English Language and Linguistics 12.3: 395–417. © Cambridge University Press 2008
doi:10.1017/S1360674308002694 Printed in the United Kingdom

The English comparative – language structure and language use[1]

MARTIN HILPERT
Freiburg Institute for Advanced Studies
(Received 20 September 2007; revised 19 December 2007)

Many English adjectives form the comparative in two ways, so that, for instance, *prouder* occurs alongside *more proud*. The availability of several forms raises the general questions of when and why speakers choose one variant over the other. The aim of this article is to identify factors of language structure and language use that underlie the comparative alternation and to determine their relative strengths on the basis of data from the BNC through a logistic regression analysis. The results suggest that the alternation is primarily governed by phonological factors, but that syntax and frequency of usage are of importance as well.

1 Introduction

This article investigates the formation of the English comparative. Many adjectives, as for example *proud*, can form the comparative in two ways, so that the morphological form *prouder* occurs alongside the periphrastic form *more proud*. To a lesser degree, these forms even combine in expressions such as *more prouder*. The availability of several ways to form the comparative raises the general questions of when and why speakers choose one variant over the other. Furthermore, it is worth asking why an adjective like *easy* has a clear preference for the morphological comparative, while a phonologically similar adjective like *queasy* has no such preference. The aim of this article is to identify factors that underlie the comparative alternation and to determine their relative strengths on the basis of data from the British National Corpus (Leech 1992) through a logistic regression analysis.

Previous treatments of the English comparative (Braun 1982; Quirk et al. 1985; Fries 1993; Leech & Culpeper 1997; Kytö & Romaine 1997; Lindquist 1998, 2000; Mondorf 2003, 2007; and Boyd 2007, amongst others) have identified many factors that appear to govern the comparative alternation to varying degrees. Basic to the alternation are phonological properties of the adjective in the positive form, such that for example monosyllabic adjectives show a greater tendency to form the morphological comparative than longer adjectives (Quirk et al. 1985: 461). Several characteristics of the final segment are also known to affect the alternation. For instance, adjectives with a final consonant cluster are, as a set, more likely to form the periphrastic comparative

[1] The author would like to thank the participants of the Workshop on Variation, Gradience and Frequency in Phonology (July 2007, Stanford, CA) for their feedback. Bas Aarts and two anonymous reviewers for *English Language and Linguistics* provided very helpful comments that improved the content and readability of this paper. As for remaining shortcomings, the usual disclaimers apply. Further thanks are due to Jeremy Boyd for sending me his dissertation. The present work was funded through the DAAD, which is hereby gratefully acknowledged.

https://doi.org/10.1017/S1360674308002694 Published online by Cambridge University Press

than adjectives that end in a vowel or a single consonant (Mondorf 2003). While the influence of phonology appears to be basic and immediate, syntactic, semantic, and pragmatic factors also play an active role. Mondorf (2003: 296) argues, for instance, that the semantic abstractness of the adjective, the complexity of the phrasal context, and discourse factors such as emphasis are determinants of the comparative alternation. The fact that the latter two of these factors are above the word level suggests that the choice between the periphrastic and the morphological comparative cannot be analyzed as a matter of morphophonology alone, a point that was already made by Braun (1982). All levels of linguistic structure have to be taken into account, and the relative importance of each level has to be determined empirically.

While linguistic structures account for some of the observed variation, factors of language use that are independent of structural characteristics also determine the comparative alternation. The fact that frequency of usage can influence the choice of a given linguistic variable is by now well known: highly frequent trisyllabic words such as *memory* and *family* are often produced as disyllabic in American English, whereas the phonologically similar but less frequent items *mammary* and *homily* tend to receive a trisyllabic pronunciation (Hooper 1976). Similarly, the irregular past-tense forms of verbs such as *weep* and *leap* are commonly regularized to *weeped* and *leaped*, whereas the corresponding forms of the more frequent verbs *keep* and *sleep* are retained as *kept* and *slept* (Bybee 1985). Recent work by Gahl & Garnsey (2004) further demonstrates that even syntactic patterns show frequency effects, as certain patterns invite phonetic reduction to a greater extent than others. Gahl & Garnsey show experimentally that verbal past-tense forms show a higher rate of final /t,d/ deletion if they are produced in the context of a complementation pattern that is typical for the respective verb. To illustrate, a reduced production of the past-tense form *confirmed* is more likely in (1), where the verb is followed by a direct object – its preferred complementation pattern. By contrast, the less typical context of a sentential complement in (2) makes it less likely that *confirmed* will undergo final /t,d/ deletion.

(1) The CIA director confirmed [the rumor]$_{DO}$ once it had spread widely.
(2) The CIA director confirmed [the rumor should have been stopped sooner.]$_{SC}$ (Gahl & Garnsey 2004: 762)

The present article thus embraces the study of usage (Barlow & Kemmer 2000; Bybee & Hopper 2001), as it holds the potential of explaining phenomena that could not be accounted for by invoking structural factors alone.

But how does the study of frequency patterns relate to the English comparative? With respect to the alternation between the two variants, it has been claimed that highly frequent adjectives exhibit a bias toward the morphological comparative, whereas less frequent adjectives tend to occur with the periphrastic variant (Braun 1982: 101; Quirk et al. 1985: 463). This would explain why for example *easy* forms the morphological comparative relatively more often than the less frequent element *noisy*, which in turn has a stronger bias toward the morphological comparative than *choosy*, which is still less frequent. As shown in table 1, a higher frequency of the adjective in its positive

https://doi.org/10.1017/S1360674308002694 Published online by Cambridge University Press

Table 1. *Correlation of frequency and morphological comparative formation*

Adjective	Positive frequency	Morphological comparatives
easy	14,760	99.2%
noisy	1,062	82.9%
choosy	49	40.0%

Table 2. *Comparative/positive ratio and morphological comparative formation*

Adjective	Comparative frequency	Positive frequency	Comp/pos ratio	Morphological comparatives
able	176	30,434	0.0058	2.8%
stable	94	3,420	0.0275	8.5%
humble	92	806	0.1141	72.8%

form correlates with a higher percentage of morphologically formed comparatives for the three adjectives in the BNC. Still, the general validity of this tendency needs to be tested systematically, preferably on an exhaustive set of alternating adjectives.

The frequency of an adjective in its positive form is only one frequency variable among several that merit consideration. One could also imagine that the ratio of positive and comparative forms functions as a determinant of the comparative alternation. For instance, the adjective *able* is not inherently gradable and does therefore not lend itself very well to the formation of the comparative. This shows itself in the fact that the positive forms of *able* by far outnumber the combined tokens of the comparative forms *abler* and *more able*. It seems motivated that a low comparative-to-positive ratio should correlate with a bias against the morphological comparative. Speakers should prefer to express unusual and novel comparisons through the variant with *more*, which is more explicit because of its additional phonological material. Again, this hypothesis can be tested against corpus data. Table 2 shows that the ratio of comparative and positive adjective forms in the BNC correlates with the percentage of morphologically formed comparatives for the adjectives *able*, *stable*, and *humble*.

As a comparison of *able* is unusual, it is typically expressed with the periphrastic comparative. By comparison, the adjective *stable*, which is relatively more often compared than *able*, shows a slightly stronger tendency to form the morphological comparative. The adjective *humble* forms the comparative still relatively more often, and of the three adjectives it shows the strongest affinity toward the morphological comparative. Again, for now this is merely anecdotal evidence that is in need of further, more systematic support.

To conclude this introductory section, the main aim of the present study is to show how the formation of the English comparative is governed by factors of language structure, i.e. phonological, morphological, and syntactic characteristics, and factors

https://doi.org/10.1017/S1360674308002694 Published online by Cambridge University Press

of language use, i.e. frequency patterns.[2] Of these two types, factors of language use have received a relatively sparse treatment in earlier studies. With the exception of Boyd (2007), all previous studies of the subject listed in the references of this article share the implicit assumption that the comparative alternation is to be explained mainly, if not exclusively, through structural factors. While frequencies from large corpora are presented in these studies, these frequencies are interpreted as the result of structural factors, not as potential explanatory factors. As will become apparent, language use, that is, statistical patterns in speakers' previous encounters with adjectives in the morphological and periphrastic comparative, has a role to play in the comparative alternation. With respect to language structure, previous research has succeeded in pointing to a variety of structural factors, but an integrated account of these factors is still missing. As many factors have not been tested as a set against a large body of data, it remains unclear which factors are more important than others, and whether some are the spurious by-product of others. In sum, the present study aims to take stock of what has been achieved so far, to test these findings against a large corpus, and to determine what matters most.

The remainder of this article is structured as follows. Section 2 introduces the factors that have been suggested in earlier research as explanatory variables for the comparative alternation. To the phonological, morphological, and syntactic variables that are mentioned in previous studies, the section adds two variables that pertain to language use. Section 3 offers a discussion of data and methodology, explaining first how the data are retrieved from the BNC and how they are annotated. After that, the section briefly details the workings of a logistic regression. Section 4 presents the results and tests whether the findings from the BNC carry over to another corpus that is different in genre and modality. Section 5 concludes and addresses the more general question of how the interplay of factors in the comparative alternation can be explained.

2 Factors underlying the comparative alternation

Previous analyses of the comparative alternation have established that the phenomenon in question is anything but trivial. Some determinants of the alternation appear to be absolute. Adjectives ending in *-ic*, *-ical*, and *-ous* are not conventionally used with the morphological comparative. Also, we do not find the periphrastic comparative where a suppletive lexical item has a blocking effect, such as *better* blocking **more good* (Poser 1992: 121).[3] Morphological comparatives are exceedingly rare with adjectives that have four or more syllables, but note the case of *untrustworthier*. By contrast, most structural factors, and certainly all of the frequency-related factors, only have gradient effects on the comparative alternation. It is therefore worthwhile to probe the relative

[2] Among factors of language use, one could also include factors of pragmatics and discourse, as, for instance, done by Mondorf (2003). For practical reasons, the discussion in the present article is restricted to factors that lend themselves to a quantitative corpus-linguistic analysis.

[3] See Boyd (2007) for a thorough critique of 'blocking' accounts.

https://doi.org/10.1017/S1360674308002694 Published online by Cambridge University Press

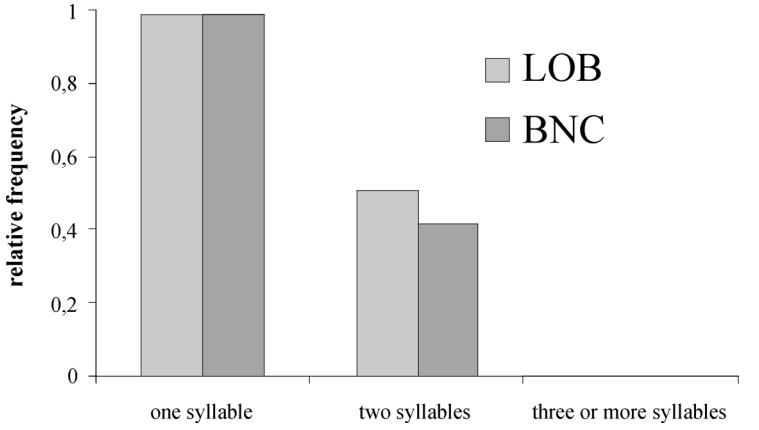


Figure 1. Relative frequency of morphological comparatives by number of syllables

strengths of these factors: a weighted analysis can resolve which factors matter most strongly to the alternation. The following sections discuss the variables that are used in this study and motivate why they should be included in the analysis.

2.1 Number of syllables

Most obviously perhaps, the number of syllables in an adjective is a determinant of the comparative alternation (Quirk et al. 1985). Leech & Culpeper (1997: 355) present evidence from the LOB (Johansson et al. 1978) and a section of the BNC, showing that most variance between the morphological and the periphrastic comparative falls into the domain of disyllabic adjectives. Figure 1 shows that in both corpora, monosyllabic and trisyllabic adjectives have a nearly uniform tendency to form only one variant (cf. table A1 in the Appendix for raw frequencies). The relative frequency of morphological comparatives with monosyllabic adjectives approaches 1, and, conversely, it approaches 0 with trisyllabic adjectives. Similar figures are reported by Kytö & Romaine (1997). The first variable to be included in the present analysis is thus the number of syllables that an adjective has in its positive form.

2.2 Characteristics of the final segment

It has been argued that several different types of final segments have an influence on the comparative alternation. Kytö & Romaine (1997) report that disyllabic adjectives ending in the sonorants /l/ and /r/ as well as disyllabic adjectives ending in /i/ show a tendency to form the morphological comparative. Later studies have shown that these generalizations need to be qualified. Data presented by Mondorf (2003: 281) show that a final /r/ is in fact more drawn toward the periphrastic variant. This is to be interpreted as a reluctance in rhotic varieties of English to concatenate two identical elements, as in *cleverer*. Lindquist (1998, 2000) further distinguishes between the ending

https://doi.org/10.1017/S1360674308002694 Published online by Cambridge University Press

/li/ as opposed to merely /i/, noting that adjectives such as *lively* show a comparatively greater affinity toward periphrastic comparative formation. Finally, adjectives ending in consonant clusters, as for example *correct* or *false*, are more likely to form the periphrastic comparative (Mondorf 2003: 283). Since final /l/ often occurs within the context of a final consonant cluster, as in *simple* or *able*, the role of /l/ in both contexts has to be reanalyzed as well.

To accommodate these distinctions, the present analysis includes a categorical variable that ranges over six different classes of final elements: /i/, /li/, /r/, /l/, a final consonant cluster, and the class of all other endings. Note that these are coded as mutually exclusive. An adjective like *subtle* ends in an /l/ and in a consonant cluster at the same time, but it only receives a positive value for the consonant cluster. Likewise, *lively* is categorized as ending in /li/, but not /i/.

2.3 Stress patterns

The factor of stress has been invoked in several accounts of the comparative alternation (Rohr 1929; Braun 1982; Leech & Culpeper 1997; *inter alia*). Leech & Culpeper (1997: 361) note that disyllabic adjectives with final stress commonly form the periphrastic comparative, pointing to examples such as *remote* and *severe*. The present analysis includes a variable that indicates whether an adjective has final stress or not. Monosyllabic adjectives automatically receive a positive value, whereas adjectives with two or more syllables can receive a negative value, depending on their stress pattern.

Mondorf (2003) further argues that stress has an influence on the phrase level. The morphological variant is said to be favored in contexts where the comparative suffix *-er* functions as a buffer between two stressed syllables. In attributive position, the comparative adjective may be followed by a noun with primary stress, as in *a frésher sálad*. Mondorf examines 23 monosyllabic adjectives, finding that these show a greater bias toward the morphological comparative when used attributively, rather than predicatively or in other syntactic contexts (2003: 275). With respect to Mondorf's argument, it needs to be added that attributive syntax is not the only context in which the morphological comparative may be used to avoid a stress clash. In an attributive construction like *It's cóoler nów than it was last week*, or in a construction such as *The drúnker Róbert gets, the more he likes Jessica*, the same phenomenon can be observed. The present study thus operationalizes Mondorf's observation through a variable that simply represents for each example in the database whether or not the word to the immediate right of the comparative adjective bears initial stress.[4] A right-side collocate with initial stress is expected to bias speakers toward the morphological comparative, while no such tendency should be observed with right-side collocates that have other stress patterns.

[4] This operationalization requires the exclusion of monosyllabic words that either cannot bear stress at all or that are only stressed in contexts of special emphasis. In the database, the right-collocates *a*, *an*, *and*, *at*, *by*, *for*, *in*, *if*, *it*, *of*, *on*, *or*, *than*, *the*, *to*, and *with* were coded as not bearing stress. Also, punctuation to the right of the comparative adjective was taken to preclude the possibility of a stress clash, since it commonly indicates the beginning of a new intonation contour.

https://doi.org/10.1017/S1360674308002694 Published online by Cambridge University Press

2.4 *Morphological complexity*

While the number of morphemes strongly correlates with word length in general, Mondorf (2003: 284) argues that morphological complexity is a factor that independently introduces a bias toward the periphrastic comparative. She compares disyllabic adjectives ending in the phoneme /l/, showing that morphologically complex items such as *careful* or *partial* have a stronger tendency to form the periphrastic comparative than the monomorphemic adjectives *gentle* or *humble*. Mondorf interprets this as support for the hypothesis that morphological complexity correlates with periphrastic comparative formation. An additional variable in the present study therefore records the number of morphemes for each adjective, in order to see if morphological complexity has an independent effect.

2.5 *Presence of a* to-*infinitive complement*

Mondorf (2003: 262) argues that a *to*-infinitive complement following the comparative adjective correlates with increased usage of the periphrastic variant. However, she notes that the effect does not hold uniformly across different classes of adjectives and different syntactic environments. She excludes attributive uses from her study, as the co-occurrence of attributive use and complementation is quite rare.[5] Mondorf studies 12 monosyllabic adjectives that robustly form the periphrastic variant more often if they are followed by a *to*-infinitive. By contrast, a set of 16 disyllabic adjectives does not yield a particular effect. Mondorf attributes this finding to the interaction of different factors. To measure the effect of infinitival complementation on the comparative alternation, this study includes a variable recording the presence or absence of such a complement for each example in the database.

2.6 *Attributive and predicative usage*

Leech & Culpeper (1997) study ten disyllabic adjectives that alternate fairly evenly between the morphological and the periphrastic comparative. Using data from the BNC, they show that attributive usage correlates with formation of the morphological comparative. Conversely, predicative usage is shown to correlate with formation of the periphrastic comparative (1997: 366). The present study will test these findings against a larger set of adjectives from the BNC, thereby also assessing the strength of the effect relative to other factors that affect the choice between the morphological and the periphrastic comparative. In order to do so, all attributive and predicative examples in the database are marked as such. Attributive examples are identified by means of a following noun, as in *a bigger boat*; predicative examples are identified through a preceding copula, as in *That is more realistic*. Since the choice of attributive and predicative usage is mutually exclusive, a single categorical variable is created

[5] This would refer to examples like *Let's have a dryer wine to go with the pork roast,* which are retained in the present analysis.

https://doi.org/10.1017/S1360674308002694 Published online by Cambridge University Press

in which all examples of the database are coded as either attributive, predicative, or other.

2.7 *A following* than

Both Leech & Culpeper (1997) and Lindquist (1998) consider the element *than* following the comparative adjective as a potential factor governing the alternation. Both studies express the initial hypothesis that a following *than* will correlate with increased usage of the periphrastic comparative. However, both studies present quantitative evidence suggesting that there is in fact no tangible effect (Leech & Culpeper 1997: 367; Lindquist 1998: 129). Should we thus simply disregard this variable? This study retains the variable for the reason that both earlier studies were based on limited sets of data and the intuitions of their authors might well be borne out by a study that takes more data into account. The presence or absence of a following *than* is therefore included as a variable.

2.8 *Premodification*

Both studies mentioned in the previous section find that premodification affects the choice between the morphological and the periphrastic comparative. Degree modifiers such as *a little*, *much*, or *marginally* tend to bias speakers toward using the periphrastic variant (Leech & Culpeper 1997: 367; Lindquist 1998: 127). It is therefore necessary to include the presence of a premodifier as a separate variable. Needless to say, premodification is instantiated by a set of forms that is difficult to define exhaustively. The present analysis includes the adverbs *altogether*, *considerably*, *even*, *ever*, *far*, *much*, *significantly*, *slightly*, *still*, *yet*, and the complex degree modifiers *a bit*, *a little*, and *a lot*, which can be assumed to account for most of the premodified examples. If any of these forms is found to the left of a comparative adjective, the example in question receives a positive value for the variable of premodification.

2.9 *Frequency values*

Two variables representing frequency values will be considered in the present analysis. The first of these concerns the frequency of the positive form. If an alternating adjective is used frequently in the positive, we expect speakers to be biased toward the morphological comparative (Quirk et al. 1985: 463). As mentioned in the introduction, the highly frequent adjective *easy* overwhelmingly forms the morphological comparative, while this is not so for the less frequent adjective *choosy*. Second, the ratio of comparative and positive forms is included as a variable. As explained above, a high ratio of comparative and positive forms means that an adjective is highly gradable. Adjectives such as *tall* or *long* fall into this category. By contrast, a low value indicates that an adjective encodes a property that is more absolute, as in *red* or *square*. Highly gradable adjectives are expected to be more likely to form the morphological comparative.

https://doi.org/10.1017/S1360674308002694 Published online by Cambridge University Press

Table 3. *Explanatory variables for the comparative alternation*

Variable	Source	Tendency
1. Number of syllables	Quirk et al. (1985)	periphrastic
2. The final element(s)		
Final /i/	Kytö & Romaine (1997)	morphological
Final /l/	Kytö & Romaine (1997)	morphological
Final /r/	Mondorf (2003)	periphrastic
Final /li/	Lindquist (1998)	periphrastic
Final consonant cluster	Mondorf (2003)	periphrastic
3. Final stress	Leech & Culpeper (1997)	periphrastic
4. Initial stress of right collocate	Mondorf (2003)	morphological
5. Number of morphemes	Mondorf (2003)	periphrastic
6. *to*-infinitive complementation	Mondorf (2003)	periphrastic
7. Attributive and predicative usage		
Attributive	Leech & Culpeper (1997)	morphological
Predicative	Leech & Culpeper (1997)	periphrastic
8. A following *than*	Lindquist (1998)	periphrastic
9. Premodification	Lindquist (1998)	periphrastic
10. Positive frequency	Quirk et al. (1985)	morphological
11. Comparative/positive ratio	Mondorf (2003)	morphological

2.10 Overview of the explanatory variables

Table 3 summarizes the explanatory variables that will be considered. Each variable is listed with a published source. The rightmost column in table 3 indicates the tendency that each variable is thought to have. For ordinal variables, this tendency correlates with a high value; e.g. a high number of syllables correlates with a high probability of the periphrastic comparative. For nominal variables, a positive value correlates with the indicated tendency.

3 Data and methodology

This section explains how the data for the present case study were gathered and analyzed. The primary source of data was the British National Corpus (Leech 1992); the chosen method of statistical analysis was a logistic regression analysis (Johnson 2007). Both data and methodology are explained in more detail below.

The BNC is a balanced corpus of 100 million words that contains both written and spoken data from a wide variety of genres. The corpus was compiled in the early 1990s, but it contains texts that were produced in earlier decades. The corpus is annotated for parts of speech, which made it possible to automate many of the searches necessary for the variables that were described above. Importantly, the corpus provides a tag for the morphological comparative, so that all inflected adjectives could be retrieved exhaustively. The periphrastic comparatives were searched for with a regular expression

https://doi.org/10.1017/S1360674308002694 Published online by Cambridge University Press

that specified the string *more* to the left of an adjective. This search pattern produces a number of false positives. In sentences of the type *John is more sad than angry*, the phrase *more sad* does not instantiate a periphrastic comparative, as the corresponding *John is sadder than angry* is not a valid paraphrase of the original sentence (di Sciullo & Williams 1987). Similarly, the sentence *We publish more fine books than ever before* does not mean that the quality of the books is getting better, but rather that the quantity of published books is increasing. Examples like these were manually identified and excluded.

The present study further restricts itself to those adjectives that occur with both of the two comparatives. In total, 247 alternating adjective types were identified in the BNC. Table A2 in the Appendix lists all alternating adjectives along with their frequencies in the morphological and the periphrastic comparative. The database includes many adjectives that could be thought of as occurring in only one variant. The sentences in (3) may serve to illustrate some of the more unusual examples from the BNC.

(3) (a) Some of these are *more easy* to exploit than others.
(b) The skewness has a positive or negative value when *more fine* or *more coarse* materials are present than in a normal distribution.
(c) the even *shadowier* lover who must have abandoned her
(d) These are *robuster* machines that record onto ¾ inch tape.

In terms of actual token frequencies, the retrieved adjective types yield 71,622 examples of morphological comparatives and 8,256 examples of periphrastic comparatives. This means that there is a clear asymmetry in frequency. With the alternating adjectives, morphological comparatives are about nine times as frequent as periphrastic comparatives. The database for the present study pairs each example with a category label that identifies it as either morphological or comparative and with a set of values for the variables summarized above. To illustrate, the example *That's far more likely now* receives the coding shown in (4).

(4) *That's far more likely now.*

Category	periphrastic
Number of syllables	2
Final element(s)	/li/
Final stress	no
Initial stress of right collocate	yes
Number of morphemes	1
to-infinitive complementation	no
Attributive or predicative usage	predicative
A following *than*	no
Premodification	yes
Positive frequency	23,667
Comparative/positive ratio	0.137

https://doi.org/10.1017/S1360674308002694 Published online by Cambridge University Press

Due to the size of the database, the coding procedures had to be operationalized in such a way that they could be performed automatically. Fortunately, many of the variables only pertain to the adjective type, that is, all instances of *likely* will end in /li/ and will share the characteristic that the frequency of their positive form in the BNC is 23,667. More challenging are variables that involve characteristics of individual examples. For instance, whether or not an example instantiates a predicative construction has to be determined on a case-by-case basis. Here, the tagging of the BNC allows the automatic coding of the syntactic variables. By contrast, a linguistic category that is not annotated in the BNC is stress. The question whether a particular example included a stress clash, as in *a more robúst finding*, hence required categorization of all right-side collocates, i.e. words to the immediate right of the comparative adjective, into words with initial stress and words with other stress patterns. An electronic version of *Webster's Unabridged Dictionary* (see references) was used to identify all right-side collocates with initial stress.

A brief note is called for to comment on how the frequencies of adjectives in their positive form were determined. It is easy enough to exhaustively retrieve each alternating adjective from the BNC, but to determine how many of these are positive forms, the respective token figures from the periphrastic comparative need to be subtracted. To illustrate, the form *profound* has a token frequency of 1,434 in the BNC. Since there are 106 instances of *more profound* in the BNC, we have to subtract this number to arrive at 1,328 as the actual frequency of the positive form.

The resulting matrix of all annotated examples was submitted to a logistic regression analysis in order to determine the significance and relative importance of each variable. The following paragraphs briefly explain how the analysis works.

A logistic regression analysis can be used to study how a choice between two mutually exclusive categories is affected by a variety of different factors. A common textbook example would be to analyze the factors that determine homeownership. Conveniently, the method can handle both explanatory variables that are *continuous*, such as age and number of children, and variables that are *categorical*, such as gender or marital status. Based on a dataset that holds information about these variables for a sample of homeowners and non-homeowners, it can be determined how each variable affects the probability of owning a home. Judging each case, the model determines the relative importance of the variables. In a second step, the derived model can be used to predict whether someone does own a home or not. To illustrate, the probability of homeownership will be substantially higher for a married 45-year-old female with three children than for an unmarried 25-year-old male without children.

How does this transfer to the analysis of language? One analogous linguistic example would be the alternation between the English ditransitive construction and the prepositional dative, both of which can express the transfer of a theme to a recipient:

(5) John gave Mary the book.
(6) John gave the book to Mary.

https://doi.org/10.1017/S1360674308002694 Published online by Cambridge University Press

Bresnan et al. (2005) use logistic regression to show that speakers' choices to use one variant rather than the other depends on several factors pertaining to both the recipient and the theme, such as their discourse accessibility, whether they are expressed pronominally or lexically, or whether they are definite or indefinite. Crucially, the relative strength of each factor can be quantified. For instance, the alternation is highly sensitive to the question whether the recipient is animate or inanimate. A transfer to an inanimate recipient, as in *I gave the clothes to the Red Cross*, has twelve times greater odds to be expressed with the prepositional dative than a transfer to an animate recipient (2005: 20). A comparatively weaker factor concerns the distinction between pronominal and lexical recipients. A transfer to a recipient that is expressed pronominally, as in *John gave her the book*, has three times greater odds to be expressed with the ditransitive than a transfer to a lexically encoded recipient (2005: 20).

In the case of the present study, use of logistic regression is appropriate for several reasons. First, the alternation between the two comparatives is a clear binary choice. With adjectives such as *proud*, speakers can either choose the morphological comparative *prouder* or the periphrastic variant *more proud*. Further, it has been shown time and again that the factors underlying the alternation are gradient, acting in a probabilistic rather than a categorical way. Third, given the amount of previous work, it would be desirable to know not only whether a given factor has a tangible effect, but also how it compares in strength to other factors.

4 Results

Table 4 summarizes the results of the analysis. For each variable, the table lists three numerical values. Each of these will be briefly explained here.

For ease of exposition, let us start with the rightmost column, which states p-values for each variable and thus tells us whether or not a given factor matters to the alternation. In the social sciences, a p-value below 0.05 is commonly taken to mean that a factor significantly contributes to an observed distribution. The column contains just one value that exceeds that limit, namely the variable of initial stress in the right collocate ($p = 0.234$). This finding contradicts Mondorf (2003: 275), who posits the avoidance of stress clash as a determinant of the comparative alternation. All other variables are judged to be significant, which allows the initial conclusion that previous analyses have detected many relevant factors. The table further shows that a following *than* does, after all, have an effect on the alternation ($p = 0.026$). Leech & Culpeper (1997) and Lindquist (1998) considered this factor, but found no evidence for its effect.

The second column in table 4 lists the coefficients.[6] Coefficients can take on positive and negative values, indicating the direction of the effect that the respective variable

[6] In table 5, the B after coefficients stands for Beta, whose numerical value represents the direction and magnitude of influence that a particular coefficient has.

https://doi.org/10.1017/S1360674308002694 Published online by Cambridge University Press

Table 4. *Coefficients of the comparative alternation*

Variable	Coefficients (B)	Odds ratio	P
(Intercept)	−13.56	0.00	<0.001
1. Number of syllables	6.33	563.41	<0.001
2. The final element(s)			
Final /i/	−1.26	0.29	<0.001
Final /l/	3.60	36.71	<0.001
Final /r/	1.82	6.18	<0.001
Final /li/	3.06	21.32	<0.001
Final consonant cluster	1.44	4.22	<0.001
3. Final stress	2.59	13.38	<0.001
4. Initial stress of right collocate	−0.06	0.94	0.234
5. Number of morphemes	0.20	1.22	0.032
6. *to*-infinitive complementation	1.28	3.60	<0.001
7. Attributive and predicative usage			
Attributive	−0.67	0.51	<0.001
Predicative	0.46	1.58	<0.001
8. A following *than*	−0.14	0.87	0.026
9. Premodification	0.19	1.21	<0.001
10. Positive frequency	−0.00002	0.99998	<0.001
11. Comparative/positive ratio	−7.18	0.00	<0.001

has on the alternation.[7] A negative value indicates a bias toward the morphological comparative, whereas, conversely, a positive value correlates with increased usage of the periphrastic comparative. As a mnemonic, read *minus – morphological* and *plus – periphrastic*. We can thus use this column to verify that the variables actually have the effect that they are proposed to have. To illustrate, the coefficient of final stress indicates the predicted effect of a bias toward the periphrastic comparative. Also, as predicted by Leech & Culpeper (1997), attributive syntax favors the morphological comparative, while predicative syntax biases speakers toward the periphrastic variant. Thirdly, the coefficient of final /l/ indicates a preference for the periphrastic comparative, which appears to contradict Kytö & Romaine (1997). However, their claim was based on adjectives like *simple* and *able*, which are coded as ending in a consonant cluster in the present analysis. Leech & Culpeper (1997) and Lindquist (1998) hypothesized that a following *than* would bias speakers toward the periphrastic comparative, but in fact the opposite seems to be the case.

[7] A brief note is called for to explain the row labeled *Intercept*. The intercept shows the odds of a morphological comparative relative to the odds of a periphrastic comparative. As the morphological comparative has a much higher text frequency than the periphrastic comparative, picking a random example from the database will likely yield a morphological comparative. Put simply, if you come across a comparative in the English language, chances are very high that it will be a morphological comparative. This asymmetry is responsible for the high odds ratio and the low p-value.

https://doi.org/10.1017/S1360674308002694 Published online by Cambridge University Press

Finally, table 4 offers odds ratios, which reflect the respective strengths of each factor. Odds ratios only take on positive values. They are interpreted as multiplicative: if the odds ratio is 1, the variable's odds of a certain outcome are multiplied by 1, i.e. they remain unchanged.[8] The more the value differs from 1, the stronger is the observed effect. For instance, the variable of *to*-infinitive complementation has an odds ratio of 3.60, indicating that the odds of examples with infinitival complements forming the periphrastic comparative are more than three times greater than the odds of examples without such complementation. To compare, the tendency for examples with *than* to form the morphological comparative rather than the periphrastic variant is not particularly strong, as the odds ratio of 0.87 is fairly close to 1.

With continuous variables, odds ratios are to be interpreted in a slightly different way, which prohibits easy comparisons to the odds ratios of categorical variables. Consider the number of syllables in table 4, which has a very high odds ratio of 563.4. This figure tells us that the odds of disyllabic adjectives forming a periphrastic comparative are about 560 times the odds of a monosyllabic adjective doing so. In turn, the odds of trisyllabic adjectives are about 560 times higher still. The high odds ratio thus mirrors the extreme rarity of morphological comparatives with adjectives that have three or more syllables.

Given that the odds ratio for the variable of positive frequency is 0.99998, and hence very close to 1, the attentive reader may wonder why this variable is found to be significant at all. The answer to that lies in the fact that odds ratios of continuous variables apply to unit changes in the variable. With regard to numbers of syllables, the attested units range from one syllable to three syllables. By contrast, the positive frequencies of the adjectives in the database range from 49 for *choosy* to 148,357 for *new*. The low odds ratio for the variable of positive frequency thus reflects that not much can be predicted from a difference of, say, 100 tokens, but that a noticeable difference obtains between adjectives that occur in altogether different orders of magnitude.

With these preliminaries in place, we can now inspect table 4 more closely. A reassuring result is that the number of syllables seems to have the most dramatic effect on the alternation. This finding corroborates our intuition that word length is the most basic factor. Morphological complexity, by comparison, has a much weaker effect. Since disyllabic and trisyllabic words have relatively similar chances of consisting of two morphemes, this variable misses a distinction that strongly influences the alternation.

Among the final elements, four of the five categories show a strong bias toward the periphrastic comparative. Final /l/ yields the strongest effect. Recall that this variable pertains only to adjectives that end in a single /l/, not a consonant cluster. The adjectives in question are thus *dull*, *frail*, *full*, *ill*, *pale*, *real*, *shrill*, *small*, *stale*, and *vile*. Despite

[8] Importantly, the odds of a certain outcome are not identical with its probability. In a coin flip, the probability of heads is 0.5. By contrast, the odds of heads are calculated as the probability of a success divided by the probability of failure, thus equaling $0.5/0.5 = 1$.

https://doi.org/10.1017/S1360674308002694 Published online by Cambridge University Press

the fact that these adjectives are monosyllabic, they show a strong tendency to occur in the periphrastic comparative (cf. table A2 in the Appendix). Final /li/ also leads to strongly increased formation of the periphrastic variant. The most frequent adjective in this category is *likely*, but also other elements such as *deadly*, *friendly*, or *worldly* show the same tendency. Final /r/ and final consonant clusters yield weaker effects. Final /i/ has an odds ratio of 0.29, which translates into the observation that adjectives such as *happy* and *noisy* have three times greater odds to form the morphological comparative than adjectives without a final /i/.

The present analysis confirms that final stress correlates with periphrastic comparative formation, as claimed by Leech & Culpeper (1997). Adjectives with final stress, such as *intense*, *mature*, and *remote* have twice the odds to form the periphrastic variant than adjectives with other stress patterns.

All four syntactic variables in the analysis are found to be significant. A following *to*-infinitive has the strongest effect of the four, lending strong credence to the observations of Mondorf (2003). The remaining syntactic factors only have limited impacts. Overall, in comparison to the phonological factors, syntax affects the comparative alternation in a relatively minor way. The analysis shows that syntax does matter, but the syntactic variables seem to be fine-tuning the model, rather than bringing about major changes.

Finally, the two frequency-based variables show the predicted effects. Increased corpus frequency and a higher ratio of comparative and positive forms both increase the odds of morphological comparative formation. As mentioned above, odds ratios of continuous variables are not easily compared to odds ratios of categorical variables. However, since it would be very informative to learn about the relative impact of frequency, we will return to this topic after addressing the accuracy of the analysis.

The accuracy of a logistic regression analysis is called the fit of the model. The fact that in the case of the present study most of the included variables are found to have significant effects does not, in and of itself, vouch for the overall quality of the analysis. It could well be that some important factor simply remains to be discovered. One way to assess the descriptive accuracy of the analysis is to determine the percentage of examples in the database for which the analysis makes a correct prediction. Note that in our case, a statistical model could achieve 89.7 percent accuracy merely by always predicting the morphological comparative, since that variant is so much more frequent. In order to earn its keep, the analysis needs to do substantially better than that. Table 5 illustrates how well the analysis predicts whether a given instance of a comparative in the BNC is formed morphologically or periphrastically. In an ideal scenario, the model would produce only very few misclassifications, i.e. predict that an instance will be periphrastic when in fact it is morphological. The fewer these errors, the better the model. Table 5 shows that 95.8 percent of all examples are categorized correctly.[9] It can thus be concluded that the analysis captures at least the most important aspects of the comparative alternation.

[9] That percentage is the weighted mean of the two percentages shown.

https://doi.org/10.1017/S1360674308002694 Published online by Cambridge University Press

Table 5. *Accuracy of the BNC analysis*

		Predicted		
		morphological	periphrastic	% correct
Observed	morphological	70,399	1,223	98.3%
	periphrastic	2,129	6,127	74.2%
			Overall correct:	95.8%

Table 6. *Comparison of three alternative models*

	Predicted		
	true	false	% correct
Model A (no phonology)	72,366	7,512	90.6%
Model B (no syntax)	76,160	3,718	95.3%
Model C (no frequency)	76,317	3,561	95.5%

How can we assess the relative impact of different categorical and continuous variables? One way to quantify their respective effects is simply to leave them out of the analysis: if the exclusion of a variable leads to impoverished results, then we know that it was responsible for the observed difference in accuracy. The same can of course be done with groups of variables. To illustrate, table 6 compares three alternative models, which were computed with subsets of the available information. Model A was not given any phonological variables, model B was not given any syntactic variables, and model C was not given the frequency values that have been entered into the original analysis. Expectedly, model A shows the poorest outcome. Ignoring the number of syllables, characteristics of the final segment, and stress patterns, the analysis deteriorates to a level of accuracy in model A that differs less than 1 percent from chance.

By contrast, ignoring the syntactic variables only leads to a minor deterioration. Model B is only slightly less accurate than the full analysis. The effect of the frequency variables is lesser still, though comparable in strength to the effect of syntax. Overall, the comparison gives us an approximate idea of how strongly each domain affects the comparative alternation.

Our discussion of model fit is not yet complete, though. The inverse problem of overlooking important factors is the inclusion of unnecessary factors. Since the above analysis is based on quite a large set of variables, it is necessary to ask whether it might not in fact be overly complex. The inclusion of too many parameters in a statistical model is referred to as overfitting. One way to test whether the analysis is overfitting the data is to confront it with a second body of previously unseen data. If the analysis based on the BNC poorly predicts what is found in other corpora, it

https://doi.org/10.1017/S1360674308002694 Published online by Cambridge University Press

Table 7. *Application of the BNC analysis to the EuroParl corpus*

		Predicted		
		morphological	periphrastic	% correct
Observed	morphological	2,558	58	97.8%
	periphrastic	142	143	50.2%
			Overall correct:	93.0%

would have to be concluded that it is overfitting the data, which would prohibit further generalization.

In order to select a reasonably challenging candidate for such a comparison, the EuroParl corpus (Koehn 2005) was chosen.[10] The EuroParl corpus differs in many ways from the BNC. First, as a record of parliamentary debate, it consists exclusively of texts that were transcribed from speech. This speech, however, was pre-written in many cases and generally falls into an elevated register. Further, some of the texts were translated into English from another European language. Besides its unique genre and modality, the corpus also shows clear thematic biases that lead to inflated frequencies of certain lexical items. For example, the adjective *fair* is more frequent in the EuroParl corpus than the adjectives *easy* and *simple* – the reverse is true for the BNC.

The procedures of data retrieval and coding described in section 3 were repeated in the same way for the EuroParl corpus. As the corpus is smaller in size, a more limited set of only 38 alternating adjectives was retrieved. Table A3 in the Appendix lists all alternating adjectives in the EuroParl corpus along with their frequencies in the morphological and the periphrastic comparative. In actual token frequencies, the retrieved examples yield 2,616 morphological comparatives and 285 periphrastic comparatives. The frequency asymmetry observed in the BNC thus extends to the EuroParl corpus; the retrieved morphological comparatives represent 90.2 percent of the data. Again, this number serves as a baseline that the analysis has to surpass. Table 7 shows how the BNC analysis carries over to the EuroParl corpus.

Overall, the BNC model correctly predicts 93 percent of the EuroParl data, which is a substantial improvement over what could have been predicted by chance only. While the categorization of periphrastic comparatives appears to be poor, a success rate of 50 percent is actually five times better than the 10 percent we would expect from chance alone. The results of the comparison are thus encouraging, given the substantial differences between the two corpora.

[10] The specific part of the EuroParl corpus that was chosen was the English part of the pair English–Swedish. The English texts comprise a total of 15.8 million words.

https://doi.org/10.1017/S1360674308002694 Published online by Cambridge University Press

5 Conclusions

The analysis presented in the preceding sections allows a fairly optimistic conclusion. The comparative alternation is a problem that is complex, but not completely elusive. It could be shown that previous analyses have succeeded in finding many of the relevant factors of language structure. Paired with two additional factors of language use, these factors could be used to predict the distribution of the two variants with a high degree of accuracy. This of course does not mean that the present analysis lays claim to having identified all relevant factors. Semantic factors such as abstractness and discourse-based factors such as emphasis or givenness, which Mondorf (2003: 296) includes in her analysis, were disregarded here not for theoretical reasons, but because these factors do not easily lend themselves to objective annotation. It is left as a challenge for future research to operationalize these variables in such a way that they can be integrated into a quantitative investigation of the two English comparatives.

A second aim of this article was to determine the relative importance of different variables that affect the comparative alternation. It was shown that structural phonological factors exert the strongest influence, while syntactic factors and factors of language use contribute in minor ways to the observed distribution.

A few more general considerations are in order at this point. Of course, an analysis that merely points out factors that lie behind a binary choice can be criticized for remaining silent on the most important question: why do we see this constellation of factors? How is all of this motivated? Mondorf (2003) realizes that questions like these need to be answered and proposes an explanation in terms of processing effort. She invokes Rohdenburg's (1996) *complexity principle*, which states that more explicit linguistic structures are favored in cognitively more complex environments. Mondorf argues that the periphrastic comparative, due to its greater explicitness, affords a processing advantage that biases speakers toward using it in situations that are cognitively complex. Many of the factors underlying the comparative alternation do relate to this: it is a more complex task to produce a trisyllabic word than it is to articulate a disyllabic word, a *to*-infinitive complement adds syntactic complexity, a less frequent adjective may not be as easily retrieved and produced as a highly frequent one. What Mondorf suggests is essentially a reductive explanation – a large array of factors can be reduced to a single one: cognitive complexity. While this is a very attractive solution, some aspects of the present analysis may prove problematic for it.

First, a reductive explanation has to stipulate that adding an *-er* suffix to a final /i/ is cognitively less complex than adding the suffix to a final /l/, which is difficult to maintain. The transition from /i/ to /ər/ may involve an intrusive glide, as in /kri:mijər/, making it more complex than the unmediated transition from /l/ to /ər/. Further, nouns ending in /l/ easily undergo *-er* nominalization, as in *caller* or *filler*, which do not seem particularly complex to produce. Whatever reason there is for final

https://doi.org/10.1017/S1360674308002694 Published online by Cambridge University Press

/l/ biasing speakers against the morphological comparative, phonological complexity is an unlikely explanation.

Second, an adjective followed by a *than*-phrase is arguably more complex than an adjective without such a phrase. Yet, a following *than* introduces a slight bias toward the morphological comparative. This contradicts the harmonic alignment of complex structures with the periphrastic comparative.

These considerations do not invalidate Mondorf's point that much of the observed variance in the comparative alternation can be explained with reference to complexity. They do, however, caution against a monocausal explanation of the comparative alternation, which carries some risk of circularity. Consider again the case of *than*. If we observe that a following *than* favors the morphological comparative, it is tempting to conclude that it does so because it alleviates some cognitive complexity, presumably by making explicit the otherwise unexpressed standard of comparison. We arrive at this conclusion because it fits our theory, not because it is supported by independent evidence. When analyzing complex phenomena, we should thus not rule out the possibility that the underlying explanation might be equally complex.

The results of the present study suggest that the comparative alternation is governed by functionally motivated factors as well as by formal phonological factors that do not necessarily reflect such a motivation. This may seem a provocative statement to the functional linguistic community, and indeed it is intended to be a strong hypothesis which I hope others will attempt to falsify.

Author's address:
Freiburg Institute for Advanced Studies (FRIAS)
Albertstr. 19
79104 Freiburg i.Br.
Germany
martin.hilpert@frias.uni-freiburg.de

Appendix

Table A1. *Relative frequency of morphological comparatives by number of syllables (Leech & Culpeper 1997: 355)*

	LOB		BNC	
	morphological	periphrastic	morphological	periphrastic
one syllable	1502	16	1213	8
two syllables	229	217	161	221
≥ three syllables	0	517	1	402

https://doi.org/10.1017/S1360674308002694 Published online by Cambridge University Press

Table A2a. *Comparative frequencies in the BNC:* able *to* mad

ADJ	*more*	*-er*	ADJ	*more*	*-er*	ADJ	*more*	*-er*
able	170	6	cosy	6	23	glossy	3	13
absurd	16	1	crafty	3	3	grand	6	175
ample	9	2	crazy	3	23	great	1	15936
angry	42	76	creamy	2	13	grey	6	67
apt	36	13	crude	5	71	grim	5	38
big	1	4466	curly	2	12	gross	1	15
black	5	85	daft	2	14	guilty	12	2
bland	5	11	damp	2	54	hairy	2	14
bleak	4	23	dark	5	802	handsome	24	10
blind	3	38	dead	19	4	happy	10	1007
bloody	2	10	deadly	30	9	hard	1	1745
blue	6	29	deaf	2	7	hardy	3	32
blunt	5	30	dense	19	147	harsh	3	166
bold	4	178	diffuse	34	14	hazy	2	1
brash	2	38	dire	5	2	healthy	26	500
brave	2	70	dirty	1	64	heavy	4	1072
brief	2	50	dreary	2	2	hefty	3	7
bright	4	642	dry	4	163	holy	5	18
brisk	2	33	dull	2	58	homely	13	4
broad	7	1588	earthy	7	4	hot	1	270
brown	4	39	easy	28	4031	humble	25	77
bulky	7	26	empty	4	14	hungry	5	43
bumpy	2	7	faint	2	139	idle	6	2
busy	8	118	fair	5	322	ill	7	5
canny	3	3	feeble	7	15	intense	169	4
cheap	1	2786	fierce	11	82	just	8	9
cheery	1	11	fine	1	627	keen	15	120
choosy	3	3	fit	3	201	kindly	4	7
chunky	2	5	fond	3	31	lax	7	10
classy	3	13	frail	11	16	leafy	2	10
clear	36	1032	free	21	270	lean	2	87
clever	21	101	fresh	3	144	leggy	2	2
close	3	2539	friendly	73	55	lengthy	6	23
cloudy	8	4	full	19	620	light	2	953
clumsy	5	12	funky	3	3	likely	3724	17
coarse	3	104	funny	8	70	lively	49	76
cold	4	383	fussy	4	2	lofty	2	24
comfy	5	4	gentle	47	221	lonely	7	13
common	594	73	ghostly	3	1	lovely	7	42
compact	18	4	glad	5	8	lowly	2	6
corrupt	5	1	glitzy	2	1	lucky	5	65
costly	82	19	gloomy	7	42	mad	3	71

https://doi.org/10.1017/S1360674308002694 Published online by Cambridge University Press

Table A2b. *Comparative frequencies in the BNC:* manly *to* yellow

ADJ	*more*	*-er*	ADJ	*more*	*-er*	ADJ	*more*	*-er*
manly	4	1	rosy	4	11	stormy	3	5
mature	141	14	rough	7	103	straight	2	92
mellow	11	14	round	3	52	strange	19	208
messy	3	27	rowdy	2	5	strict	11	266
mighty	2	30	rude	4	26	strong	11	3194
moist	3	6	sad	7	94	sturdy	2	18
murky	2	5	safe	12	933	subtle	339	114
narrow	14	550	salty	4	5	supple	15	1
nasty	1	76	scarce	5	55	sure	19	61
naughty	2	5	scary	7	11	sweet	2	157
new	10	601	secure	156	4	tasty	3	16
noble	9	35	severe	227	9	tender	15	6
noisy	6	34	sexy	5	37	tense	9	3
obscure	40	2	shadowy	8	2	thin	3	544
odd	8	56	shaky	3	8	tidy	8	36
oily	3	7	shallow	6	125	tight	1	579
pale	3	296	sharp	3	368	tricky	13	43
patchy	3	2	shrill	2	13	ugly	4	41
petty	2	4	shy	2	12	unhappy	18	15
pleasant	90	50	silly	1	23	unlikely	27	1
polite	23	7	simple	60	1115	unruly	3	1
poor	3	1115	sincere	12	2	untidy	2	2
profound	106	11	sleepy	7	7	vague	8	39
proud	4	14	slender	15	44	vain	4	3
pure	3	94	slight	2	35	vast	4	21
quick	8	770	small	5	8816	vile	3	3
quiet	7	454	sober	50	4	warm	2	344
quirky	3	4	soft	5	563	wary	21	6
racy	2	6	sorry	12	18	weak	2	913
rare	22	231	sour	3	4	wealthy	15	148
raw	5	4	sparse	7	20	weary	3	9
ready	55	24	speedy	6	69	weighty	8	24
real	109	4	spicy	2	7	weird	5	27
red	8	100	spooky	2	1	wet	1	134
remote	179	87	sporty	2	14	white	9	118
rich	4	877	sprightly	2	1	wild	1	189
right	7	5	stable	86	9	worldly	20	1
risky	42	47	stale	2	6	worthy	38	17
robust	95	1	stark	9	17	yellow	8	9
rocky	3	16	steady	4	31			
roomy	4	5	sticky	5	6			

https://doi.org/10.1017/S1360674308002694 Published online by Cambridge University Press

Table A3. *Comparative frequencies in the European Parliament Proceedings (EN-SV)*

ADJ	*more*	*-er*	ADJ	*more*	*-er*	ADJ	*more*	*-er*
angry	1	1	grave	2	2	sad	2	6
bold	3	25	harsh	2	15	safe	3	211
brief	1	8	healthy	1	61	scarce	2	6
calm	1	10	just	20	1	simple	11	165
clear	13	472	keen	1	8	slow	1	35
common	40	2	likely	111	1	sound	1	18
easy	2	912	needy	3	1	speedy	1	21
fair	5	150	noble	2	2	stark	1	2
firm	1	40	noisy	1	2	strict	6	356
frank	2	1	prompt	2	2	vague	1	7
free	3	35	rare	1	2	wealthy	5	25
friendly	7	2	remote	23	5			
gloomy	1	1	risky	2	3			

References

Barlow, Michael & Suzanne Kemmer (eds.). 2000. *Usage-based models of language*. Stanford: CSLI Publications.

Boyd, Jeremy K. 2007. Comparatively speaking: A psycholinguistic study of optionality in grammar. PhD dissertation, UCSD.

Braun, Albert. 1982. *Studien zur Syntax und Morphologie der Steigerungsformen im Englischen*. Bern: Francke.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina & R. Harald Baayen. 2005. Predicting the dative alternation. In G. Boume, I. Kraemer & J. Zwarts (eds.), *Cognitive foundations of interpretation*, 69–94. Amsterdam: Royal Netherlands Academy of Science.

Bybee, Joan L. 1985. *Morphology: A study of the relation between meaning and form*. Amsterdam: John Benjamins.

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

di Sciullo, Anna Maria & Edwin Williams. 1987. *On the definition of word*. Cambridge, MA: MIT Press.

Fries, Udo. 1993. The comparison of monosyllabic adjectives. In A. H. Jucker (ed.), *The noun phrase in English: Its structure and variability*, 25–44. Heidelberg: Winter.

Gahl, Susanne & Susan Garnsey. 2004. Knowledge of grammar, knowledge of usage: Syntactic probabilities affect pronunciation variation. *Language* 80(4), 748–75.

Hooper, Joan B. 1976. Word frequency in lexical diffusion and the source of morphophonological change. In W. Christie (ed.), *Current progress in historical linguistics*, 96–105. Amsterdam: North Holland.

Johansson, Stig, Geoffrey Leech & Helen Goodluck. 1978. Manual of information to accompany the Lancaster–Oslo/Bergen Corpus of British English, for use with digital computers. Unpublished document: Department of English, University of Oslo.

Johnson, Keith. 2007. *Quantitative methods in linguistics*. Oxford: Blackwell.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. *MT Summit*.

https://doi.org/10.1017/S1360674308002694 Published online by Cambridge University Press

Kytö, Merja & Suzanne Romaine. 1997. Competing forms of adjective comparison in Modern English: What could be more quicker and easier and more effective? In T. Nevalainen & L. Kahlas-Tarkka (eds.), *To explain the present: Studies in the changing English language in honour of Matti Rissanen*, 329–52. Helsinki: Mémoires de la Société Néophilologique de Helsinki.

Leech, Geoffrey N. 1992. 100 million words of English: The British National Corpus. *Language Research* 28(1), 1–13.

Leech, Geoffrey N. & Jonathan Culpeper. 1997. The comparison of adjectives in recent British English. In T. Nevalainen and L. Kahlas-Tarkka (eds.), *To explain the present: Studies in the changing English language in honour of Matti Rissanen*, 353–74. Helsinki: Mémoires de la Société Néophilologique de Helsinki.

Lindquist, Hans. 1998. The comparison of English disyllabic adjectives in *-y* and *-ly* in Present-day British and American English. In H. Lindquist et al. (eds.), *The major varieties of English. Papers from MAVEN 97*, 205–12. Växjö: Acta Wexionensia.

Lindquist, Hans. 2000. *Livelier* or *more lively?* Syntactic and contextual factors influencing the comparison of disyllabic adjectives. In J. M. Kirk (ed.), *Corpora galore: Analyses and techniques in describing English*, 125–32. Amsterdam: Rodopi.

Mondorf, Britta. 2003. Support for *more*-support. In G. Rohdenburg and B. Mondorf (eds.), *Determinants of grammatical variation in English*, 251–304. Berlin: Mouton de Gruyter.

Mondorf, Britta. 2007. Recalcitrant problems of comparative alternation and new insights emerging from internet data. In M. Hundt et al. (eds.), *Corpus linguistics and the Web*, 211–32. Amsterdam: Rodopi.

Poser, W. J. 1992. Blocking of phrasal constructions by lexical items. In I. Sag and A. Szabolsci (eds.), *Lexical matters*, 111–30. Stanford: CSLI Publications.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. 1985. *A comprehensive grammar of the English language*. New York: Longman.

Rohdenburg, Günter. 1996. Cognitive complexity and increased grammatical explicitness in English. *Cognitive Linguistics* 7(2), 149–82.

Rohr, Anny. 1929. Die Steigerung des neuenglischen Eigenschaftswortes im 17. und 18. Jahrhundert mit Ausblicken auf den Sprachgebrauch der Gegenwart. PhD dissertation, Universität Giessen.

Webster's unabridged dictionary (The Project Gutenberg etext). 1996. Plainfield, NJ: MICRA Inc.

https://doi.org/10.1017/S1360674308002694 Published online by Cambridge University Press