CrossMark

English Language and Linguistics 20.3: 559–583. © Cambridge University Press 2016
doi:10.1017/S1360674316000368

Determinants of the synthetic–analytic variation across English comparatives and superlatives[1]

LAWRENCE CHEUNG and LONGTU ZHANG
The Chinese University of Hong Kong

(Received 18 December 2015; revised 15 June 2016)

Some English adjectives accept both synthetic and analytic comparative and superlative forms (e.g. *thicker* vs *more thick*, *happiest* vs *most happy*). As many as 20+ variables have been claimed to affect this choice (see Leech & Culpeper 1997; Lindquist 2000; Mondorf 2003, 2009). However, many studies consider one variable at a time without systematically controlling for other variables (i.e. they take a monofactorial approach). Further, very little research has been done on superlatives. Following Hilpert's (2008) multifactorial study, we investigate the simultaneous contribution of 17 variables towards comparative and superlative alternation and further measure the strength(s) of the predictors. On the whole, phonological predictors are much more important than syntactic and frequency-related predictors. The predictors of the number of syllables and final segments in <-*y*> consistently outrank other predictors in both models. Important differences have also been identified. Many syntactic variables, such as predicative position and presence of complements, are weak or non-significant in the comparative model but have stronger effects in the superlative model. Further, higher frequency of an adjective leads to a preference for the synthetic *-er* variant in comparatives but the analytic *most* variant in superlatives. The study shows that generalizations about comparatives do not straightforwardly carry over to superlatives.

1 Introduction

It has long been recognized that monosyllabic adjectives strongly prefer synthetic comparatives/superlatives (i.e. *-er*/*-est*) whereas adjectives longer than two syllables almost always take analytic forms (i.e. *more/most*) (see Thomson & Martinet 1980; Quirk *et al.* 1985: 461; Huddleston & Pullum 2002: 1583; Biber *et al.* 1999: 522, among others). Nevertheless, some adjectives can take both synthetic and analytic forms, as shown in the following examples from the *British National Corpus* (BNC).

[1] We want to express our gratitude to Britta Mondorf, Javier Pérez-Guerra and three anonymous reviewers for comments and suggestions on an earlier draft of this article. An earlier version of the article was presented at the 5th International Conference on the Linguistics of Contemporary English (ICLCE 5), University of Texas, Austin, 25–29 September 2013. We want to thank the audience for their comments. Last but not least, thanks go to Mercy Wong and Carleon Mendoza for their hard work and patience with the annotation of the examples. All remaining errors are ours.

Downloaded from http:/www.cambridge.org/core. Chinese University of Hong Kong, on 27 Oct 2016 at 12:48:53, subject to the Cambridge Core terms of use, available at http:/www.cambridge.org/core/terms. http://dx.doi.org/10.1017/S1360674316000368

(1) (a) Today he is to visit Gettysburg, Pennsylvania, the site in 1863 of the bloodiest battle of the American Civil War. (BNC: A1V-785)
(b) Later they went to the scene of one of the most bloody battles in the 1950–53 Korean war affecting British troops. (BNC: CEM-524)
(2) (a) Low rateable values in some of the remoter rural areas also played a part. (BNC: FPR-847)
(b) Everyone knows how good he is, or nearly is, but somehow he remains a more remote figure. (BNC: AHK-217)

A number of studies have examined determinants of the S-A variation, e.g. presence of complements, final segments of the respective adjective and frequency (see section 3 for relevant studies).

There are at least three gaps in the literature. First, previous research has shown that as many as 20+ variables can potentially have an effect on the S-A variation. Many researchers study the effect of one single variable at a time without systematically controlling for other variables. It is important to examine the S-A variation using a model that can take into consideration a range of variables simultaneously. Hilpert (2008) is a notable exception in that he takes into account a wide variety of variables using logistic regression. Our methodology is similar to his. Second, previous research on S-A alternation is heavily skewed towards comparatives. Superlatives have largely been neglected. Third, some are based on relatively small data samples (e.g. Leech & Culpeper 1997; Lindquist 2000).

Our research will address these limitations in several ways. We have adopted a multifactorial methodology to systematically compare the S-A variation in English comparatives and superlatives. Special attention will be paid to the difference in the respective predictors' contribution to the analysis of comparatives and superlatives. We included additional predictor variables that have been mentioned previously but were not studied in Hilpert (2008). To ensure that findings are reliable and representative, two large datasets derived from the BNC have been used to build the regression models. Three research questions are to be tackled in this article.

(3) (a) What are the relevant predictor variables affecting S-A variation of English comparatives and superlatives?
(b) Among those identified in (a), which are the strongest predictors?
(c) Is S-A variation in comparatives and superlatives determined by the same factors?

The remainder of the article is organized as follows. Section 2 reviews the approaches of previous studies on S-A alternation and illustrates some shortcomings. The list of predictor variables identified is briefly described in section 3. In section 4, we introduce the corpus data extracted from the BNC and the regression methodology used to determine variable strength. Section 5 provides the results of the regression models. Section 6 discusses similarities and differences between S-A variation of comparatives and superlatives. A conclusion will be drawn in section 7.

Downloaded from http:/www.cambridge.org/core. Chinese University of Hong Kong, on 27 Oct 2016 at 12:48:53, subject to the Cambridge Core terms of use, available at http:/www.cambridge.org/core/terms. http://dx.doi.org/10.1017/S1360674316000368

2 Synthetic–analytic alternation of comparatives and superlatives

2.1 Monofactorial studies

Most studies on comparative alternation (see Leech & Culpeper 1997; Lindquist 2000; Mondorf 2003, 2009; González-Díaz 2008, among others) and superlatives (see Claridge 2007; Lindquist 2000) focus on the identification of variables that affect S-A alternation. As documented in Mondorf (2009), at least 25 variables (in eight categories) have been reported in the literature,[2] as shown in (4).

(4) (a) Phonological: e.g. avoidance of stress clash, avoidance of certain final segments
(b) Morphological: e.g. morphological complexity
(c) Syntactic: e.g. various types of complements, attributive vs predicative position
(d) Semantic: e.g. abstractness, figurativity, gradability
(e) Lexicon: e.g. length, frequency, parallel structure
(f) Pragmatic: e.g. end-weight, emphasis, gradual increase
(g) Historical: e.g. Old English, Middle English, Early Modern English
(h) Variety: British vs American English

Some previous studies adopt a quantitative approach based on relatively small data samples. For instance, among the variables studied, Lindquist (2000) finds that premodification of the comparative (e.g. 'the *single* deadliest strike') seems to favour the analytic form.[3] In his corpus of *The Independent*, 46 per cent (57 out of 123) of analytic comparatives are premodified while only 33 per cent (42 out of 127) of synthetic comparatives are premodified.

Mondorf (2003, 2009) and González-Díaz (2008) adopt a more stringent methodology. Although they also focus on the effect of one independent variable at a time, they control for one or two potentially intervening variables. For example, to study the effect of *to*-infinitive complements on comparative alternation, word length (monosyllabic vs disyllabic) and positional distribution (attributive vs non-attributive) are systematically controlled for (see Mondorf 2003: 261–3). These measures help isolate the contribution of individual variables to S-A variation.

Nevertheless, monofactorial approaches have inherent limitations.[4] First, since 20+ variables may simultaneously contribute to the variation, it is not feasible to systematically control for all other 20+ variables when one studies a particular variable. Even with Mondorf's (2009) methodology, only a few variables can be controlled for at a time, thus limiting our understanding of the simultaneous contribution of the other variables. Multifactorial models offer more explanatory potential by permitting the comparison of multiple – though not all 20+ – variables at the same time. Second, it is difficult to compare the contribution of different variables

[2] The variables tested in our study are described in section 3. Interested readers are referred to Mondorf (2009) for a comprehensive description of the variables.

[3] Lindquist (2000: 127, table 4) is based on (i) 212 comparative examples from the *New York Times* 1995, and (ii) 250 comparative examples from *The Independent* 1995 (Lindquist 2000: 125).

[4] Some of the weaknesses of monofactorial analyses have also been mentioned in Hilpert (2008: 398).

Downloaded from http:/www.cambridge.org/core. Chinese University of Hong Kong, on 27 Oct 2016 at 12:48:53, subject to the Cambridge Core terms of use, available at http:/www.cambridge.org/core/terms. http://dx.doi.org/10.1017/S1360674316000368

as determined in a monofactorial analysis. One cannot easily make predictions about the choice of comparative variant if individual variables lead to conflicting S-A predictions. Third, some previous studies (e.g. Lindquist 2000) are based on relatively small datasets.

Apart from identifying various variables, Mondorf (2003, 2009) attempts to provide a principled account for why the synthetic and analytic comparatives are correlated with various properties or contexts. Drawing heavily on processing research (see Hawkins 1994; Rohdenburg 1996), Mondorf argues that one of the important principles determining the relationship between the variables and the S-A outcome is processing efficiency. She proposes the notion of *more*-support:

> *More*-support: In cognitively more demanding environments which require an increased processing load, language users tend to make up for the additional effort by resorting to the analytic (*more*) rather than the synthetic (*-er*) comparative. (Mondorf 2009: 6)

More-support thus works as a support strategy, mitigating the processing load in hard-to-process environments (for further examples of analytic variants as support strategies in English, Spanish and German see Mondorf 2014). The analytic form is more explicit and incurs a lower processing demand. As a result, the analytic variant is preferred when the comparative is more complex phonologically, lexically, syntactically, semantically or pragmatically. For example, the presence of an adjectival complement tends to favour the use of analytic *more*. In Mondorf's (2009) analysis, the complement results in a higher degree of syntactic complexity, leading to the higher frequency of the analytic form. We will report in section 6 whether our findings are consistent with this hypothesis.

2.2 Multifactorial study

Hilpert (2008) utilizes logistic regression to provide an integral account of the influence of 16 variables on S-A alternation. The advantage of regression analysis is that it can evaluate the significance of each predictor variable simultaneously. The approach represents a significant improvement over previous studies. Hilpert's regression analysis fitted to 79,878 comparatives from the BNC shows that 15 out of 16 variables have a statistically significant influence on S-A selection. To evaluate the relative importance of three variable groups (i.e. phonological vs syntactic vs frequency variables), three additional regression models were constructed, each of which excludes one group of variables. The model with the lowest accuracy suggests that the deleted variable group has the strongest influence on S-A alternation. Hilpert concludes that phonological variables are strong determinants, and that syntactic variables and lexical frequency are much weaker predictors.

The findings of Hilpert (2008) will be compared with ours in sections 5 and 6 as both studies use similar methodology to study S-A variation of comparatives. It is noteworthy that the predictors and the data used in the two studies are not exactly the same. Each study has a few predictors such as `Complement(P)` and `FS_s_st_sh`

Downloaded from http:/www.cambridge.org/core. Chinese University of Hong Kong, on 27 Oct 2016 at 12:48:53, subject to the Cambridge Core terms of use, available at http:/www.cambridge.org/core/terms. http://dx.doi.org/10.1017/S1360674316000368

that are exclusive to the study. Further, the definitions of some variables differ slightly. `Complement` in Hilpert's study is restricted to only *to*-infinitive complements. Also, our study distinguishes three types of positions (attributive, predicative and postnominal) whereas Hilpert only distinguishes the former two.

Although our multifactorial analysis largely mimics Hilpert's, we will take on three issues not tackled in his studies. First, while Hilpert (2008) has only dealt with comparatives, we will highlight the similarities and differences between comparatives and superlatives. Second, Hilpert's analysis of predictive power was run for variable groups but not for individual variables. For example, as both final stress and final consonant cluster are part of the phonological group, one cannot easily tell which one exerts a stronger influence on S-A alternation. Third, even though our study cannot include all variables proposed in the past, we include a few variables proposed in other studies that were not considered in Hilpert's (2008) study, e.g. prepositional complement and final sibilant segments (see Mondorf 2009).

3 Description of predictor variables

The predictor variables selected in this study were previously reported to be significant to the S-A alternation of comparatives or superlatives. All predictors are categorical variables, except `SyllNum`, `Pos_Freq`, `CompPos_Ratio` and `SuplPos_Ratio`.

3.1 Phonological predictors

A. *Number of syllables (SyllNum).* It is well established that the number of syllables of an adjective is an important determinant of S-A alternation (see Huddleston & Pullum 2002; Biber *et al.* 1999; Mondorf 2009). Shorter adjectives (e.g. *free*) strongly prefer the synthetic form, whereas longer ones (e.g. *comfortable*) favour the analytic form. `SyllNum` refers to the number of syllables of the positive adjective and is a continuous variable.

B. *Final segments.* Based on earlier research, Kytö & Romaine (1997) report that the final segments of adjectives influence S-A preference. Final segments, such as *-y*, *-ly*, *-le* and *-er* have been examined. Final segments in /i/ (e.g. *easy*, *friendly*), /l/ (e.g. *cruel*) or /ɹ/ (e.g. *clever*) favour the synthetic form. Mondorf (2009: 19) and Hilpert (2008: 309), on the other hand, point out that adjectives ending in /ɹ/ are more likely to be rendered analytically. Lindquist (1998) as well as Leech & Culpeper (1997: 359) distinguish *-ly* from *-y* (even though they both end in /i/), claiming that *-ly* adjectives are more analytic-oriented. The list of final segments is provided in table 1. The predictor `FS_ly` is reserved specifically for adjectives ending in *-ly* (e.g. *friendly*), and the predictor `FS_y`, for /i/-endings other than *-ly*. Syllabic /ɹ/ and /l/ final segments are also separated.[5]

[5] Kytö & Romaine (1997) group the syllabic /ɹ/ and /l/ final segments as one single variable *-le*/*-er*.

Downloaded from http:/www.cambridge.org/core. Chinese University of Hong Kong, on 27 Oct 2016 at 12:48:53, subject to the Cambridge Core terms of use, available at http:/www.cambridge.org/core/terms. http://dx.doi.org/10.1017/S1360674316000368

Table 1. *Phonological variables for various adjective final segments (A = analytic; S = synthetic)*

Predictor	Adjective final segment	S-A preference
`FS_l`	/l/ (e.g. *–le*)	A
`FS_ly`	*–ly*	S
`FS_r`	/ɹ/ (e.g. *-r/-re*)	mixed
`FS_y`	*-y*	S

C. *Avoid haplology*. Some studies (Sweet [1891] 1968: 327, as cited in Mondorf 2009: 24; Plag 1998; Kytö & Romaine 2000: 185) note a tendency for adjectives ending in /ɹ/ and /s, st, ʃ/ to take analytic comparatives and superlatives respectively. It has been suggested that the (near-)adjacent repetition of identical features such as /ɹ/ (e.g. *clearer, severer*) and sibilant consonants (e.g. *harshest*, *vastest*) makes the sequence more difficult to pronounce. As a result, the analytic variant is favoured in order to avoid phonological haplology. Two predictors are introduced to capture the effect in both the comparative and superlative logistic regression models.

- `FS_r`: [same as the predictor mentioned in B]
- `FS_s_st_sh`: adjective ending in a sibilant, i.e. /s/, /st/ or /ʃ/.

One may wonder why `FS_r` is needed in the superlative model. After all, the *-est* morpheme does not clash with an adjective X_r ending in /ɹ/. We still include `FS_r` because /ɹ/ could potentially disfavour the use of the *-er* morpheme. The rarer use of the synthetic comparative marker *-er* on X_r may result in a dispreference of the synthetic superlative *-est* on X_r for consistency across the two related categories. Including both predictors in the analyses allows us to detect such an effect. By the same token, `FS_s_st_sh` is included in the comparative model.

D. *Consonant cluster (Cons_Cluster)*. Consonant cluster endings /pt/ (e.g. *apt*, *corrupt*) and /kt/ (*direct*, *correct*) have been claimed to increase processing effort (Görlach 1991: 93; Mondorf 2009: 31–3). As a support strategy which compensates for the added cognitive effort, the analytic variant is more often used among these adjectives (see Mondorf 2009: 31–3). The predictor indicates the presence of consonant clusters at the end of the adjective. Following Hilpert (2008: 400), adjectives ending in /l/ and in a consonant cluster at the same time such as *ample* only receive a positive value for `Cons_Cluster` but not `FS_l`.

E. *Stress on final syllable (Final_Stress)*. Adjectives with stress on the final syllable are more likely to take the synthetic variant *-er* (see Leech & Culpeper 1997). Mondorf's (2009: 23) explanation for the preference is that the *-er* suffix can serve as a 'buffer' to avoid two consecutive stresses[6] in a row when the comparative is immediately followed by an initially stressed noun, as shown in (5).

[6] This is referred to as the *Principle of Rhythmic Alternation* (Schlüter 2005: 18).

Downloaded from http:/www.cambridge.org/core. Chinese University of Hong Kong, on 27 Oct 2016 at 12:48:53, subject to the Cambridge Core terms of use, available at http:/www.cambridge.org/core/terms. http://dx.doi.org/10.1017/S1360674316000368

(5) (a) a sin**ce**rer **ge**sture (two stressed syllables separated by an unstressed one)
(b) a more sin**cere ge**sture (two consecutive stressed syllables)

Although Mondorf (2009: 22) notes this tendency, the effect is not very strong. The variable `Final_Stress` helps to check if the final stress of an adjective plays a role in the S-A alternation.

3.2 Syntactic predictors

F. *Positional distribution (Position)*. Predicative comparatives are positively correlated with the analytic form whereas attributive comparatives correlate with the synthetic form (see Leech & Culpeper 1997: 357; Lindquist 2000: 126; González-Díaz 2008: 110–11; Mondorf 2003: 286–7, 2009: 80). Leech & Culpeper (1997) suggest that the attributive use of *more* could result in ambiguity between a comparative marker (e.g. [*more happy*] *stories*) and a determiner (e.g. *more* [*happy stories*]). To avoid the ambiguity, synthetic comparatives are favoured in attributive position. Leech & Culpeper (1997) as well as Mondorf (2009) include a third category for 'postnominal' position, as in (9). They observe that postnominal comparatives pattern with predicative comparatives, preferring the analytic form. Therefore, the predictor 'positional distribution' has four possible values:

- Predicative (P): adjective serving as a predicate,[7] e.g. (7)
- Attributive (A): adjective preceding and modifying a nominal, e.g. (8)
- Postnominal (N): adjective occurring after a nominal,[8] e.g. (9)
- Others (O): all positions other than the above, e.g. correlative comparatives (10a) and nominalized superlatives, e.g. *the coldest of stares*.

(6) I should have been more strict with you. (BNC: HWC-1030)
(7) Where these signs are not obvious, subtler discriminations can be made … (BNC: CAH-1184)
(8) She had put on something rather more dressy. (BNC: H8F-1369)
(9) … the angrier Gebrec became, the more Erdle taunted him. (BNC: GVP-2998)

G. *Premodification*. When the comparative/superlative is immediately preceded by a modifying adverb,[9] like *considerably*, *even*, *ever*, *far*, *much*, *significantly*, *slightly*, *still*, *yet*, *a bit* or *a lot*, the example will be marked as being premodified, as illustrated in the examples below.

(10) The note taken from the victim's purse was even briefer. (BNC: G3E-1264)
(11) The very simplest kind of perceptron consists of a single node. (BNC: FNR-1772)

[7] Predicative comparatives are generally preceded by a copula or verbs like *seem* and *remain*. We also coded those examples where the comparative/superlative is the predicate of a small clause as 'Predicative', e.g. *To make them easier to serve*, … (BNC: ED4-2945).
[8] The comparative/superlative typically follows a noun and can be paraphrased as 'who/which is X-er/more X'.
[9] Note that the range of modifiers for superlatives is much more restricted than that for comparatives.

Downloaded from http:/www.cambridge.org/core. Chinese University of Hong Kong, on 27 Oct 2016 at 12:48:53, subject to the Cambridge Core terms of use, available at http:/www.cambridge.org/core/terms. http://dx.doi.org/10.1017/S1360674316000368

H. *Complement type (Complement)*. Poutsma (1914: 477, cited in Mondorf 2009: 57) observed early on that adjectives with a complement[10] tend to form comparatives analytically. `Complement` has four possible values, namely, *to*-infinitive complement (T) (13), PP complement (P) (14, 15), other complements (O) (16) and no complements (Z).

(12) Telecommunications were easier to build, service, and tap in this way. (BNC: A64-1512)
(13) One set of parents were very understanding but the others were more angry about it, although they are calmer now. (BNC: CBF-5492)
(14) But when I go from hence to this testing I must leave affairs here in one pair of hands, the closest and dearest to me. (BNC: K8S-2138)
(15) And in any case, it's most likely that it was a completely misleading headline in the press. (BNC: KGR-458)

The proportion of analytic comparatives is higher when the comparative has a *to*-infinitive complement (see Mondorf 2003: 262, 2009: 59) or a prepositional complement (see Mondorf 2003: 266, 2009: 72). Hilpert (2008: 408) confirms that the *to*-infinitive complement is a significant predictor of the alternation, but he does not test for prepositional complements.

I. *Presence of* than *phrase (Than_Phrase)*. There has been speculation that the *than* phrase is a complement of the comparative and might lead to more frequent use of *more*. Lindquist (2000) considers the *than* phrase a 'postmodification', not a complement. Its effect is debatable, though. While many studies (see Leech & Culpeper 1997; Lindquist 2000; Mondorf 2009) found no noticeable correlation, Hilpert (2008) did. Mondorf (2009: 78–80) explains that the stipulation of a correlation results from a misconception, because the *than* phrase may not be treated as an adjectival complement in the first place, as it is dependent on the degree marker rather than the adjective itself. This predictor is included in our comparative model, though we are aware of its doubtful status as an adjectival complement.

3.3 Lexical frequency predictors

High-frequency adjectives tend to be shorter in length and prefer synthetic *-er*. Mondorf's (2009: 40) data shows that highly frequent comparatives correlate with the synthetic variant. Hilpert (2008: 402) furthermore uses the ratio of comparative/positive in his logistic regression model. He argues that a high ratio of comparative/positive implies that the adjective is highly gradable, and highly gradable adjectives are more likely to form comparatives synthetically. The following continuous predictors are used in our analysis.

[10] Quirk *et al.* (1985: 1220–31) claim that there are six types of complementation in English adjectives, i.e. prepositional phrase, *that*-clause, *wh*-clause, *than*-clause, *to*-infinitive clause and *-ing* participle clause. However, Mondorf (2009) argues that *than* clauses should not be treated as the complement of the adjective.

Downloaded from http:/www.cambridge.org/core. Chinese University of Hong Kong, on 27 Oct 2016 at 12:48:53, subject to the Cambridge Core terms of use, available at http:/www.cambridge.org/core/terms. http://dx.doi.org/10.1017/S1360674316000368

Table 2. *Search queries for comparatives and superlatives (AJ0 = positive adjectives, AJC = comparative adjectives, AJS = superlative adjectives, AV0 = adverbs)*

Form	Comparatives	Superlatives
Synthetic	`*_AJC`	`*_AJS`
Analytic	`more_AV0 *_AJ0`	`most_AV0 *_AJ0`

- `Pos_Freq`: frequency of the positive form of the adjective
- `CompPos_Ratio`: comparative/positive ratio (for comparatives only)
- `SuplPos_Ratio`: superlative/positive ratio (for superlatives only)

4 Methodology

4.1 Corpus data

The data used were extracted from the BNC XML edition, a 100-million-word collection of spoken and written texts. It contains British English near the end of the twentieth century (see Aston & Burnard 1997; Burnard 2007). The BNC includes 90 million words of written English and 10 million words of spoken English. Its size makes it possible to collect a large amount of tokens. Also, the BNC has part-of-speech tags for positive adjectives (`AJ0`), comparative adjectives (`AJC`), and superlative adjectives (`AJS`) (Burnard 2007), making it easy to retrieve examples with the search queries[11] in table 2. Mondorf (2009: 202) comments that due to possible tagging errors in the BNC, some relevant examples may not be correctly retrieved. However, she also notes that the use of a large amount of examples from the BNC would likely not considerably skew the overall results.

A total of 324,005 comparatives and 141,641 superlatives were retrieved using the search patterns in table 2. Adjective types whose analytic-to-synthetic or synthetic-to-analytic ratio is smaller than 0.005[12] (i.e. 1:200) were removed. This procedure removes adjectives that do not display S-A variation[13] and adjectives for which one variant is extremely rare. For example, the analytic-to-synthetic ratio of *fast* is 2:1,130 (= 0.0018). The analytic form is so rare that the two examples are possibly due to production errors or characteristic of one particular speaker's idiolect. If *fast* were retained, the 1,130 synthetic examples would have to be included, possibly inflating the synthetic form in the dataset. This step also substantially reduces the total amount of

[11] Hilpert (2008: 403–4) adopts a similar search method to retrieve examples.
[12] The threshold is kept relatively small because S-A variation is generally strongly biased towards one form. It is safer to choose a small value to avoid the elimination of relevant adjectives from the dataset.
[13] Thus adjectives such as *more efficient* and *more effective*, which only accept the analytic form, were excluded.

Downloaded from http:/www.cambridge.org/core. Chinese University of Hong Kong, on 27 Oct 2016 at 12:48:53, subject to the Cambridge Core terms of use, available at http:/www.cambridge.org/core/terms. http://dx.doi.org/10.1017/S1360674316000368

Table 3. *Statistics of the two databases*

Properties	Comparatives (CompDB)	Superlatives (SuplDB)
1. Number of tokens	22,320	9,469
2. % of synthetic tokens	85.8%	73.0%
3. Number of positive adjective types	308	161
4. Adjective types favouring synthetic form	200 (64.9%)	113 (70.2%)
5. % of mono-, di-, polysyllabic base adjectives (by token)	49.2% / 50.1% / 0.6%	38.6% / 58.7% / 2.7%

examples to be processed. In a final step, irrelevant examples were manually discarded, e.g. mis-taggings (17), determiner uses of *more* (18) and adverbial uses (19).

(16) What meanest thou by this word Sacrament? (BNC: AC7-456)
(17) This makes sense in Japan where (because of stellar house prices) most young single professionals still live at home with their parents and have cash to spend, ... (BNC: ABG-2192)
(18) Somehow the message isn't getting through – or worse, it's getting twisted. (BNC: G35-1788)

It should be noted that nominal uses of comparatives and superlatives are included, e.g. *the funnier of the two* (Mondorf 2009: 12) and *the coldest of stares* (BNC: HGE-31). However, we discarded examples of elatives, e.g. *a most lovely situation*. Such intensification use of superlatives is functionally different from the superlative use. Further, elatives may skew the count of analytic superlative because the *most* variant is always used in such use (Mondorf 2009: 25).

The collected examples are stored in two databases, namely, CompDB (comparatives) and SuplDB (superlatives). The basic statistics of the databases are provided in table 3. A total of 22,320 tokens (308 types) of comparatives and 9,469 tokens (162 types) of superlatives were collected. Comparative examples are more than twice as frequent as superlative examples, which is in line with Claridge (2007).[14] The synthetic form accounts for 85.82 per cent of the comparative data and 72.56 per cent of the superlative data.[15] Although the synthetic variety dominates both databases, the bias is stronger in the comparative dataset.

4.2 Data annotation

Each example in the databases is annotated with respect to a set of predictor variables. Table 4 shows the annotation of examples (20) and (21).

[14] Claridge's (2007) calculation is based on the spoken-demographic part of the BNC only.
[15] The synthetic bias of comparatives is weaker than the bias reported by Hilpert (2008). He recorded that 89.7 per cent of the comparatives collected are synthetic. One possible reason is that we removed adjective types where one variant is extremely rare.

Downloaded from http:/www.cambridge.org/core. Chinese University of Hong Kong, on 27 Oct 2016 at 12:48:53, subject to the Cambridge Core terms of use, available at http:/www.cambridge.org/core/terms. http://dx.doi.org/10.1017/S1360674316000368

Table 4. *Two sample entries from the databases ('N/A' means the variable is not applicable)*

	Fields	Example 20 (CompDB)	Example 21 (SuplDB)
1	`Pos_Adjective`	'busy'	'costly'
2	`SA_Form`	S (synthetic)	A (analytic)
3	`SyllNum`	2 (disyllabic)	2 (disyllabic)
4	`FS_l`	N (no final /l/)	N (no final /l/)
5	`FS_ly`	N (no final *-ly*)	Y (with final *-ly*)
6	`FS_r`	N (no final /ɹ/)	N (no final /ɹ/)
7	`FS_y`	Y (with final *-y*)	N (no final *-y*)
8	`FS_s_st_sh`	N (no final sibilant)	N (no final sibilant)
9	`Cons_Cluster`	N (no final cons. cluster)	N (no final cons. cluster)
10	`Final_Stress`	N (no final stress)	N (no final stress)
11	`Position`	P (predicative)	A (attributive)
12	`Premodification`	Y (with premodification)	N (without premodification)
13	`Complement`	Z (no complement)	Z (no complement)
14	`Than_Phrase`	N (no *than* phrase)	N/A
15	`Pos_Freq`	4,705	1,118
16	a. `CompPos_Ratio`	0.017853	N/A
	b. `SuplPos_Ratio`	N/A	0.040250

(19) In later years, after the oil fields were established, he was to be even busier when a helicopter pad was built in Unst. (BNC: H0C-1132)

(20) At a cost of at least $1,900 million, the implementation of UNTAC would be the largest and most costly operation in the UN's history. (BNC: HLG-992)

4.3 Logistic regression

Logistic regression is a statistical model that predicts a binary classification using a set of categorical and/or continuous predictor variables (see Burns & Burns 2008; Field & Miles 2010). SAS Studio Release 3.5[16] was used to construct a regression model for comparatives and another one for superlatives. The predictor variables described in section 3 serve as predictors in the regression model for predicting the choice between synthetic and analytic forms.

When a logistic regression model is constructed, it is important to ensure that it is free of multicollinearity.[17] Multicollinearity occurs when at least one predictor variable highly correlates with another predictor variable. In other words, the value of one variable can be predicted from the other with a very high degree of accuracy. Strong multicollinearity renders the estimation of the explanatory power of predictor variables unreliable (see Baayen 2008: 181). One way of diagnosing multicollinearity is to refer

[16] SAS Studio website: www.sas.com/en_us/software/university-edition.html

[17] We thank an anonymous reviewer for highlighting the potential problem of multicollinearity in regression analysis.

Downloaded from http:/www.cambridge.org/core. Chinese University of Hong Kong, on 27 Oct 2016 at 12:48:53, subject to the Cambridge Core terms of use, available at http:/www.cambridge.org/core/terms. http://dx.doi.org/10.1017/S1360674316000368

Table 5. *Variables entered into the regression models*

		Comparative model	Superlative model	Hilpert's model
Dependent variables				
1	SA_Form	✔	✔	✔
Predictor variables				
2	SyllNum	✔	✔	✔
3	FS_l	✔	✔	✔
4	FS_ly	✔	✔	✔
5	FS_r	✔	✔	✔
6	FS_y	✔	✔	✔
7	FS_s_st_sh	✔	✔	×
8	Cons_Cluster	✔	✔	✔
9	Position (A/P)	✔	✔	✔
	Position (N/Z)	✔	✔	×
10	Premodification	✔	✔	✔
11	Complement (T/Z)	✔	✔	✔
	Complement (O/P)	✔	✔	×
12	Than_Phrase	✔	×	✔
13	Pos_Freq	✔	✔	✔
14	a. CompPos_Ratio	✔	–	✔
	b. SuplPos_Ratio	–	✔	–

to the 'Variance Inflation Factors' (VIF) of variables (see Allison 2012). A VIF of 5 or greater indicates a reason to be concerned about multicollinearity. Before the regression analysis is performed, the VIFs are computed for each predictor variable in the two models. Two pairs of variables are found to be highly correlated, namely, (i) `Final_Stress` and `FS_y` (comparative model) and (ii) `Final_Stress` and `FS_ly` (superlative model). A possible way to resolve the multicollinearity problem is to discard a correlated variable. We decided to remove `Final_Stress` from both models. After that, multicollinearity disappears.

4.4 Predictor variables

The variables entered into the models are given in table 5.

4.5 Statistical tests for logistic regression models

Several sets of parameters are used to evaluate the logistic regression model and the contribution of predictors in the model.

4.5.1 Goodness of fit

The goodness of fit of a logistic regression model assesses whether the model is a better fit to the data over the intercept-only model (i.e. a model without any predictor

Downloaded from http:/www.cambridge.org/core. Chinese University of Hong Kong, on 27 Oct 2016 at 12:48:53, subject to the Cambridge Core terms of use, available at http:/www.cambridge.org/core/terms. http://dx.doi.org/10.1017/S1360674316000368

variables). If the logistic regression model is significantly different from the null hypothesis, it means at least one predictor in the model contributes to the improvement of the fit of the model. Three chi-square tests (Likelihood Ratio, Score and Wald) are used to indicate whether the regression model differs significantly from the null hypothesis.

4.5.2 Pseudo R^2 statistics

According to Allison (2012: 68), it is not 'uncommon to have a model that fits well, as judged by … goodness-of-fit statistics, yet has very low predictive power'. To test whether this is the case, 'R-Square' and 'Max-rescaled R-Square' are also included in our report. This class of statistics assesses how well the model predicts the dependent variable based on the predictor variables.

4.5.3 Classification accuracy

Another important metric of the model is its classification accuracy. It informs us how well the model predicts the outcome of the dependent variable using sample observations. The evaluation procedure uses the model coefficients to predict values. Afterwards, the predicted target variable is compared with the observed values for each observation.

4.6 Statistical tests for contribution of predictors

4.6.1 Effect of individual predictors

The Wald chi-square statistic tests the null hypothesis that an individual predictor's regression coefficient is zero (i.e. no contribution), provided that the other predictor variables are in the model. If the predictor makes a statistically significant contribution to the alternation, one expects its *p*-value to be smaller than 0.05 (i.e. the null hypothesis is very unlikely). Note, however, that one cannot assess the relative magnitude of the contribution of predictors based on the *p*-value of the Wald statistic (Allison 2012).

4.6.2 Relative contribution of individual predictors

The relative contribution of predictors can be compared with reference to their standardized coefficients, which measure 'how many standard deviations the dependent variable *y* changes for a 1-standard deviation increase in each of the *x* variables' (Allison 2012: 89). These coefficients serve as a metric for comparing the relative importance of individual predictor variables in the logistic regression model (see Menard 2011; Tonidandel & LeBreton 2011; Allison 2012: 89), rather than comparing them at the variable group level (see section 4.6.3). The coefficients also allow comparison regardless of variable type (i.e. continuous vs categorical).

4.6.3 Contribution of predictors in groups

To evaluate the relative importance of predictors as a group (i.e. phonological vs syntactic vs frequency), Hilpert constructed three additional logistic regression models, each of which leaves out one group of predictors, i.e.:

Downloaded from http:/www.cambridge.org/core. Chinese University of Hong Kong, on 27 Oct 2016 at 12:48:53, subject to the Cambridge Core terms of use, available at http:/www.cambridge.org/core/terms. http://dx.doi.org/10.1017/S1360674316000368

- all variables except phonological ones,
- all variables except syntactic ones, and
- all variables except frequency ones.

By observing the degradation of the accuracy after a variable group is removed, one can estimate the relative impact of the respective group. This 'variable-group removal' method will also be used in our study.

5 Results

5.1 Comparatives

5.1.1 Evaluation of overall model

Table 6 summarizes the results of the full regression model fitted to comparatives.

The goodness of fit tests above shows that the comparative model with the predictor variables fit the data better than the model without the variables. The *p*-values of the three chi-square tests (Likelihood Ratio, Score and Wald) are all less than 0.0001. The R-Square (0.222) and Max-rescaled R-Square (0.397) suggest that the predictor variables are moderately good predictors of the dependent variables. The model improves the classification accuracy from 85.8 to 88.7 per cent.

5.1.2 Evaluation of contribution of predictors

In table 7, each predictor variable demonstrates an effect on the S-A variation with *p*-value smaller than 0.0001. For comparison, the last column shows Hilpert's (2008: 407) findings of the comparative form that the variables favour. Among the 17 predictors, 13 of them are significant but `FS_ly`, `Premodification`, `Complement(O)` and `Than_clause` are non-significant.

For convenience of comparing their relative importance, table 8 ranks predictors in descending order of the absolute value of standardized coefficients. Phonological and frequency variables are generally ranked high, whereas syntactic ones are low on

Table 6. *Assessment of logistic regression model of comparatives*

A. Goodness of fit	Chi-square	DF	*p* value
i. Likelihood ratio	5591.230	17	<.0001
ii. Score	5141.016	17	<.0001
iii. Wald	3080.824	17	<.0001
B. Pseudo R-squares			
i. R-square	0.222		
ii. Max-rescaled R-square	0.397		
C. Classification accuracy	88.7%		

Downloaded from http:/www.cambridge.org/core. Chinese University of Hong Kong, on 27 Oct 2016 at 12:48:53, subject to the Cambridge Core terms of use, available at http:/www.cambridge.org/core/terms. http://dx.doi.org/10.1017/S1360674316000368

Table 7. *Assessment of individual predictors' contribution to the model of comparatives. Positive coefficients (or '+ve') indicate a preference for analytic more; negative coefficients (or '–ve') indicate a preference for synthetic -er. # indicates that the significance of the predictor in our model is different from that in Hilpert's model.*

Predictor		df	Coeff.	Standard error	Wald chi-square	P	Standardized coefficient	Hilpert's findings
Intercept		1	−4.176	0.139	897.864	<.0001		
SyllNum		1	2.470	0.064	1508.892	<.0001	0.711	+ve
FS_l		1	2.208	0.113	380.321	<.0001	0.252	+ve
FS_ly		1	0.142	0.104	1.867	non-sig.	0.013	+ve#
FS_r		1	2.065	0.078	707.009	<.0001	0.337	+ve
FS_y		1	−1.372	0.075	330.629	<.0001	−0.361	−ve
FS_s_st_sh		1	1.286	0.106	145.974	<.0001	0.123	unavailable
Cons_Cluster		1	0.284	0.070	16.214	<.0001	0.055	+ve
Position	A	1	−0.339	0.105	10.519	<.01	−0.092	−ve
Position	N	1	0.617	0.152	16.435	<.0001	0.053	unavailable
Position	P	1	0.240	0.103	5.402	<.05	0.066	+ve
Premodification		1	0.032	0.057	0.307	non-sig.	0.007	+ve#
Complement	O	1	0.151	0.263	0.330	non-sig.	0.011	unavailable
Complement	P	1	0.620	0.092	45.154	<.0001	0.081	unavailable
Complement	T	1	0.687	0.125	30.418	<.0001	0.115	+ve
Than_clause		1	−0.108	0.059	3.333	non-sig.	−0.023	−ve#
Pos_Freq		1	−0.000	0.000	715.464	<.0001	−1.080	−ve
CompPos_Ratio		1	−11.802	0.544	471.401	<.0001	−0.713	−ve

Table 8. *Relative importance of predictors (in descending order of the absolute value of standardized coefficients)*

Predictor		p	Standardized coefficients
Pos_Freq		<.0001	−1.080
CompPos_Ratio		<.0001	−0.713
SyllNum		<.0001	0.711
FS_y		<.0001	−0.361
FS_r		<.0001	0.337
FS_l		<.0001	0.252
FS_s_st_sh		<.0001	0.123
Complement	T	<.0001	0.115
Position	A	<.01	−0.092
Complement	P	<.0001	0.081
Position	P	<.05	0.066
Cons_Cluster		<.0001	0.055
Position	N	<.0001	0.053

Downloaded from http:/www.cambridge.org/core. Chinese University of Hong Kong, on 27 Oct 2016 at 12:48:53, subject to the Cambridge Core terms of use, available at http:/www.cambridge.org/core/terms. http://dx.doi.org/10.1017/S1360674316000368

Table 9. *Comparison of contribution of variable groups in the comparative model*

Logistic regression model	Classification accuracy
A. All variables	88.7%
B. All variables except phonological ones	85.8%
C. All variables except syntactic ones	89.0%
D. All variables except frequency ones	88.2%
E. Baseline[18](no variables)	85.8%

the list. The standardized coefficients allow us to derive the following ranking of the relative strength of individual predictors.

(21) A. Phonological variables
`SyllNum > FS_y > FS_r > FS_l > FS_s_st_sh > Cons_Cluster`
B. Syntactic variables
`Complement(T)>Position(A)>Complement(P)>Position(P)>Position(N)`
C. Frequency variables
`Pos_Freq > CompPos_Ratio`

Frequency and phonological variables are generally stronger predictors. Even though phonological variables are generally influential, `Cons_Cluster` is weaker than the others in the group. The impact of `Position(P)` and `Position(N)` is relatively small.

Table 9 shows the results of the contribution of variable groups using the 'variable-group removal' method. With all the predictors, the model has a classification accuracy of 89.0 per cent. This percentage should be compared to the rate of correct predictions achieved by always guessing the most frequent outcome (see model E), which is 85.8 per cent.

The phonological group exerts the greatest influence. The classification accuracy decreases most (–2.9%) when phonological variables are discarded, i.e. in model B. Frequency variables are fairly useful (–0.5%). Surprisingly, the inclusion of syntactic variables actually has a slight adversary effect on the classification (compare models A and C).

Last, let us compare our model with Hilpert's (2008) model. `FS_s_st_sh`, `Position(N)` and `Complement(O/P)` are not investigated in Hilpert's regression model. Our model largely replicates Hilpert's model in terms of the significance and the S-A preference of predictors. The major difference is that `FS_ly`,[19] `Premodication`, `Complement(O)` and `Than_clause` are significant predictors in Hilpert's model but the variable is not significant in our model.

[18] Even if all cases are always predicted to be rendered synthetically without considering the predictors, one can achieve this 85.8 per cent baseline accuracy.

[19] One possibility of the divergent results of FS_ly is that the very high-frequency adjective *likely* (>3,600 comparative tokens) is included in Hilpert's (2008: 414) dataset but not in our dataset due to its strong analytic bias.

Downloaded from http:/www.cambridge.org/core. Chinese University of Hong Kong, on 27 Oct 2016 at 12:48:53, subject to the Cambridge Core terms of use, available at http:/www.cambridge.org/core/terms. http://dx.doi.org/10.1017/S1360674316000368

Table 10. *Assessment of logistic regression model of superlatives*

A. Goodness of fit	Chi-square	DF	*p* value
i. Likelihood ratio	4941.712	16	<.0001
ii. Score	4475.766	16	<.0001
iii. Wald	2377.584	16	<.0001
B. Pseudo R-squares			
i. R-square	0.407		
ii. Max-rescaled R-square	0.590		
C. *Classification accuracy*	88.3%		

5.2 Superlatives

5.2.1 Evaluation of overall model

The assessment of the regression model for superlatives is given in table 10.

The *p*-values of the three chi-square tests (Likelihood Ratio, Score and Wald) for the goodness of fit of the regression models are all less than 0.0001. Again, the regression model with the predictor variables fit the data better than the model without the variables. The R-square (0.407) and Max-rescaled R-square (0.590) also suggest that the predictor variables are strong predictors of the dependent variables. The model improves the classification accuracy from 73.0 to 88.3 per cent.

5.2.2 Evaluation of contribution of predictors

Among the 16 predictor variables, 11 of them have a *p*-value smaller than 0.05. `FS_ly`, `Position(A)`, `Complement(O)`, `Complement(P)`, and `SuplPos_Ratio` do not have an effect on the choice of the superlative forms, as shown in table 11.

As in the comparative model, phonological variables are generally ranked high in the superlative model. Syntactic factors are found in the upper and lower half of the list. However, the frequency variable `POS_Freq` is pretty low in the superlative model but is very high in the comparative model (see table 12). The relative strength of individual predictors is ranked in each variable group in (23). `SyllNum` and `FS_y` emerge as more important variables in the phonological variable group, which is similar to what we found in the comparative model. The contribution of syntactic predictors in the comparative and superlative models is quite different. Important syntactic predictors in the superlative model, e.g. `Position(P)` and `Position(N)`, are pretty low in the comparative model. The fact that `SuplPos_Ratio` is non-significant is unexpected considering the high ranking of `CompPos_Ratio`.

(22) A. Phonological variables
`SyllNum > FS_y > Cons_Cluster > FS_r > FS_l > FS_s_st_sh`
B. Syntactic variables
`Position(P) > Position(N) > Premodification > Complement(T)`
C. Frequency variables
`Pos_Freq`

Downloaded from http:/www.cambridge.org/core. Chinese University of Hong Kong, on 27 Oct 2016 at 12:48:53, subject to the Cambridge Core terms of use, available at http:/www.cambridge.org/core/terms. http://dx.doi.org/10.1017/S1360674316000368

Table 11. *Assessment of individual predictors' contribution to the model fitted to superlatives*

Predictor		df	Coeff.	Standard error	Wald chi-square	p	Standardized coefficient
Intercept		1	− 6.018	0.209	832.039	<.0001	
SyllNum		1	3.135	0.102	948.909	<.0001	0.921
FS_l		1	1.522	0.252	36.465	<.0001	0.090
FS_ly		1	− 0.104	0.115	0.820	non-sig.	− 0.022
FS_r		1	− 0.591	0.114	26.915	<.0001	− 0.126
FS_y		1	− 3.481	0.124	788.283	<.0001	− 0.768
FS_s_st_sh		1	0.780	0.196	15.843	<.0001	0.071
Cons_Cluster		1	− 2.378	0.123	373.474	<.0001	− 0.528
Position	A	1	0.036	0.106	0.114	non-sig.	0.009
Position	N	1	2.116	0.287	54.337	<.0001	0.177
Position	P	1	1.518	0.151	100.662	<.0001	0.315
Premodification		1	− 2.113	0.793	7.111	<.01	− 0.097
Complement	O	1	0.423	0.368	1.320	non-sig.	0.025
Complement	P	1	− 0.086	0.179	0.231	non-sig.	− 0.010
Complement	T	1	0.531	0.170	9.775	<.01	0.085
Pos_Freq		1	0.000	0.000	15.449	<.0001	0.089
SuplPos_Ratio		1	− 0.617	0.877	0.494	non-sig.	− 0.019

Table 12. *Relative importance of predictors (in descending order of the absolute value of standardized coefficients)*

Predictor		p	Standardized coefficients
SyllNum		<.0001	0.921
FS_y		<.0001	− 0.768
Cons_Cluster		<.0001	− 0.528
Position	P	<.0001	0.315
Position	N	<.0001	0.177
FS_r		<.0001	− 0.126
Premodification		<.01	− 0.097
FS_l		<.0001	0.090
Pos_Freq		<.0001	0.089
Complement	T	<.01	0.085
FS_s_st_sh		<.0001	0.071

As shown in table 13, the predictors in the superlative model significantly improve the accuracy from 73.0 per cent (model E) to 88.3 per cent (model A). Discarding the phonological group has led to the greatest degradation in classification accuracy (model B), followed by the frequency group (model D) and the syntactic group (model C).

Downloaded from http:/www.cambridge.org/core. Chinese University of Hong Kong, on 27 Oct 2016 at 12:48:53, subject to the Cambridge Core terms of use, available at http:/www.cambridge.org/core/terms. http://dx.doi.org/10.1017/S1360674316000368

Table 13. *Comparison of contribution of variable groups in the superlative model*

Logistic regression model	Classification accuracy
A. All variables	88.3%
B. All variables except phonological ones	80.1%
C. All variables except syntactic ones	87.4%
D. All variables except frequency ones	87.2%
E. Baseline (no variables)	73.0%

6 Discussion

6.1 Identification of predictors

We begin to address the first research question by identifying the predictors that are statistically significant in the two models based on sections 5.1.2 and 5.2.2.

Comparatives (13 predictors)
- `SyllNum, FS_y, FS_r, FS_l, FS_s_st_sh, Cons_Cluster`
- `Complement(T), Position(A), Complement(P), Position(P), Position(N)`
- `Pos_Freq, CompPos_Ratio`

Superlatives (11 predictors)
- `SyllNum, FS_y, Cons_Cluster, FS_r, FS_l, FS_s_st_sh`
- `Position(P) > Position(N) > Premodification > Complement(T)`
- `Pos_Freq`

(Note: underlined predictors are significant only in one model.)

The set of significant predictors in the two models consists of the following ten variables: `SyllNum, FS_y, FS_r, FS_l, FS_s_st_sh, Cons_Cluster, Complement(T), Position(P), Position(N)` and `Pos_Freq`. The phonological variables mostly have an effect in both models. In contrast, only three out of seven syntactic variables are significant predictors in both models.

6.2 Contribution of predictors

This section addresses the second research question about the relative contribution among the predictors. The standardized coefficient offers a measure to compare the contribution of individual predictors. Here are the top five predictors of the two models from tables 8 and 12.

Comparatives (top five in descending order)
- `Pos_Freq > CompPos_Ratio > SyllNum > FS_y > FS_r`

Superlatives (top five in descending order)
- `SyllNum > FS_y > Cons_cluster > Position(P) > Position(N)`

Downloaded from http:/www.cambridge.org/core. Chinese University of Hong Kong, on 27 Oct 2016 at 12:48:53, subject to the Cambridge Core terms of use, available at http:/www.cambridge.org/core/terms. http://dx.doi.org/10.1017/S1360674316000368

A more detailed discussion and comparison of important predictors can be found in the next section.

6.3 *Comparative vs superlative*

The crucial question of our study is how far S-A variation differs between comparatives and superlatives. The findings so far suggest that the two share some similarities but also display obvious differences.

6.3.1 *Degree of syntheticity bias*

The syntheticity bias has some significant implications for the role of the predictors in the models. The use of the same methodology in both models enables us to compare their degree of syntheticity bias. Among the tokens collected, 85.8 per cent in CompDB are synthetic, which contrasts with only 73.0 per cent in SuplDB. Thus, our parallel analysis of both the comparative and the superlative alternation empirically validates the claim that the synthetic bias is stronger in comparatives than in superlatives. The discrepancy has some implications for regression models. A stronger bias entails a higher baseline accuracy. Improvement of prediction accuracy becomes more difficult. Despite the inclusion of 17 predictors and the larger dataset in the comparative model, classification accuracy only raises from 85.8 to 88.7 per cent. The remaining misclassified cases are possibly those involving idiolectal or lexical preferences. In contrast, with a lesser syntheticity bias, the predictors in the superlative model play a greater role in improving the accuracy from 73.0 to 88.3 per cent.

6.3.2 *Predictor contribution*

Variable group level. The impact of the three variable groups looks very similar in the comparative and superlative models at the variable group level using the variable-group removal method. The influence of the three groups in both models can be summarized by the hierarchy in (24) (see also tables 9 and 13).

(23) Phonological group > Frequency group > Syntactic group

The phonological group is consistently more important than the other two, as the removal of the phonological group has led to a larger reduction in classification accuracy. The frequency group and syntactic group are weaker with the former being slightly stronger than the latter. Hilpert (2008) comes to a similar conclusion about the hierarchy in (24) for S-A variation in comparatives.

Readers may note that frequency variables rank even higher than phonological variables with respect to standardized coefficients in comparatives, which seems to contradict the hierarchy in (24). One explanation is that even though frequency variables tend to score high on the standardized coefficient in the comparative model, the frequency group has only two variables. In contrast, the phonological group has seven variables, possibly making the overall contribution of the phonological group much more prominent in the hierarchy.

Downloaded from http:/www.cambridge.org/core. Chinese University of Hong Kong, on 27 Oct 2016 at 12:48:53, subject to the Cambridge Core terms of use, available at http:/www.cambridge.org/core/terms. http://dx.doi.org/10.1017/S1360674316000368

Individual predictor level. The standardized coefficient offers an additional measure to compare the contribution at the individual predictor level. Here are the top five and bottom five predictors from the two models (see tables 8 and 12).

Comparatives (in descending order)

- Top five: `Pos_Freq > CompPos_Ratio > SyllNum > FS_y > FS_r`
- Bottom five: `Position(A) > Complement(P) > Position(P) > Cons_Cluster > Position(N)`

Superlatives (in descending order)

- Top five: `SyllNum > FS_y > Cons_cluster > Position(P) > Position(N)`
- Bottom five: `Premodification > FS_l > Pos_Freq > Complement(T) > FS_s_st_sh`

More phonological predictors are found among the top five predictors. `SyllNum` and `FS_y` outrank most other predictors on both models. Some divergences at the predictor level stand out in the results. One notable difference between the models is that although many syntactic variables are weak or non-significant predictors in comparative choice, syntactic variables seem to contribute more strongly in the superlative model, e.g. `Position(N/P)`. While `CompPos_Ratio` contributes much to the comparative model, the influence of its counterpart `SuplPos_Ratio` is negligible in the superlative model. Moreover, even though frequency variables are important predictors in both models, higher frequency favours syntheticity in comparatives but analyticity in superlatives. These discrepancies call for further scrutiny in future research. The empirical findings show that claims about comparative variation do not carry over to superlative variation.

6.4 *More support for* more-*support?*

Recall that Mondorf (2003, 2009) proposes the processing view of *more*-support to explain the S-A variation of comparatives. In essence, cognitively more complex or difficult structures tend to align with the analytic variant. The question arises whether analytic forms also serve as a support strategy with superlatives, i.e. whether Mondorf's claims can be extended to superlatives. To examine how Mondorf's (2009: 197–9) prediction fares in our two models, we have listed the predicted preferences and the actual preferences in table 14. Predictors about which Mondorf has made explicit claims are included and their equivalents in our models are shown below.

Mondorf's predictions are borne out much more accurately in the comparative model. In the superlative model, five predictors clearly support her predictions that analytic variants act as a support strategy mitigating the processing load. However, the observed preferences of `Pos_Freq` and `Cons_Cluster` are opposite to those predicted by *more*-support. Also, three syntactic variables have no significant influence on S-A choice. It would be interesting to scrutinize in the future whether and why '*most*-support' does not seem to fare as well as *more*-support.

Downloaded from http:/www.cambridge.org/core. Chinese University of Hong Kong, on 27 Oct 2016 at 12:48:53, subject to the Cambridge Core terms of use, available at http:/www.cambridge.org/core/terms. http://dx.doi.org/10.1017/S1360674316000368

Table 14. *Predicted and actual preference of predictors (A = analytic, S = synthetic)*

	Comparative			Superlative		
Predictor	Sig.	Actual preference according to our model	Predicted preference according to *more*-support	Sig.	Actual preference according to our model	Predicted preference according to *more*-support
`SyllNum`	*	A	A	*	A	A
`FS_r`	*	A	A	not applicable		
`FS_s_st_sh`	not applicable			*	A	A
`Cons_Cluster`	*	A	A	*	A	S
`Position(A)`	*	S	S	non sig.	–	S
`Position(N)`	*	A	A	*	A	A
`Position(P)`	*	A	A	*	A	A
`Complement(O)`	non-sig.	–	A	non-sig.	–	A
`Complement(P)`	*	A	A	non-sig.	–	A
`Complement(T)`	*	A	A	*	A	A
`Pos_Freq`	*	S	S	*	A	S

6.5 Context-(in)sensitivity of predictors

The regression results may imply that one should focus on phonological and frequency variables in the investigation of the S-A alternation of comparatives and superlatives. After all, syntactic variables contribute relatively little. This tendency is even stronger in the comparative model. However, it is still important not to neglect the syntactic variables even if their contribution is relatively small. Why? Recall that the goal of our research is to understand under which conditions a comparative/superlative is realized synthetically or analytically. For example, 88 per cent of the superlatives of *healthy* in SuplDB are rendered synthetically and the remainder analytically. Ideally, the regression model can inform us about the likelihood of each form, such as in (23) and (24).

(24) Taking regular exercise is one of the healthiest things that women can do. (A0J-1692)
(25) Today's population in their 20's is the most healthy, […] (CC3-30)

Most phonological variables, however, are not able to do this as they are based on the inherent properties of adjectives, i.e. they are context-insensitive. The values for the phonological and frequency variables, such as `SyllNum`, `FS_y` and `Pos_Freq`, are the same for (23), (24) and all other examples involving *healthy* in SuplDB. In contrast, syntactic variables are context-sensitive and thus have the potential to unveil whether there is a relationship between S-A choice and context. Unless we believe that S-A variation is independent of context, syntactic variables should remain a crucial issue of the research on S-A variation. Here we want to highlight one syntactic variable which deserves more attention in future research: Leech & Culpeper (1997) as well as

Downloaded from http:/www.cambridge.org/core. Chinese University of Hong Kong, on 27 Oct 2016 at 12:48:53, subject to the Cambridge Core terms of use, available at http:/www.cambridge.org/core/terms. http://dx.doi.org/10.1017/S1360674316000368

Lindquist (2000) claim that coordinated comparatives or superlatives tend to promote matching S-A forms in both conjuncts, which, in turn, affects the likelihood of how comparatives/superlatives are rendered.[20] In our datasets, coordinated comparatives account for 11.8 per cent of CompDB, whereas coordinated superlatives yield 8.3 per cent of SuplDB. If the conjunct parallelism claim is correct, it can potentially explain some of the patterns we cannot account for now. However, due to the complexity of data annotation, the issue has not been taken up in the present study.

7 Conclusion

Logistic regression analysis enables us to assess and compare the simultaneous contribution of a set of predictors towards the S-A variation. Applying the same corpus methodology, we have constructed two regression models to analyse the S-A alternation of comparatives and superlatives. Our methodology is similar to that of Hilpert (2008) but our study of S-A alternation covers not only comparatives but also superlatives. We have also included some predictors not considered in Hilpert's (2008) study which have been reported to be relevant elsewhere, such as `FS_s_st_sh`, `Position(N)` and `Complement(P)`; and they turn out to be helpful to some extent in explaining some of the variation. On the whole, the syntheticity bias is stronger in comparatives than in superlatives.

Based on the classification accuracy of the models, we find that the predictors of superlatives contribute more notably over the baseline accuracy in predicting the S-A choice than predictors of comparatives. Thirteen out of 17 predictors have a significant influence on comparative variation, whereas only 11 out of 16 predictors have a significant influence on superlative choice. Phonological variables generally contribute most to the variation, followed by frequency variables and syntactic variables. In both models, `SyllNum` and `FS_y` consistently outrank other predictors. Notable differences between comparatives and superlatives have been detected: even though some syntactic variables, e.g. `Premodification` and `Position(N)`, are non-significant or weak for comparatives, they contribute much more in the superlative model.

For comparatives, the distribution of synthetic vs analytic variants is in line with Rohdenburg's Complexity Principle (1996) and analytic support (Mondorf 2014), which predict that the more explicit variant tends to be preferred in cognitively complex environments. For superlatives, analytic variants also served as a support strategy with most factors, but more frequent adjectives triggered syntheticity in comparatives and analyticity in superlatives, a finding that calls for further research on superlative alternation.

[20] The idea is supported by Szmrecsanyi (2005), who finds that speakers tend to reuse forms recently used or heard. If this is correct, the S-A choice of the comparative/superlative in the second conjunct should be affected by the choice in the first conjunct of coordinated comparatives and superlatives.

Downloaded from http:/www.cambridge.org/core. Chinese University of Hong Kong, on 27 Oct 2016 at 12:48:53, subject to the Cambridge Core terms of use, available at http:/www.cambridge.org/core/terms. http://dx.doi.org/10.1017/S1360674316000368

Authors' addresses:

Department of Linguistics and Modern Languages
G/F, Leung Kau Kui Building
The Chinese University of Hong Kong
Shatin
Hong Kong
yllcheung@cuhk.edu.hk
vincentzlt@gmail.com

References

Allison, Paul D. 2012. *Logistic regression using SAS: Theory and application*. Cary, NC: SAS Institute.

Aston, Guy & Lou Burnard. 1997. *The BNC handbook: Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Baayen, R. Harald. 2008. *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman grammar of spoken and written English*. London and New York: Longman.

Burnard, Lou (ed.). 2007. *Reference guide for the British National Corpus (XML edition)*. Research Technologies Service at Oxford University Computing Services. www.natcorp.ox.ac.uk/docs/URG

Burns, Robert P. & Richard Burns. 2008. Logistic regression. In *Business research methods and statistics using SPSS*, chapter 24. Thousand Oaks, CA: Sage. (Extra online chapter: www.uk.sagepub.com/burns/chapters.htm)

Claridge, Claudia. 2007. The superlative in spoken English. *Language and Computers* 62, 128–67.

Field, Andy & Jeremy Miles. 2010. *Discovering statistics using SAS*. Los Angeles: Sage.

González-Díaz, Victorina. 2008. *English adjective comparison: A historical perspective*. Amsterdam: John Benjamins.

Görlach, Manfred. 1991. *Englishes: Studies in varieties of English, 1984–8* (Varieties of English around the World 9). Amsterdam: John Benjamins.

Hawkins, John A. 1994. *A performance theory of order and constituency*. Cambridge: Cambridge University Press.

Hilpert, Martin. 2008. The English comparative: Language structure and language use. *English Language and Linguistics* 12(3), 395–417.

Huddleston, Rodney & Geoffrey K. Pullum. 2002. *The Cambridge grammar of English*. Cambridge: 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 Nevalainen & Kahlas-Tarkka (eds.), 329–52.

Kytö, Merja & Suzanne Romaine. 2000. Adjective comparison and standardization processes in American and British English from 1620 to the present. In Laura Wright (ed.), *The development of Standard English 1300—1800: Theories, descriptions, conflicts*, 171–94. Cambridge: Cambridge University Press.

Leech, Geoffrey & Jonathan Culpeper. 1997. The comparison of adjectives in recent British English. In Nevalainen & Kahlas-Tarkka (eds.), 353–73.

Downloaded from http:/www.cambridge.org/core. Chinese University of Hong Kong, on 27 Oct 2016 at 12:48:53, subject to the Cambridge Core terms of use, available at http:/www.cambridge.org/core/terms. http://dx.doi.org/10.1017/S1360674316000368

Lindquist, Hans. 1998. The comparison of English disyllabic adjectives in -y and -ly in Present-day British and American English. In Hans 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. *Language and Computers* 30, 125–32.
Menard, Scott. 2011. Standards for standardized logistic regression coefficients. *Social Forces* 89(4), 1409–28.
Mondorf, Britta. 2003. Support for *more*-support. In Rohdenburg & Mondorf (eds.), 251–304.
Mondorf, Britta. 2009. *More support for* more-*support: The role of processing constraints on the choice between synthetic and analytic comparative forms*. Amsterdam: John Benjamins.
Mondorf, Britta. 2014. (Apparently) competing motivations in morpho-syntactic variation. In Brian MacWhinney, Andrej Malchukov & Edith Moravcsik (eds.), *Competing motivations in grammar and usage*, 209–28. Oxford: Oxford University Press.
Nevalainen, Terrtu & Leena Kahlas-Tarkka (eds.). 1997. *To explain the present: Studies in the changing English language in honour of Matti Rissanen*. Helsinki: Société Néophilologique.
Plag, Ingo. 1998. Morphological haplology in a constraint-based morpho-phonology. In Wolfgang Kehrein & Richard Wiese (eds.), *Phonology and morphology of the Germanic languages*, 199–215. Tübingen: Niemeyer.
Poutsma, Hendrik. 1914. *A grammar of late modern English*. Groningen: Noordhoff.
Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. 1985. *A comprehensive grammar of the English language*. London: Longman.
Rohdenburg, Günter. 1996. Cognitive complexity and increased grammatical explicitness in English. *Cognitive Linguistics* 7(2), 149–82.
Rohdenburg, Günter & Britta Mondorf (eds.). 2003. *Determinants of grammatical variation in English*. Berlin: Mouton de Gruyter.
Schlüter, Julia. 2005. *Rhythmic grammar: The influence of rhythm on grammatical variation and change in English*. Berlin: Mouton de Gruyter.
Sweet, Henry. [1891] 1968. *A new English grammar: Logical and historical*. Oxford: Clarendon Press.
Szmrecsanyi, Benedikt. 2005. Language users as creatures of habit: A corpus-linguistic analysis of persistence in spoken English. *Corpus Linguistics and Linguistic Theory* 1(1), 113–50.
Thomson, Audrey J. & Agnes V. Martinet. 1980. *A practical English grammar*. Oxford: Oxford University Press.
Tonidandel, Scott & James LeBreton. 2011. Relative importance analysis: A useful supplement to regression analysis. *Journal of Business and Psychology* 26(1), 1–9.

Downloaded from http:/www.cambridge.org/core. Chinese University of Hong Kong, on 27 Oct 2016 at 12:48:53, subject to the Cambridge Core terms of use, available at http:/www.cambridge.org/core/terms. http://dx.doi.org/10.1017/S1360674316000368