principle suggests – a likely spaced format for

conceptually distant, opaque compounds, and a likely concatenation for transparent compounds – would lead to a suboptimal system in which semantic processing would take place when it is most detrimental, i.e., in compounds whose semantics makes them difficult for comprehension.

Conclusions

The present data demonstrate that spelling preferences in English compounds are codetermined by the morphemic salience of constituents in compounds, the analogical influence of the constituents' paradigms, as well as by the semantic relationship between constituents. Thus, the bias towards one of the three spelling formats is a weighted sum of correlations with orthographic, statistical and semantic properties of the compounds' constituents. In line with evidence obtained for syntactic choices reviewed above, we thus confirm that the word-level alternation between equivalent meanings and variable orthographic forms is neither random nor arbitrary (Mondorf, 2009a, 2009b; Rakic, 2009; Sepp, 2006). Moreover, we argue that preferences in the production system mirror the demands for the least effort of both production and comprehension (Gennari & MacDonald, 2009). This also becomes evident in the attested diachronic change from a longer and thus more effortful spaced representation to concatenation that accompanies lexicalization of compounds. Likewise, economy appears to underlie the fact that some spelling formats are preferred over others not because they are iconic with respect to the compound's meaning, or necessarily effort-saving for the writer. They are preferred because they fit best the strategies -- including morphemic segmentation and semantic integration -- that readers employ when processing written complex words. Importantly, going against these orthographic preferences in production comes with a high cost in recognition, as follows from the visual lexical decision and word naming data. The processing disadvantage in recognizing a concatenated compound that is more likely to occur in another spelling is on par with the effects of strongest known predictors of lexical decision latencies, i.e. word frequency and word length. This body of findings is compatible with the view that language structure is shaped through language use. While multiple competing motivations were proposed to drive the use-driven change in language structure, the present study is a demonstration that economy is apparently behind the fact that the written production of English compounds is attuned to their comprehension.

Acknowledgments

Thanks are due to Valentin Spitkovsky for extensive and insightful discussions of computational aspects of this work, and to Barbara Juhasz and Emmanuel Keuleers for their comments on an earlier draft.

FOOTNOTES

1. Throughout the paper, we will use the plus sign in the compound spelling (e.g., girl+friend) to refer to the compound regardless of how it is spelled.
2. While this procedure meets our goal of describing orthographic alternation, it cannot be used to identify non-alternating concatenated compounds. Unlike spaced and hyphenated compounds, concatenated compounds cannot be told apart from morphologically simple singular common nouns on the basis of part-of-speech tags or orthography in parsed Wikipedia, and the size of the corpus prohibits manual identification of concatenated compounds. To roughly estimate the number of non-alternating concatenated compounds, we extracted all concatenated compounds identified in the morphological coding of the lexical database CELEX (Baayen, Piepenbrock, & Gulikers, 1995). We found 500 compounds that were not part of our alternating set in Wikipedia. The sum of the alternating concatenated compounds and the non-alternating ones is 2,100 (= 1,600 + 500) and is a more precise – though potentially still too low -- estimate of the total type count of (alternating or non-alternating) concatenated compounds in the Wikipedia corpus.
3. The list of alternating compounds comprises: *audio+book, call+center, care+giver, chick+pea, coffee+house, copy+cat, cyber+cafe, data+base, die+cast, field+house, fund+raiser, help+line, house+cat, jump+start, race+day, road+show, salary+cap, shoe+box, show+biz, slide+show, soap+box, sound+board, steak+house, paint+ball*.
4. We are indebted to Emmanuel Keuleers for raising this possibility.
5. We also considered entropy as the measure of uncertainty in the choice of one of several alternatives. Entropy is minimal (zero) when a compound occurs in only one of the available formats, and it is maximal when all alternatives are equiprobable (see Milin et al., 2009b). Entropy of orthographic choice was not a significant predictor (at the 0.05-level) of either lexical decision or word naming response times.

References

Andrews, S., Miller, B., & Rayner, K. (2004). Eye movements and morphological segmentation of compound words: There is a mouse in mousetrap. *European Journal of Cognitive Psychology, 16,* 285-311.

Baayen, R. H. (2008). *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge University Press.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). The CELEX lexical database [CD-ROM]. Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

Baayen, R. H., Wurm, H. L., & Aycock, J. (2007). Lexical dynamics for low-frequency complex words. A regression study across tasks and modalities. *The Mental Lexicon, 2*, 419-463.

Balling, L., & Baayen, R.H. (2008) Morphological effects in auditory wordrecognition: Evidence from Danish. *Language and Cognitive Processes*, *23*,1159-1190.

Balota, D.A., Yap, M.J., Cortese, M.J., Hutchison, K.I., Kessler, B., Loftis, B., Neely, J.H., Nelson, D.L., Simpson, G.B., & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods, 39*, 445-459.

Bauer, L. (1988). *Introducing Linguistic Morphology*. Edinburgh: Edinburgh University Press.

Bauer, L., & Renouf, A. (2001). A Corpus-Based Study of Compounding in English. *Journal of English Linguistics*, 29, 101-123.

Bertram, R., & Hyönä, J. (2003). The length of a complex word modifies the role of morphological structure: Evidence from eye movements when reading short and long Finnish compounds. *Journal of Memory and Language, 48,* 615-634.

Bertram, R., Kuperman, V., Hyönä, J., & Baayen, R. H. (2011). The hyphen as a segmentation cue: It's getting better all the time. *Scandinavian Journal of Psychology*, *52*, 530-544.

Bertram, R., Pollatsek, A., & Hyönä, J. (2004). Morphological parsing and the use of segmentation cues in reading Finnish compounds. *Journal of Memory and Language, 51,* 325-345.

Bien, H., Levelt, W. M. J., & Baayen, R. H. (2005) Frequency effects in compound production. *Proceedings of the National Academy of Sciences, 102*, 17876-17881.

Blevins, J.P. (2006). *Word-based morphology*. Cambridge: Cambridge University Press.

Borjars, K., & Burrige, K. (2001). *Introducing English grammar*. London: Arnold.

Bresnan, J., Cueni, A., Nikitina, T., & Baayen, R. H. (2007). Predicting the dative alternation. In G. Boume, I. Kraemer, & J. Zwarts (Eds.), *Cognitive foundations of interpretation*, pp. 69-94. Royal Netherlands Academy of Science: Amsterdam.

Bresnan, J., & Ford, M. (2010). Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language, 86*, 168-213.

Brinton, L.J., & Traugott, E.C. (2005). *Lexicalization and language change*. Cambridge: Cambridge University Press.

Bybee, J. L. (2003). Mechanisms of change in grammaticalization: the role of frequency. In J. Brian, & R. Janda (Eds.). *Handbook of historical linguistics*, pp. 602-623. Oxford: Blackwell.

Bybee, J. L., & Hopper, P. (2001). *Frequency and the emergence of linguistic structure*. Amsterdam: John Benjamins.

Bybee, J. L., & Scheibman, J. (1999). The effect of usage on degrees of consituency: the reduction of don't in English. *Linguistics*, *37*, 575–596.

Cherng, M. (2008). *The Role of Hyphenation in English Compound Word Processing*. Unpublished bachelor thesis, Wesleyan college: Middletown, CT.

Croft, William. 2003. *Typology and universals*, 2nd edition. Cambridge: Cambridge Universitiy Press.

Croft, W. (2008). On iconicity of distance. *Cognitive Linguistics*, *19*, 49-57.

Crystal, D. (2001). *Language and the Internet*. Cambridge: Cambridge University Press.

Cunnings, I., & Clahsen, H. (2007). The Time-Course of Morphological Constraints: Evidence from Eye-Movements During Reading. *Cognition, 104,* 467-494.

Fabb, N. (1998). Compounding. In A. Spencer, & A. M. Zwicky, *The Handbook of Morphology*, pp. 66-83. Oxford: Blackwell.

Ferreira, V. S. (2003). The persistence of optional complementizer mention: Why saying a ''that" is not saying ''that" at all. *Journal of Memory and Language, 48*, 379–398.

Frank, A., & Jaeger, T.F. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proceedings of the 30th annual meeting of the Cognitive Science Society*, pp. 939-944. Washington, DC.

Gagné, C. L., & Spalding, T. L. (2009). Constituent integration during the processing of compound words: Does it involve the use of relational structures? *Journal of Memory and Language*, *60*, 20-35.

Gennari, S. P., & MacDonald, M. C. (2009). Linking production and comprehension processes: The case of relative clauses. *Cognition, 111*, 1-23.

Gries, S. (2003). *Multifactorial analysis in corpus linguistics: A study of particle placement*. New York: Continuum International Publishing Group Ltd..

Haiman, J. (1983). Iconic and Economic Motivation. *Language*, 59, 781-819.

Haiman, J. 2008. In defense of iconicity. *Cognitive Linguistics, 19,* 59-66.

Haspelmath, M. 2008. Frequency vs. iconicity in explaining grammatical asymmetries. *Cognitive Linguistics, 19,* 1-34.

Hawkins, J. A. (2004). *Efficiency and Complexity in Grammars*. Oxford: Oxford University Press.

Huddleston, R. (1984). *Introduction to the grammar of English*. Cambridge: Cambridge University Press.

Inhoff, A.W., & Radach, R. (2002). The biology of reading: The use of spatial information in the reading of complex words. *Comments on Modern Biology. Part C., Comments on Theoretical Biology, 7*, 121-138.

Inhoff, A. W., Radach, R., & Heller, D. (2000). Complex compounds in German: Interword spaces facilitate segmentation but hinder assignment of meaning. *Journal of Memory and Language, 42,* 23–50.

Inhoff, A.W., Starr, M.S., Solomon, M., & Placke, L. (2008). Eye movements during the reading of compound words and the influence of lexeme meaning. *Memory & Cognition 36,* 675-687.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59,* 434–446.

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology, 61,* 23-62.

Jaeger, T. F., & Tily, H. (2011). On language 'utility': Processing complexity and communicative efficiency'. *WIRE: Cognitive Science, 2(3)*, 323-335.

Jespersen, O. (1977). *Essentials of English Grammar*. London: George Allen and Unwin.

Ji, H., Gagné, C. L., & Spalding, T. L. (2011). Benefits and costs of lexical decomposition and semantic integration during the processing of transparent and opaque English compounds. *Journal of Memory and Language, 65,* 406-430.

Juhasz, B. J. (2007). The influence of semantic transparency on eye movements during English compound word recognition. In R. P. G. Van Gompel, M. H. Fischer, W. S. Murray, & R. L. Hill (Eds.), *Eye movements: A window on mind and brain*, pp. 373–390. Oxford, UK: Elsevier.

Juhasz, B.J. (2008). The processing of compound words in English: Effects of word length on eye movements during reading. *Language and Cognitive Processes, 23,*1057-1088.

Juhasz, B. J., Inhoff, A. W., & Rayner, K. (2005). The role of interword spaces in the processing of English compound words. *Language and Cognitive Processes, 20,* 291–316.

Keuleers (2008). *Memory-based learning of inflectional morphology*. Antwerp: University of Antwerp.

Klein, D. & Manning, C. D. (2003). Accurate Unlexicalized Parsing. *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pp. 423-430.

Krott, A., Baayen, R. H., & Schreuder, R. (2001). Analogy in morphology: modeling the choice of linking morphemes in Dutch. *Linguistics*, *39*, 51-93.

Krott, A., Schreuder, R., & Baayen, R.H., & Dressler, W.U. (2007). Analogical effects on linking elements in German compounds. *Language and Cognitive Processes*, *22*, 25-57.

Kuperman, V., Bertram, R., & Baayen, R. H. (2008). Morphological dynamics in compound processing. *Language and Cognitive Processes, 23*, 1089-1132.

Kuperman, V., Bertram, R., & Baayen, R. H. (2010). Processing trade-offs in the reading of Dutch derived words. *Journal of Memory and Language*, *62*, 83-97.

Kuperman, V., Schreuder, R., Bertram, R., & Baayen, R. H. (2009). Reading Poly-morphemic Dutch Compounds: Toward a Multiple Route Model of Lexical Processing. *Journal of Experimental Psychology: Human Perception and Performance, 35*, 876–895.

Landauer, T.K., & Dumais, S. (1997). A Solution to Plato's Problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, *104*, 211-240.

Landauer, T.K., Foltz, P., & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes*, *25*, 259-284.

Laudanna, A., & Burani, C. (1995). Distributional properties of derivational affixes: Implications for processing. In L. B. Feldman (Ed.), *Morphological aspects of language processing*, pp. 345–364. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lieber, R. (1983). Argument linking and compounds in English. *Linguistic Inquiry, 14*, 251-285

Lohmann, A., (2011). *Help* vs *help to*: a multifactorial, mixed-effects account of infinitive marker omission. *English Language and Linguistics*, *15*, 499-521.

Lohse, B., Hawkins, J., & Wasow, T. (2004). Domain minimization in English verb-particle constructions. *Language*, *80*, 238-261.

Milin, P., Filipovic Durdevic, D., & Moscoso del Prado Martin, F. (2009a). The simultaneous effects of inflectional paradigms and classes on lexicalrecognition: Evidence from Serbian. *Journal of Memory and Language, 60*, 50-64.

Milin, P., Kuperman, V., Kostic, A., & Baayen, R. H. (2009b). Words and paradigms bit by bit: An information-theoretic approach to the processing of inflection and derivation. In J. Blevins & J. Blevins (Eds.), *Analogy in grammar: Form and acquisition,* pp. 214-252. Oxford: Oxford University Press.

Moscoso del Prado, F., Deutsch, A., Frost, R., de Jong, N. H., Schreuder, R., and Baayen, R. H. (2005). Changing places: a cross-language perspective on frequency and family size in Hebrew and Dutch. *Journal of Memory and Language, 53*, 496-512

Mondorf, B. (2009a). How lexicalization reflected in hyphenation affects variation and word-formation. In: A. Dufter, J. Fleischer, & G. Seiler. (Eds.), *Describing and modeling variation in grammar*, pp. 361-388. Berlin: Mouton De Gruyter.

Mondorf, B. (2009b). *More Support for More-support: The Role of Processing Constraints on the Choice Between Synthetic and Analytic Comparative Forms*. Amsterdam: John Benjamins.

Newmeyer, F.J. (1992). Iconicity and Generative Grammar. *Language*, *68*, 756-796.

Plag, I, Kunter, G. & Lappe, S. (2007). Testing hypotheses about compound stress assignment in English: a corpus-based investigation, *Corpus Linguistics and Linguistic Theory, 3*, 199-233.

Quirk, R. Greenbaum, S., Leech, G. & Svartvik, J. (1985). *A comprehensive grammar of the English language*. London: Longman.

Rakic, S. (2009). Some Observations on the Structure, Type Frequencies and Spelling of English Compounds. *SKASE Journal of Theoretical Linguistics* [online]. 2009, vol. 6, no. 1. Available http://www.skase.sk/Volumes/JTL13/pdf_doc/04.pdf.

Rastle, K., Davis, M. H., & New, B. (2004). The broth in my brother's brothel: Morpho-orthographic segmentation in visual word recognition. *Psychonomic Bulletin & Review*, *11*, 1090-1098

Roland, D., Elman, J., & Ferreira, V. S. (2006). Why is that? Structural prediction and ambiguity resolution in a very large corpus of English sentences. *Cognition, 98*, 245-272.

Sahel, S., Nottbusch, G., Grimm, A. & Weingarten, R. (2008). Written production of German compounds: Effects of lexical frequency and semantic transparency. *Written Language and Literacy, 11,* 221-228.

Sepp, M. (2006). *Phonological constraints and free variation in compounding: A corpus study of English and Estonian noun compounds*. Unpublished doctoral dissertation, City University of New York, NY.

Shie, J. S. (2002). English hyphenated compounds. *Journal of the Da-Jeh University*, *11*, 89-98.

United States Government Printing Office. (2008). *Style Manual. 30th edition*. Washington, DC.

Wasow, T. (2002). *Postverbal Behavior*. Stanford: CSLI.

Zipf, G. K. (1929). Relative frequency as a determinant of phonetic change. *Harvard Studies in Classical Philology*, *15*, 1-95.

Zipf, G. K. (1935). *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. Boston: Houghton Miffin.

Table 1: Summary of the distribution of spelling variants across 2306 types of alternating compounds in Wikipedia: C stands for concatenated, S for spaced, H for hyphenated. The table reports type counts for all kinds of two-way alternations, and for the three-way alternation.

Alternation	Type count
C and S (H = 0)	984
S and H (C = 0)	856
C and H (S = 0)	34
C and S and H	432

Table 2: Summary of the logistic regression model for the spaced/concatenated alternation, fitted to 984 compounds, including only those predictors whose removal significantly decreased model's performance. Standard deviation of compound as a random effect was 1.60. Positive coefficients indicate the bias towards concatenated spelling, negative ones towards the spaced alternant.

	Regression coeffcient	Standard Error	z value	Pr(>|z|)
Intercept	0.48	0.05	9.26	0.00
Orthographic				
Left length	-0.25	0.05	-4.68	0.00
Distributional				
JointFreq	0.72	0.05	13.54	0.00
LeftFreq	-0.15	0.05	-2.75	0.01
RightFreq	-0.11	0.05	-2.02	0.04
BiasLeftFamFreqS	-0.16	0.05	-2.97	0.00
BiasRightFamFreqS	-0.40	0.05	-7.55	0.00

Table 3: Summary of the logistic regression model for spaced/hyphenated alternation, fitted to 856 compounds, including only those predictors whose removal significantly decreased the model's performance. Standard deviation of compound as a random effect was 1.13. Positive coefficients indicate the bias towards hyphenated spelling, negative ones towards the spaced alternant.

	Regression coeffcient	Standard Error	z value	Pr(>|z|)
Intercept	-1.73	0.04	-43.55	0.00
Orthographic				
Left length	-0.07	0.04	-1.77	0.08
Right length	-0.06	0.04	-1.61	0.11
Distributional				
JointFreq	-1.19	0.04	-29.64	0.00
BiasLeftFamFreqH	0.29	0.04	7.23	0.00
BiasRightFamFreqH	0.12	0.04	2.88	0.00

Table 4: Summary of the model for lexical decision times in the English Lexicon Project, fitted to 495 concatenated compounds (subset after trimming) which also occur in other spelling variants. Only those predictors are reported whose removal significantly decreased the model's performance.

	Regression coefficient	Standard Error	t value	Pr(>|t|)
Intercept	784.60	49.25	15.93	0.00
rBiasC	-99.21	19.14	-5.18	0.00
lJointFreq	-17.43	3.93	-4.43	0.00
lLeftFamFreq	-7.97	3.48	-2.29	0.03
lRightFamFreq	-11.53	3.57	-3.23	0.00
LeftLength	23.69	5.16	4.59	0.00
RightLength	23.06	5.45	4.23	0.00

Table 5: Summary of the model for word naming times in the English Lexicon Project, fitted to 483 concatenated compounds (subset after trimming) which also occur in other spelling variants. Only those predictors are reported whose removal significantly decreased the model's performance.

	Regression coefficient	Standard Error	t value	Pr(>|t|)
Intercept	704.46	28.34	24.86	0.00
rBiasC	-44.65	12.78	-3.49	0.00
lJointFreq	-9.73	2.49	-3.92	0.00
lLeftFamFreq	-8.56	1.73	-4.96	0.00
lRightFamFreq	-5.52	1.80	-3.07	0.00
LeftLength	17.13	3.54	4.84	0.00
RightLength	18.39	3.50	5.25	0.00

Table 6: Factors that drive writer's orthographic choices in compound spelling in the English Wikipedia.

Likely Concatenated vs. Likely Spaced			Likely Hyphenated vs. Likely Spaced	
Orthographic			*Orthographic*	
	Longer left constituent			Longer left and right constituent
Distributional			*Distributional*	
Higher-frequency compound	Higher-frequency left and right constituent			Higher-frequency compound
Analogical			*Analogical*	
Family bias towards concatenation	Family bias towards spacing		Family bias towards hyphenation	Family bias towards spacing
Semantic			*Semantic*	
	Greater semantic similarity of constituents			

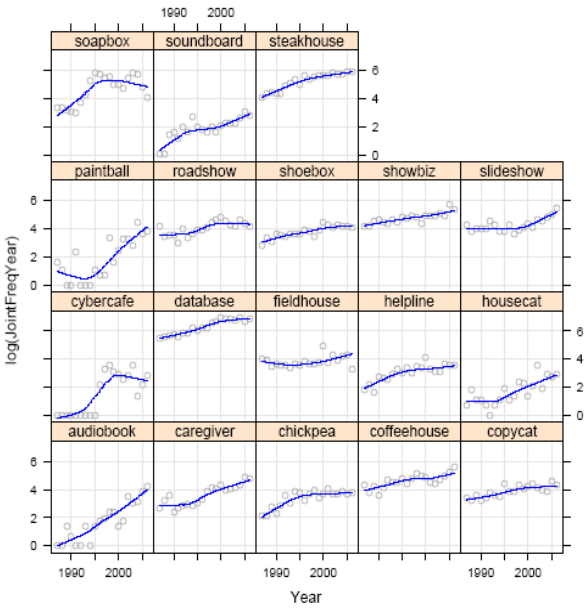
FIGURE CAPTIONS

Figure 1. Change in the frequency of use across all formats of 18 compounds in the NYT-corpus over 20 years (Panel A) and the bias (BiasU) towards concatenated/unspaced spelling for these compounds (Panel B).

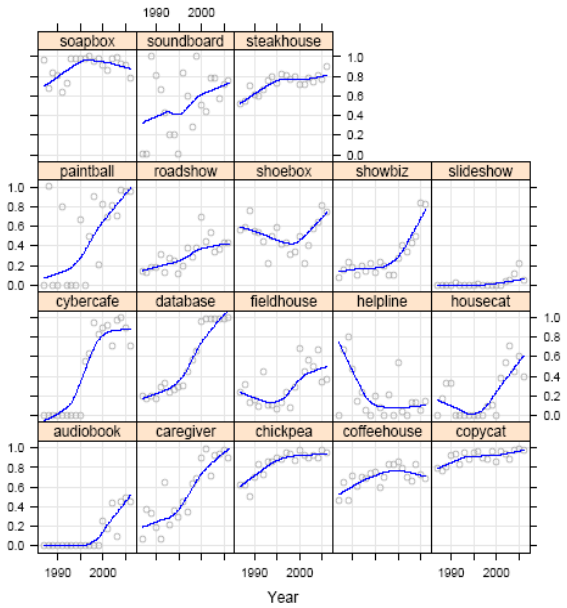
Figure 2. Partial effects on the mean lexical decision latencies to 495 English concatenated compounds. Dotted lines indicate boundaries of the 95% confidence interval.

Figure 3. Partial effects on the mean word naming latencies to 493 English concatenated compounds. Dotted lines indicate boundaries of the 95% confidence interval.

Figure 1



Panel A



Panel B

Figure 2

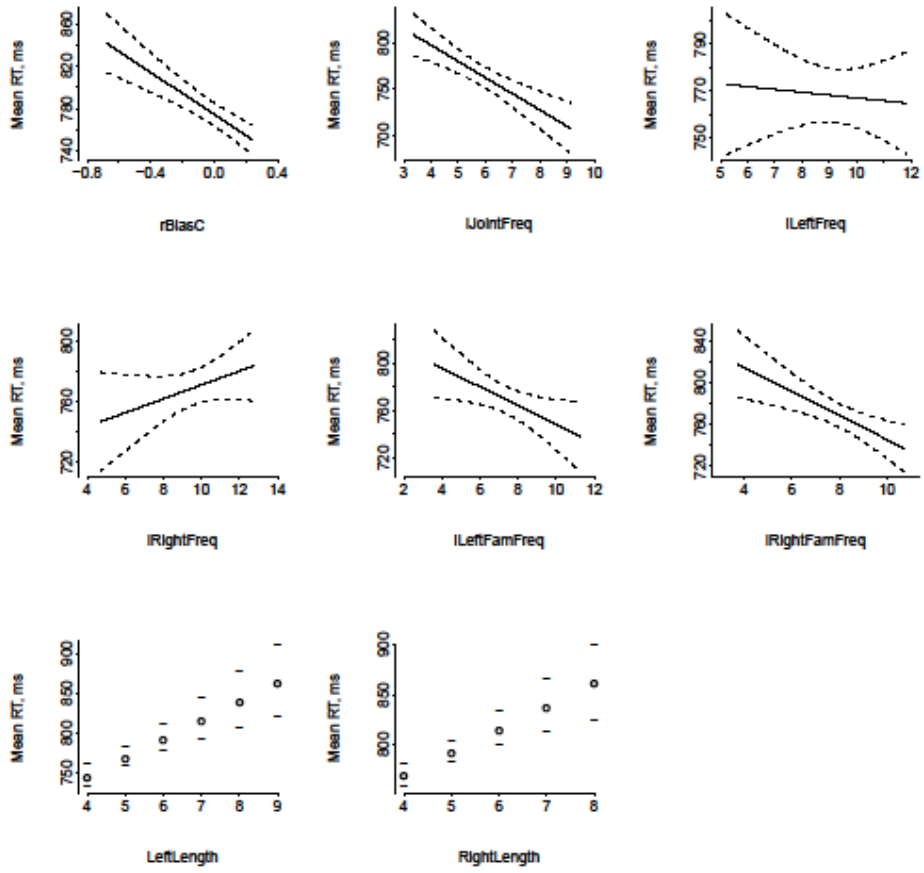
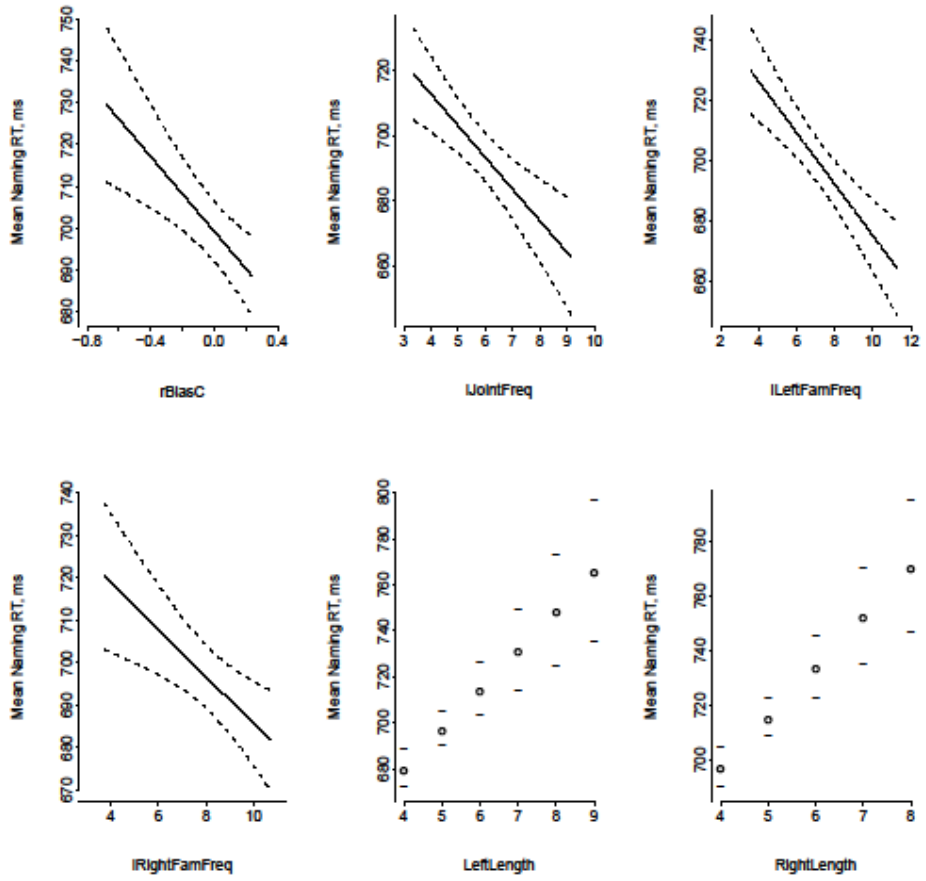

Mean RT, ms
rBiasC
lJointFreq
lLeftFreq
lRightFreq
lLeftFamFreq
lRightFamFreq
LeftLength
RightLength

Figure 3.


Mean Naming RT, ms
rBiasC
lJointFreq
lLeftFamFreq
lRightFamFreq
LeftLength
RightLength

Appendix 1. Computing the family bias

To illustrate how the family bias towards each of spelling variants is computed, consider the left constituent family of *bird* and a compound from that family *bird+cage* with 36 occurrences in the spaced format, 48 in the concatenated format, and 0 in the hyphenated format. The calculation was as follows:

1. Subtract the frequency of each spelling variant of the compound from the summed family frequency of those spelling variants (concatenated: 343 - 36 = 307; spaced: 247 - 48 = 199; hyphenated: 0 - 0 = 0).
2. Subtract the frequency of the compound summed across all spelling variants from the family frequency, i.e, the summed frequency of all family members across spelling variants: 590 - 84 = 506. In steps 1 and 2, subtracting the frequency of a compound from the family-based estimates of the bias is necessary to avoid circularity.
3. For each spelling variant of the compound, divide the value obtained in 1 by the result of 2: if the denominator is 0, assign 0 as the outcome. The family bias towards the concatenated format for bird+cage was 307/506 = 0.61; the bias towards the spaced format was 199/506 = 0.39; and the bias towards the hyphenated format was 0.

Importantly, many of our compounds were the only members of the respective constituent family, and thus would always have their family bias estimated at zero (see steps 2 and 3 above). Instead, we estimated their biases based on all compounds in the dataset, namely, as the ratio of the sum of family frequencies in the given spelling variants and the sum of family frequencies across all spelling variants. The resulting estimates of biases computed for a spelling variant for the left/right constituent family ranged from 0 to 1 and were labeled, for the left constituent family, BiasLeftFamC, BiasLeftFamS, and BiasLeftFamH, where "C, S, H" stand for concatenated, spaced and hyphenated.