Lexis
Journal in English Lexicology

14 | 2019
Blending in English

Improving on observational blends research: regression modeling in the study of experimentally-elicited blends

Stefanie Wulff and Stefan Th. Gries

Electronic version
URL: http://journals.openedition.org/lexis/3625
ISSN: 1951-6215

Publisher
Université Jean Moulin - Lyon 3

Electronic reference
Stefanie Wulff and Stefan Th. Gries, « Improving on observational blends research: regression modeling in the study of experimentally-elicited blends », *Lexis* [Online], 14 | 2019, Online since 16 December 2019, connection on 16 December 2019. URL : http://journals.openedition.org/lexis/3625

Improving on observational blends research: regression modeling in the study of experimentally-elicited blends

Stefanie Wulff and Stefan Th. Gries

We thank Dylan Attal, Anna Bjorklund, Steven Critelli, Erica Drayer, Corinne Futch, Hali Lindsay, and Noah Rucker for their invaluable help with running the experiments, transcribing the recordings, and annotating the data.

0. Introduction

This study examines blending, that is "an intentional fusion of usually two words where a part of the first source word (sw_1) – usually the beginning of sw_1 – is combined with a part of the second source word (sw_2) – usually the end of sw_2 – where at least one source word is shortened and/or the fusion may involve overlap of sw_1 and sw_2" [Gries 2012: 146]. Blending presents a curious case of word formation as it does not appear to be rule-governed as other derivational processes – usually, blending involves a conscious effort that involves word play, which often violates rigid morphological rules; it is less productive, yet at the same time, arguably more creative than most other derivational processes; and while it is superficially quite similar to other intentional word formation processes such as compounding, clipping, abbreviations, and acronyms, blending has received far less attention, maybe because the intricate interplay between orthography and pronunciation at play in blending is not a centerpiece of linguistic theory (see Gries [2012: 145]).

Much of our previous knowledge of blends is based on observational data that, ultimately, may entail the risk of being opportunistic data samples. Therefore, in order to validate previous work based on these kinds of data, this paper compares results

from an experimentally solicited sample of blends to the results from previously-collected observational data by studying the following three hypotheses:

- Hypothesis 1: the shorter source word contributes more of itself to the blend;
- Hypothesis 2: sw_2 determines the stress of the blend (more than sw_1); and
- Hypothesis 3: blending maximizes similarity between source words and blends.

We first summarize previous research in particular with regard to the classificatory and descriptive questions they discussed and the kinds of data they used, before we turn to a description of our experiment and the data it provided (Section 2). Then, each of the three hypotheses is discussed (in Sections 3, 4, and 5 respectively) before we conclude (Section 6).

1. Previous research

Earlier research on blends focused mostly on classifying different types of blends, and how to distinguish blending from other word formation processes. One of the first such studies on blends is Pound [1914], who analyzed 314 blends. Pound defines a blend as "two or more words, often of cognate sense, telescoped as it were into one; as factitious conflations which retain, for a while at least, the suggestive power of their various elements" [Pound 1914: 1]. She argues that blends, which clearly fuse meanings and are consciously coined, should be considered distinct from analogical extensions and enlargements since these do not exhibit semantic fusion and are created unintentionally. Less clear to Pound is the distinction between blends and contractions [Pound 1914: 11]. Pound proposes qualitative labels such as literary coinages, speech errors, and conscious folk formations. In terms of structural analysis, however, Pound offers no groupings and even warns that "no very definite grouping seems advisable" since source words are combined in (apparently) unpredictable ways [1914: 22].

Algeo [1977: 48] defines blends as "a combination of two or more forms, at least one of which has been shortened in the process of combination." This definition is based solely on blend structure and does not account for troublesome cases like *meritocracy*, which Algeo [1977: 54] defines as a derivative that combines with the form *-ocracy*. However, under Algeo's own definition, *meritocracy* could be understood as a blend of *merit* and *aristocracy*. Algeo goes on to say that in some cases blends cannot be clearly distinguished from other derivational processes. For example, *breadth* could be understood as a blend (OE *brede* x *length*) but also as an analogical extension *long* → *length* : *broad* → *x* [1977: 51]. Differently from Pound [1914], Algeo offers two main classifications of blends. A first classification groups blends structurally into blends with phonemic overlap, blends with clipping, and blends with both phonemic overlap and clipping. A second classification distinguishes between syntagmatic blends, or telescope blends, of words that usually co-occur sequentially, like *radarange* (*radar* × *range*), and associative blends, which are blends whose source words are usually semantically linked in the coiner's mind.

Many researchers have articulated doubt that the formation of blends obeys any systematic rules – Bauer, for example, states that "in blending, the blender is apparently free to take as much or as little from either base as is felt to be necessary or desirable. [...] Exactly what the restrictions are, however, beyond pronounceability and spellability is far from clear" [Bauer 1983: 225]. Nevertheless, several studies have made an effort to identify the cognitive determinants of blend formation. For example,

Kelly's [1998] analysis of 426 blends supports the idea that blending is predictable by revealing the following systematicities:

- the first source words found in the blends are significantly shorter, significantly more frequent, and denote significantly more prototypical category members than the second source words;
- the breakpoints of blends occur significantly more at syllable/word breaks than elsewhere (e.g., *fool* × *philosopher* → *foolosopher*, *scum* × *company* → *scumpany*, or *sun* × *umbrella* → *sunbrella*); furthermore, within-syllable breaks usually preserved the rime (e.g., *breakfast* × *lunch* → *brunch*, *channel* × *tunnel* → *chunnel*, or *flight* × *plane*→ *flane*).

Gries [2004a] further examines the amount of information each source word contributes and the similarity of the source words to the blend. Building on Kaunisto [2000], Gries considers not only the graphemic, but also phonemic contributions of the source words to the blend, alongside length and medium as additional variables. The results of a loglinear analysis revealed that sw_2 tends to be longer and contribute more of itself than sw_1. Interestingly, his analysis reveals that the interaction between length and contribution is strong enough to not be affected by medium; that is, his results reveal a strong graphemic influence on blend formation, which is not observed in many other linguistic processes.

Gries [2004b] investigates the degree of recognizability of the blend and the similarity of the sws to the blend. Examining the stress patterns of 614 blends with up to four syllables, Gries finds that sw_2 plays a dominant role in determining the blend's stress pattern.

Gries [2012] distinguishes, if only heuristically, three stages of the blending process – (i) the selection of the source words to blend, (ii) the decision how to order them in the blend, and (iii) the decision how to split them up for the fusion – and shows that each of the stages exhibits distinct and significant patterns in their own right, but also when compared to (authentic and induced) error blends with regard to lengths and frequencies of source words, the similarities of source words to each other, and to the resulting blend (e.g., in terms of shared substrings, string-edit distances, and stress patterns).

While these and other previous studies (see Renner *et al.* [2012]) have produced a wealth of results, they were all based on observational samples of blends collected by the researcher. This is potentially problematic in the same way that speech error collections often studied in the 1970s and 1980s are: it is not clear that the collection of the data is not affected by the ease with which certain blends can be perceived or memorized. In other words, not all blends occurring in real life – the population of blends, so to speak – have an equal chance of ending up in the researcher's observational sample. A first experimental approach to follow up on the many observational studies was conducted by Arndt-Lappe & Plag [2013], who had 29 speakers of Irish English write and then pronounce blends in response to 60 written pairs of words; their source words systematically varied syllabic lengths and stress placements. They obtained altogether 1357 blend tokens from 107 ordered word pairs and largely corroborate existing observational studies (mostly focusing on Cannon [1986], Kubozono [1990], Gries [2004a-c], and Bat-El & Cohen [2006]): blends are typically as long as the longer source word, source words are often, but not always, split at constituent boundaries, and sw_2 determines the stress of the blend more than sw_1.

In this paper, we will also discuss experimentally-obtained blends; in this first case study, the focus will be on validating previous observational research.

2. Data and methods

2.1. Experimental design

The source words to be used as stimuli came from four distinct semantic domains that represented plausible scenarios for intentional blending: fruit, vegetables, dog breeds, and car brands (see Appendix A for the full prompts participants were given). The specific source words selected for each domain were controlled for syllabic, graphemic and phonemic length as well as frequency to the best extent possible. Eight mono-, bi-, and tri-syllabic source words were rotated in each participant form. Monosyllabic source words had 3-4 graphemes/phonemes; bisyllabic source words had 5-7 graphemes and 4-6 phonemes (with the exception of *kia* with 3 graphemes and phonemes); and trisyllabic source words had 6-10 graphemes and 6-8 phonemes. The source words were presented in such a way that once all participants completed the experiment, mono-, bi-, and trisyllabic source words were blended together an equal number of times. Source word frequencies were obtained from the *Corpus of Contemporary American English* (COCA), a publicly accessible corpus covering spoken news reports and interviews, fiction writing, magazines, newspapers, and academic writing. Of the potential source words in each domain, the most frequent and the least frequent were selected. The resulting list of source words was the following:

- fruit: *banana, cantaloupe, cherry, grape, guava, plum*
- vegetables: *bean, garbanzo, lentil, onion, potato, yam*
- dog breeds: *chihuahua, lab, mastiff, poodle, pug, retriever*
- car brands: *dodge, honda, jeep, kia, mercedes, pontiac*

To avoid potential priming effects, source words were never presented twice in a row as stimuli. Each participant saw an experimental form that contained 30 pairs of source words (15 pairs each from two out of the four semantic domains) and 30 filler items that served to shift participants' attention from the blending task to a sufficiently dissimilar task. The filler items were simple math problems such as divisions and multiplications, rounding of numbers, and fractions. 12 unique experimental forms were created so that in a group of 12 participants, two participants saw source word pairs from the same two domains, yet in different order of presentation of sw_1 and sw_2.

2.2. Procedure

All experiments took place in the laboratory of Stefanie Wulff and were approved by the University's Internal Review Board. All participants were college students enrolled at Stefanie Wulff's university, and all were native English speakers between the ages of 18 and 25. A research assistant walked participants through the informed consent form and a participant information form that asked for personal information such as language background, age, and sex. Participants were then seated in front of a computer screen for the experiment. The experiment was conducted in two rounds. In Experiment 1 (E1), participants were presented with the stimuli and filler items on the computer screen and then asked to record their response in writing using pen and

paper. In Experiment 2 (E2), a new group of participants were presented with the stimuli and filler items on a computer screen and then asked to articulate the stimulus or filler item out loud before recording their response in writing, and then to sound out their responses as well. To capture participants' oral productions, the entire experimental session was tape-recorded. 72 students participated in E1, yielding 2,188 blends; 84 students participated in E2, yielding 2,520 blends (in both experiments, discarded responses included the participant saying "I don't know" and repeating one or both source words without blending them). All written blends were copied into a spreadsheet, and all oral productions of source words and blends were transcribed using the CELEX phonetic alphabet [Baayen, Piepenbrock & Gulikers 1995].

2.3. Data annotation

Regarding the blend type, we determined for each grapheme/phoneme of the blend where its elements come from (we henceforth use the terms grapheme and letter interchangeably). For instance, consider Table 1 for our treatment of the well-known blend *brunch*. In this format modeled after Gries [2004c], the first two rows represent for each of the letters in sw_1, *breakfast*, whether it is in the blend (lower row) or not (upper row); the then next two rows do the same for sw_2, *lunch*, just in the reverse order, which is so that the middle two rows highlighted in bold comprise the blend. The resulting annotation for BLENDTYPE is shown in the last row, namely for each letter in the blend which of the two source words – 1 or 2 – it is from. The current example highlights how our annotation identifies what is often considered the prototypical kind of blend – the beginning of sw_1 followed by the end of sw_2 – namely as a sequence of one or more 1s followed by a sequence of one or more 2s; in regular expressions, might one might summarize this as "1+2+".

Table 1: Annotation of BLENDTYPE for *breakfast* × *lunch* → *brunch*

Letter slot	1	2	3	4	5	6	7	8	9
Letters from sw_1 in the blend			e	a	k	f	a	s	t
Letters from sw_1 in the blend	**b**	**r**							
Letters from sw_2 in the blend			**u**	**n**	**c**	**h**			
Letters from sw_2 not in the blend		l							
Annotation for letter BLENDTYPE	1	1	2	2	2	2			

This annotation can be extended to handle the maybe next most prototypical kind of blend, namely one that, around the point of fusion, involves overlap, i.e. graphemes or phonemes that occur in both source words, such as the *l* in *fool* × *philosopher* → *foolosopher*. These were marked with a 3, as shown in Table 2 for a blend from our data, *potato* × *lentil* → *potatil*.

Table 2: Annotation of letter BLENDTYPE for *potato* × *lentil* → *potatil*

Letter slot	1	2	3	4	5	6	7		

Letters from sw_1 in the blend						o			
Letters from sw_1 in the blend	**p**	**o**	**t**	**a**	**t**				
Letters from sw_2 in the blend					**t**	**i**	**l**		
Letters from sw_2 not in the blend		l	e	n					
Annotation for letter BLENDTYPE	1	1	3	1	3	2	2		

Finally, there was a very small number of blends where the subjects coined a blend on the basis of the letters, but when they pronounced it, that blend contained a phoneme that was not represented in either source word, but instead resulted from the subjects 'making phonemic sense' of their graphemically-motivated creation; those were coded as 4; consider Table 3 and Table 4 for the letter and phoneme annotation of the blend *jeep* × *honda* → *jenda*, respectively.

Additionally, for the oral responses, all source words and blends were also annotated for stress.

Table 3: Annotation of letter BLENDTYPE for *jeep* × *honda* → *jenda*

Letter slot	1	2	3	4	5				
Letters from sw_1 in the blend			e	p					
Letters from sw_1 in the blend	**j**	**e**							
Letters from sw_2 in the blend			**n**	**d**	**a**				
Letters from sw_2 not in the blend	h	o							
Annotation for letter BLENDTYPE	1	1	2	2	2				

Table 4: Annotation of phoneme BLENDTYPE for *jeep* × *honda* → *jenda*

Phoneme slot	1	2	3	4	5				
Phonemes from sw_1 in the blend		i	p						
Phonemes from sw_1 in the blend	_								
Phonemes from no sw in the blend		**e**							
Phonemes from sw_2 in the blend			**n**	**d**	**%**				
Phonemes from sw_2 not in the blend	h	Q							
Annotation for phoneme BLENDTYPE	1	4	2	2	2				

3. Hypothesis 1: the shorter source word contributes more of itself to the blend

In this section, we are revisiting the first hypothesis from above, which was first proposed by Kaunisto [2000] and then studied in, for instance, Gries [2004a-c].

3.1. Preparation of the data

In order to test Hypothesis 1, we needed the lengths of the source words in graphemes and phonemes as well as how much in percent they contributed to the blend. The graphemic lengths of the source words were straightforward to obtain from our master spreadsheet by just counting the number of characters for all source words. The contributions to the blends required a slightly more elaborate approach based on the blend lengths and their types as outlined above in Section 2.3. Based on that annotation, the contribution of

- sw_1 to the blend was the number of 1s and 3s in $_{BlendType}$ divided by the length of sw_1;
- sw_2 to the blend was the number of 1s and 3s in $_{BlendType}$ divided by the length of sw_2.

That is, for *brunch* (recall Table 1), the graphemic contributions of sw_1 and sw_2 are $^2/_9$ and $^4/_5$ respectively, for *potatil* (recall Table 2), the graphemic contributions of sw_1 and sw_2 are $^5/_6$ and $^3/_6$ respectively, etc.

3.2. Statistical analysis

In the existing literature on this hypothesis, the lengths of the source words and their contributions were expressed in a ternary format. That means, comparisons were made between the source words of each blend to determine

- for lengths, whether $sw_1>sw_2$, $sw_1=sw_2$, or $sw_1<sw_2$;
- for contributions to the blend as computed above, whether $sw_1>sw_2$ (i.e., whether sw_1 contributed more of itself than sw_2), $sw_1=sw_2$, or $sw_1<sw_2$.

Then, the frequencies for each combination were tallied and subjected to a chi-squared test or a Poisson regression. This approach is simple, but quite defensible for the observational data of previous work. If we apply this method here to the grapheme-based blends of E1, which we will use to outline our statistical methodology for this section, we get Table 5. The frequency distribution is significantly different from chance (X^2=266.51, df=4, p<10^{-10}, V=0.25) and the only positive Pearson residuals are in precisely the highlighted cells one would expect from, say, Gries [2004a: 654]: in summary, the shorter source word contributes more of itself to the blend and when both are equally long, they contribute equally much.

Table 5: Cross-tabulation of source words' lengths and contributions: observed frequencies (and Pearson residuals in parentheses)

Contribution Length	$sw_1<sw_2$	$sw_1=sw_2$	$sw_1>sw_2$	Totals

sw_1<sw_2	436	36	**504 (+7.56)**	976
sw_1=sw_2	120	**54 (+9.98)**	44	218
sw_1>sw_2	**672 (+4.83)**	62	260	994
Totals	1228	152	808	2188

However, the assumption of independence of data points that a chi-squared test relies on was already violated in the observational data. There, that violation was probably fairly inconsequential because the data comprised only a few blends that share certain source words; for instance, there were several blends with *sex* as sw_1. But in the present experimental data, the amount of repeated-measurements structure of this type is of course much higher: all blends were created from the same set of source words, and every speaker contributed many data points. Thus, while the above results are suggestive, a better approach is needed.

As an alternative, we adopted an ordinal mixed-effects modeling approach. For the dependent variable we first computed the following contribution percentage difference: contribution % sw_1 – contribution % sw_2. The resulting value ranged from -1 to +1: when it is high, sw_1 contributes much more of itself to the blend than sw_2; when it is low, sw_1 contributes much less of itself to the blend than sw_2; and when it is 0 or close to 0, both source words contribute about equally much. However, this set of values is very diverse (200 difference values with some less than 0.001 apart), many of them are only minimally different while at the same time meaning the same thing. Two differences of, say, 0.5 and 0.46 both mean sw_1 contributes much more than sw_2 – we do not need a linear regression to try to 'explain' that difference of 0.04 and would in fact not have much of a theoretical account at the level of quantitative resolution. Thus, we converted the difference values into a more useful ordinal response variable such that

- if -1 < difference < -0.25, the response variable was set to "sw_2 contributes more";
- if -0.25 ≤ difference ≤ 0.25, the response variable was set to "both contribute equally";
- if 0.25 < difference < 1, the response variable was set to "sw_1 contributes more".

This response variable was then modeled as a function of each source word's length (each as an orthogonal polynomial to the second degree to allow for curvature) and their interaction. As for the random effect structure, the only one that did not cause modeling problems consisted of varying intercepts for both sw_1 and sw_2 – additional varying intercepts for subjects exhibited very little variance in initial simple models and led to convergence problems with the fixed effects mentioned above.

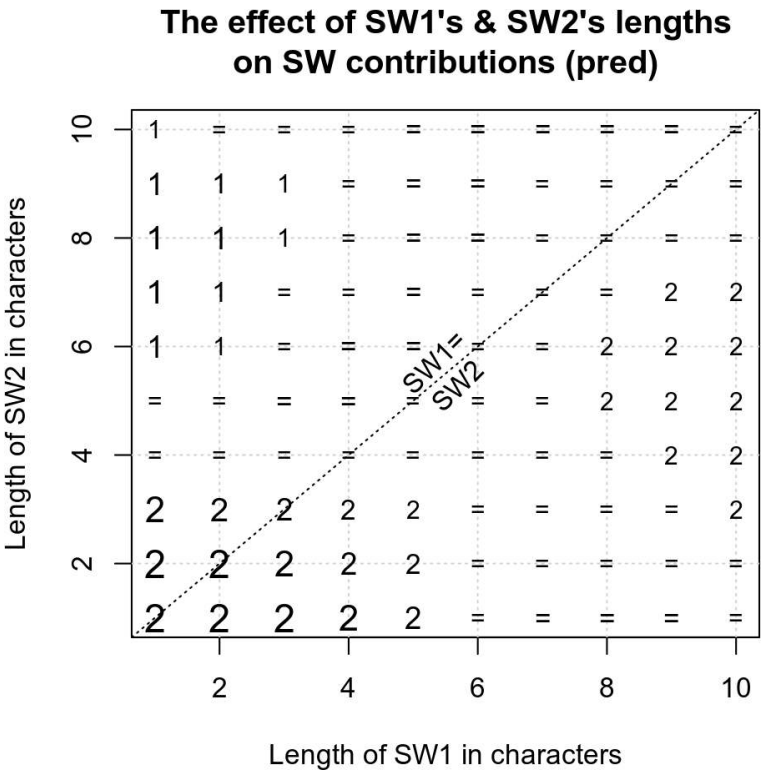
3.3. Results

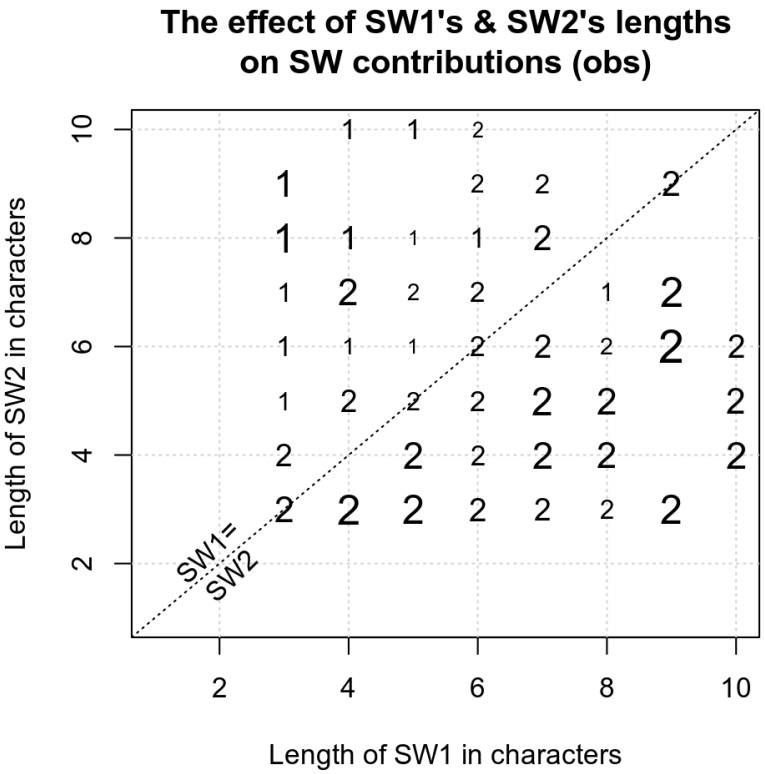
The above model provides for a highly significant fit to the data (LRT=94.84, df=8, $p<10^{-15}$), with the interaction of the two polynomials being significant as well (LRT=70.94, df=4, $p<10^{-13}$), allowing for no obvious simplification to the model. However, the strength of the effect is small: Nagelkerke's R^2=0.05. While a higher R^2 would have been desirable, the smallness of the value is not really surprising given that blend production is affected by many different and consciously manipulated factors, while we are testing only a single and very specific hypothesis here. Nevertheless, in order to be

safe, we computed two other mixed-effects models – one with the actual difference values as the response variable (e.g., the above 0.04) and one with a binary response variable ('sw_2 contributes more' vs. 'it does not'). While the numerical results differ, their implications with regard to Hypothesis 1 do not, which is why we proceed with our interpretation from what we considered to be the 'best' response variable.

Given the nature of our model – an ordinal model with polynomials interacting – its interpretation on the basis of the numerical results is impossible. We therefore proceed on the basis of predicted probabilities of the three outcomes, but since we have two numeric predictors and three levels in our response variable, the resulting 3-dimensional graphs are instructive (and beautiful), but cannot be used in a non-interactive print medium. Instead, we represent the results in two 2-dimensional plots. Each plot in Figure 1 has the lengths of sw_1 on the *x*-axis and the lengths of sw_2 on the *y*-axis, and within the coordinate systems we are plotting 1s and 2s (when sw_1 or sw_2 is predicted/observed to contribute more of itself respectively) and "=" (when both are predicted/observed to contribute equally). In the upper panel, we plot the results predicted by the model, with greater font sizes indicating that the predictions are more confident (i.e., the predicted probabilities are higher). In the lower panel, we plot whether for each observed combination of source word lengths, the contribution of sw_1 or sw_2 was higher (plotting 1s and 2s respectively); empty slots in the lower panel mean that no such combination of source word lengths was observed in the data (e.g., we had no situation where both sw_1 and sw_2 were 7 characters long).

Figure 1: Summary of the final model (graphemes, E1): predicted outcomes (upper panel) and observed outcomes (lower panel)





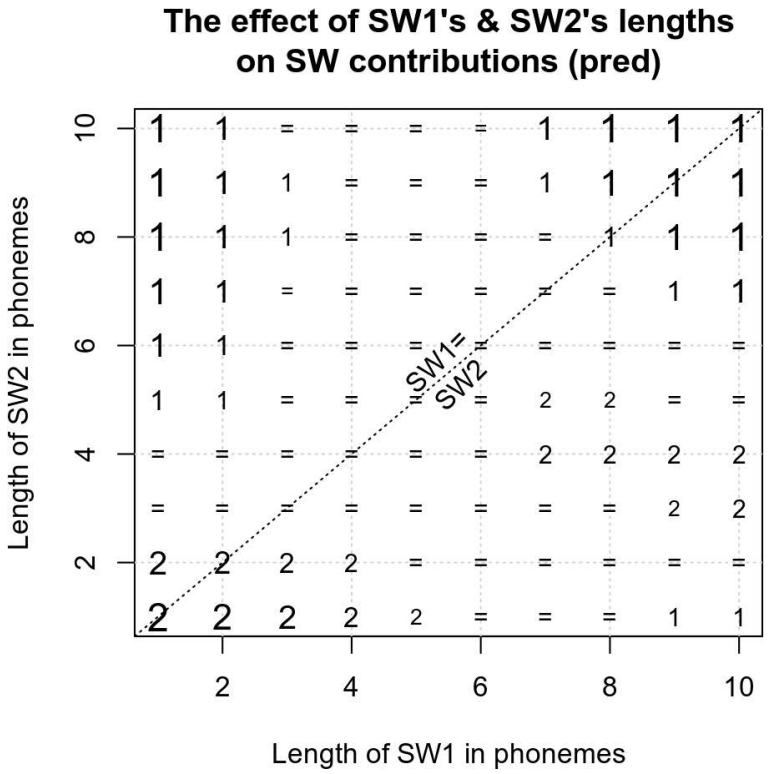
The overall complexity of the model notwithstanding, the results are interpretable and, in this case, fairly compatible with the simplistic chi-squared analysis, as is particularly clear from the lower panel: 1s (i.e. cases where sw_1 contributes more) and especially big 1s are mostly found in the top left part of the plot, where sw_1 is shorter than sw_2, and the situation is the reverse for 2s. While the lower panel does not show any "="s, it does show that many of the physically smaller 1s and 2s (i.e., when the distribution of the data is not clearly biased in favor of 1 or 2) are close to the main diagonal, where both source words are equally long.

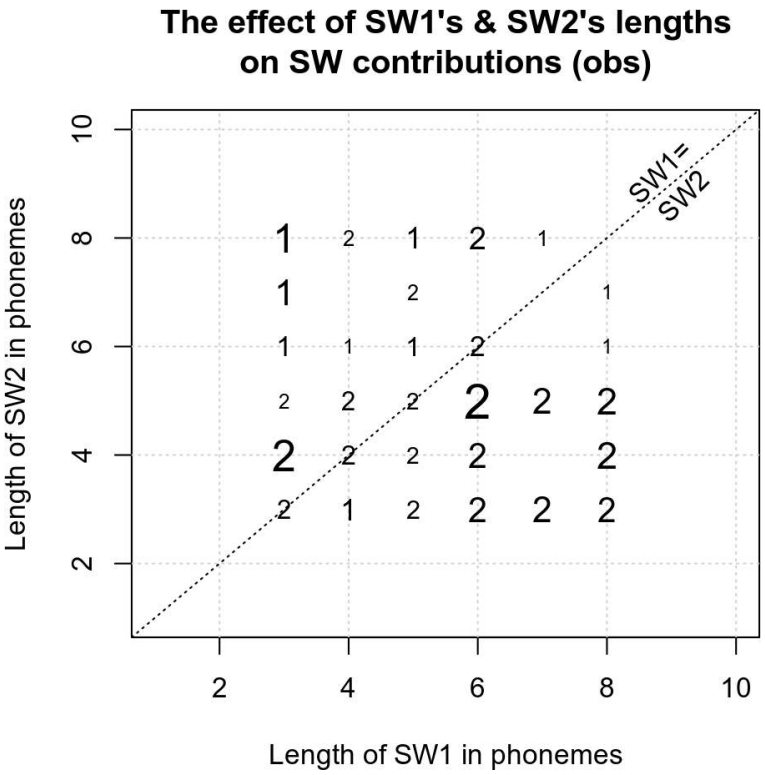
What about Hypothesis 1 for the phonemic contributions of source words in E2? The result of the initial exploratory chi-squared test for the phonemic lengths and contributions from E2 was extremely similar to that of E1: X^2=194.7, df=4, p<10^{-10}, V=0.1987, with the same three positive residuals only. For the same reasons as above, however, we proceeded with the ordinal mixed-effects model with the same fixed-effects predictors (just for the phoneme data in E2) and the same random-effects structure (this time, however, varying intercepts per subject could be included in the model without problems).

This model, too, provides for a highly significant fit to the data (LRT=48.96, df=8, p<10^{-7}), with the interaction of the two polynomials being significant as well (LRT=28.81, df=4, p<10^{-5}), allowing for no obvious simplification to the model; however, the strength of the effect is even smaller than before: Nagelkerke's R^2=0.022. Again we computed two other mixed-effects models and again their results led to the same implications with regard to Hypothesis 1. Consequently, the visualization of the results in Figure 2 is the same as above. The model predictions in the upper panel are not particularly

instructive, which is not surprising given the very low R^2, but the lower panel is a bit more informative: there are more and bigger 1s in the top left triangle (where sw_1 is shorter than sw_2) and there are more and bigger 2s in the bottom right triangle (where sw_2 is shorter than sw_1), which is indeed as expected.

Figure 2: Summary of the final model (phonemes, E2): predicted outcomes (upper panel) and observed outcomes (lower panel)





In sum, the effects obtained from the experimental data are in the hypothesized direction – the shorter source word contributes more of itself to the blend – but they are noticeably weaker than they were in the observational data. In other words, while the previous results are supported, the present data also raise the specter that the convenience-sampling kind of approach that accounts for part of the observational data appears to amplify certain effects, maybe because the people who identified the blends unwittingly were more likely to notice formations as blends if they exhibited the hypothesized structure.

4. Hypothesis 2: sw_2 determines BLENDSTRESS (more)

4.1. (Additional) Preparation of the data

A first exploration of Hypothesis 2 consisted again of 3-dimensional cross-tabulation, namely cross-tabulating the four stress patterns of each sw_1 and sw_2 (stressed: S, stressed-unstressed: SU, stressed-unstressed-unstressed: SUU, and unstressed-stressed-unstressed: USU) with the stress patterns of the blends provided by the participants in E2. However, the 12 stress patterns of the blends were quite Zipfian-distributed, which would be problematic for most kinds of categorical data analyses, which do not usually respond well to response variable with 12 levels, four of which are attested less than four times. At the same time, the four most frequent blend stress patterns not only accounted for nearly 91% of all tokens, but were also exactly the stress patterns exhibited by the source words. In order to describe the data best, we proceeded to do

both analyses. In what we will now call *analysis*$_1$, we created a variable BLENDSTRESSWHENCE, which stated for each blend where it got its stress pattern from; for that we needed to distinguish four levels:

- a level *sw1*, if the blend had sw$_1$'s stress pattern, and sw$_2$'s stress pattern was different;
- a level *sw2*, if the blend had sw$_2$'s stress pattern, and sw$_1$'s stress pattern was different;
- a level *sw1sw2* if both sw$_1$ and sw$_2$ had the same stress pattern as the blend;
- a level *neither*, if the blend had a stress pattern different from sw$_1$ and sw$_2$.

This variable then became the response variable in a first statistical analysis discussed presently. However, in the other analysis, which we will now call *analysis*$_2$, we 'reduced' the data by discarding the ≈9% of cases where the blend had a stress pattern that was neither that of sw$_1$ nor that of sw$_2$.

4.2. Statistical analysis

As before, a 3-dimensional chi-squared test or a hierarchical configural frequency analysis of either data set would have been possible, but also probably problematic, given the repeated measures structure in this data set. Therefore, we opted again for a mixed-effects model, this time – given our response variables BLENDSTRESSWHENCE (in *analysis*$_1$) and BLENDSTRESS (in *analysis*$_2$) had four levels – Bayesian multinomial mixed-effects models. The predictors were SW1STRESS and SW2STRESS as well as their interaction, the random-effects structure consisted of varying intercepts of each sw$_1$, each sw$_2$, and each participant; our modeling parameters were four chains each with 2000 iterations (after a burn-in for each of 1000); these numbers may seem low, but see the convergence results below.

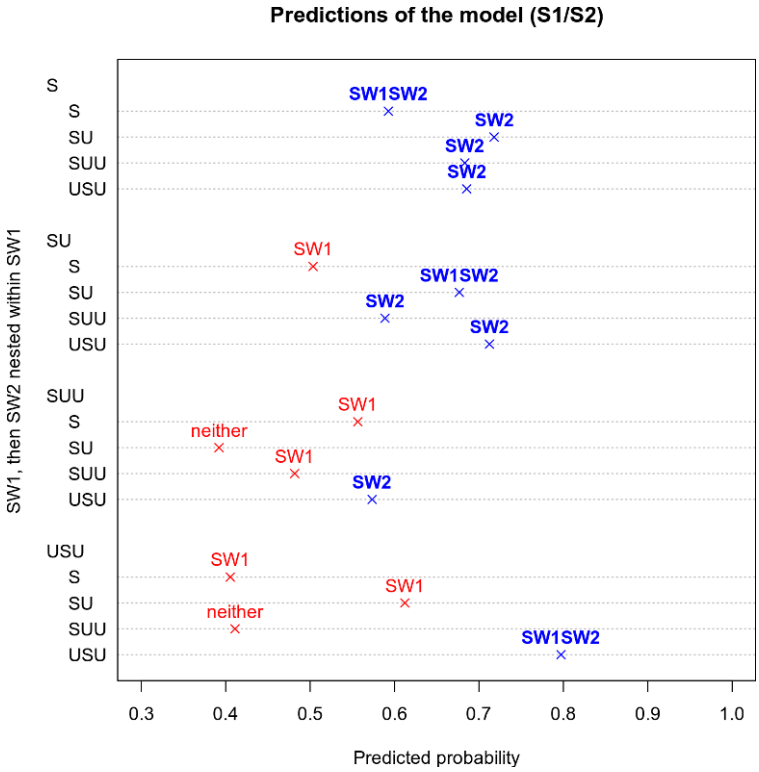
4.3. Results

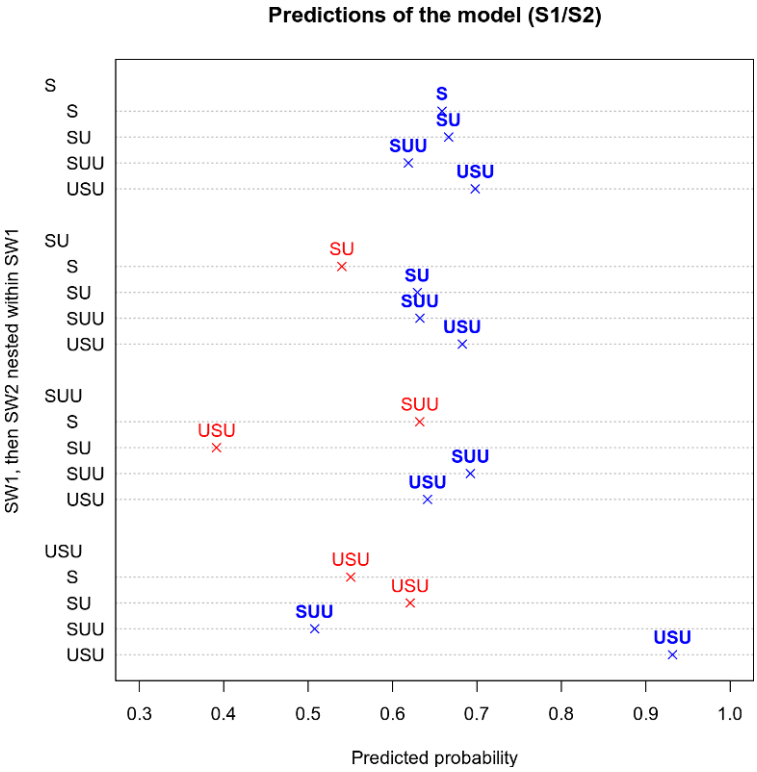
The models from both analyses converged just about perfectly (no R-hat values >1.01) and resulted in quite a good fit and accuracy. While, to the best of our knowledge, R^2-values for this kind of model are not available, the classification accuracy of both models are quite good: 61.5% for *analysis*$_1$, 66.7% for *analysis*$_2$, which according to an exact binomial test is highly significantly better ($p<10^{-96}$ and $p<10^{-184}$ respectively) than the baselines of 40.75% or 36.8% (the frequency of the most frequent level of the response variables).

Interpreting the results of such models is more difficult than those of ordinal models: A multinomial model with an interaction like ours would return four (levels of predictor 1) times four (levels of predictor 2) times four (levels of the dependent variable) = 64 predicted probabilities and/or 64 observed probabilities. We therefore proceeded as follows (described here first for *analysis*$_1$): First, for all 16 combinations of the four levels of the two predictor variables SW1STRESS and SW2STRESS, we identified which level of BLENDSTRESSWHENCE had the highest predicted probability and what that predicted probability was. These are represented in the upper panel of Figure 3, which shows the levels of SW1STRESS in the 'outer' *y*-axis and the levels of SW2STRESS nested within those, with the predicted probability indicated by "×" on the *x*-axis with a small label on top of the "×" representing which level of BLENDSTRESSWHENCE is predicted for that combination of SW1STRESS and SW2STRESS. That means that the second row from the top indicates the

following: 'When SW1STRESS is *S* and SW2STRESS is *SU*, then the blend is predicted to have the stress pattern of *SW2* (with a probability of >0.7, 0718 to be precise).' The crosses and labels are printed in blue when the blend did indeed exhibit the stress pattern of sw_2 or of both sw_1 and sw_2, and else in red. For $analysis_2$, we show the equivalent in the lower panel of Figure 3.

Figure 3: Predicted probabilities of predicted outcomes of BLENDSTRESSWHENCE for all the data (upper panel) and predicted probabilities of predicted outcomes of BLENDSTRESS for the reduced data (lower panel)





The results are supportive of Hypothesis 2, but maybe not as strong as expected and maybe with a twist: Both panels show that, when SW1STRESS is S, sw_2 – whatever its stress pattern – determines BLENDSTRESS, and when SW1STRESS is SU, then sw_2 determines BLENDSTRESS unless sw_2 is monosyllabic. However, when SW1STRESS is SUU, sw_2 only determines BLENDSTRESS when it also is trisyllabic, and when SW1STRESS is USU, sw_2 only 'co-determines' BLENDSTRESSWHENCE when it also does in *analysis*$_1$ and only determines BLENDSTRESS in *analysis*$_2$ when sw_2 also is trisyllabic.

While these results support that, on the whole, sw_2 is a stronger determinant of BLENDSTRESS than sw_1, part of the results is also compatible with the alternative (if only at times coincidental) account that the longer source word determines BLENDSTRESS. This is supported by the observation that trisyllabic sw_1s determine BLENDSTRESS more than shorter sw_1s. So, our above result of mostly sw_2 determining BLENDSTRESS could be an artefact resulting from (i) BLENDSTRESS really being determined by the longer source word and (ii) the fact that previous studies have shown that sw_2 is on average a bit longer than sw_1 (e.g., Kelly [1998], Gries [2004c, 2012]).

We therefore ran three separate models on the *analysis*$_2$ version of the data: They all featured BLENDSTRESS as the response variable and the same random-effects structure as above, but the first one had the length difference in phonemes as a predictor, the second one the length difference in syllables between sw_1 and sw_2 as a predictor, and the third one the length difference in syllables as well as SW1STRESS and SW2STRESS and all their interactions as predictors. The results were unambiguous: All models converged but the classification accuracies of the first two did not even reach baseline

performance; the third model had a good classification accuracy of 67.1% (expectable since it featured the same two predictors that were already successful without the added length difference), but (i) that classification accuracy is not significantly better than the one above for the model without the length-difference predictor (p_{binomial}=0.3375) and (ii) a *WAIC* comparison showed that adding the length difference and its interaction to the model with 'just' SW1STRESS and SW2STRESS did not make the model reliably better (ELPD difference = -5.6, but with a standard error of 6.2). Therefore, this case study delivers results that are largely supportive of Hypothesis 2 and, thus, previous analyses based on the observational data. In addition, in our first multifactorial study of blend stress assignment, we also find that sw_2's dominance, so to speak, does *not* seem to be reducible to a length effect and does not benefit from being augmented with a length effect.

5. Hypothesis 3: blending maximizes similarity between source words and blends

As discussed in much previous work, the similarity of blends to source words can be measured on a variety of dimensions as, for instance, in terms of stress pattern as in the previous section. This section focuses on similarity/distance in terms of graphemes (E1) and phonemes (E2). Here, we can adopt two perspectives:

- similarity can be enhanced by picking two source words to blend that are already more similar to each other than random words are to each other;
- similarity can be enhanced by blending the two source words in such a way that the resulting blend retains a high degree of similarity (and, thus, recognizability) to the source words.

5.1. The similarity of source words to each other

The first perspective is in fact requires observational data because it can only be studied if one has a wide range of source word-blend combinations to look at. Previous work has confirmed that source words of blends are more similar to each other than random word pairs (or source words of complex clippings, for that matter; see Gries [2006, 2012]). However, we are also returning to this briefly here even with our experimental source words so as to offer at least an idea of how the source words we used compare to previous findings.

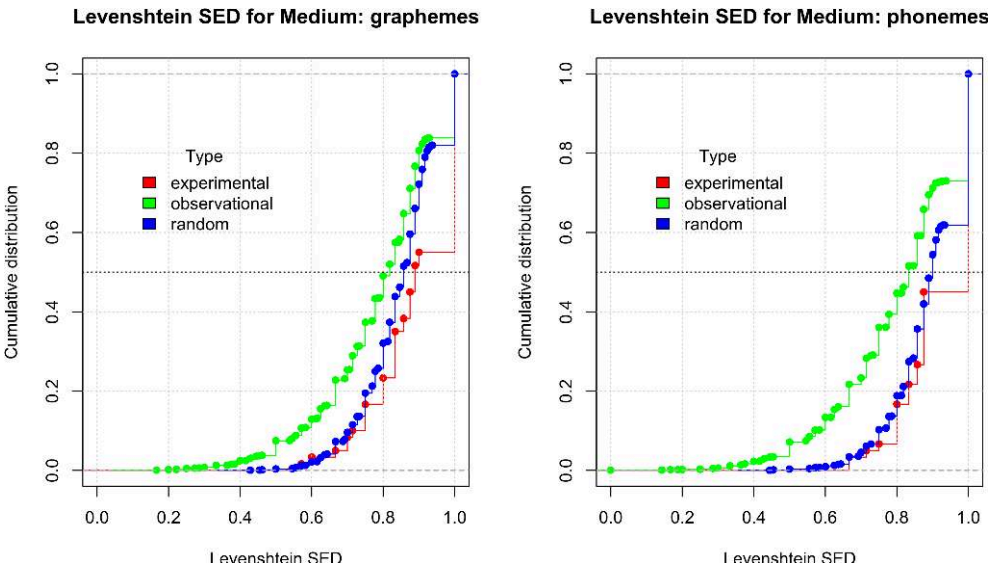
To do that, we computed pairwise Levenshtein string-edit distances (SEDs) – i.e. the inverse of similarity – of the source words to each other for

- the blends in our experimental data;
- the blends in the latest version of Gries's collection of observational data, which was last discussed in Gries [2012];
- 4708 pairs of randomly-chosen words, a number which corresponds to the number of experimental blends collected in both E1 and E2.

In all these cases we computed both grapheme- and phoneme-based similarity. For instance, the SED for *channel* and *tunnel* is $^3/_7$ because one the longer of the two source words has seven characters and one needs three operations to get from *channel* to *tunnel*: deleting the *c*, replacing the *h* with a *t*, and replacing the *a* with a *u*. Then, we visually compared their empirical cumulative distributions, which are represented in

Figure 4. It is plain to see that the string-edit distances of words have relatively similar medians (of around 0.8 or 0.85 at y=0.5) and similar curves. Somewhat unsurprisingly, the source words of the blends from the observational data have the lowest distances – i.e. the highest degrees of similarity – but the reassuring finding is that our stimulus source words do not already behave very differently (in either direction).

Figure 4: Ecdf plots for string-edit distances between words based on graphemes (upper panel) and phonemes (lower panel)



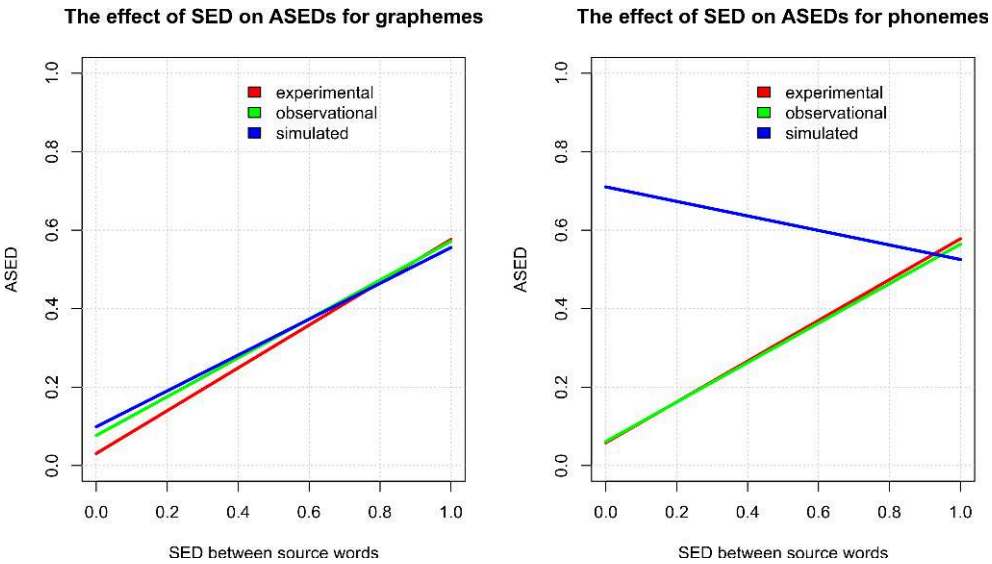
5.2. The similarities of source words to blends

As for the second perspective and to measure how much the blending of the source words leads to a similar blend, we followed Gries's [2012] general logic and computed for each blend an average Levenshtein string-edit distance (ASED) value, which means, we took the average of the SED between sw_1 and the blend and the SED between sw_2 and the blend. For *channel* × *tunnel* → *chunnel*, the ASED is 3/14=0.2143, namely the mean of the SED of *channel* and *chunnel* ($^1/_7$) and the SED of *tunnel* and *chunnel* ($^2/_7$). We did this for

- all experimental blends of our data, using graphemes for the blends in E1 and phonemes for those of E2;
- all observational blends of Gries's [2012] data, using both graphemes and phonemes;
- all blends one could hypothetically generate from 6 pairs of source words (sw_1s: *strong*, *rich*, *television*, *flight*, *sloppy*, *Chevrolet*; sw_2s: *powerful*, *handsome*, *armchair*, *suitcase*, *medical*, *Cadillac*) while respecting phonotactic rules of English, meaning we did not include a hypothetical blend *rich* × *handsome* → *rndsome*.

Then we fit a linear model to see how much the ASEDs – the similarity-preserving ways in which blends are formed from the source words – vary as a function of Medium (*graphemes* vs. *phonemes*), Type (*experimental* vs. *observational* vs. *hypothetical*/simulated) and the SEDs between the source words. The model revealed a significant three-way interaction between these predictors (p=0.016), which is represented in Figure 5.

Figure 5: The effect of SED on ASED for graphemes in the final model (upper panel) and the effect of Type:Medium on ASED in the final model (lower panel)



The upper panel indicates that for graphemes, all three blend types behave the same: the more similar the source words are, the more similarly they also are jointly to the blend. This is reassuring because it confirms previous results based on observational blends, namely that blend creation involves this kind of using similarity to enhance word play and recognizability. At the same time, it is surprising that the mechanistically-created simulated blends, which by definition do not heed to this, reveal the same trend. An exploration of means does suggest, however, that as expected, the simulated blends scored lower on ASEDs than the other two kinds of blends.

For the phonemes, the results are reassuring: the experimental blends behave just like the observational ones, and both are significantly different from the simulated blends. We did not include confidence intervals to reduce visual clutter, but the 95%-CI for simulated blends (phonemes) includes 0, reflecting that their similarity to the source words does not increase even as the source words become more similar to each other.

All in all, we find that previous results based on the observational blends are supported. While there is one effect we cannot at present account for – the fact that simulated blends score as high on similarity between source words and blends as experimental and observational blends – this effect does not undermine the main point of this section, namely that the experimental blends pattern like the observational ones from prior studies.

6. Concluding remarks

In sum, the results are first rather encouraging. While many studies, including several of Stefan Th. Gries, have proceeded using collections of blends that were often accrued under less-than-ideal sampling conditions, the results of our three case studies join most of those by Arndt-Lappe & Plag [2013] and lend credence to this kind of previous work. Section 3 showed that the shorter source word indeed contributes more of itself to the blend (using ordinal mixed-effects modeling); Section 4 showed that sw_2 is indeed most influential in determining blends' stress patterns (using multinomial

mixed-effects modeling); and Section 5 showed that blending attempts to increase similarity between source words and blends (using traditional linear modeling).

That being said, we have also seen at least a bit of evidence that the observational data studied much in the past can, under certain circumstances, impart anticonservative results in the sense that effects appear stronger in the observational data than in the more controlled experimental data. The fear that this might happen motivated this study in the first place, but then also means that much more such 'validational work' needs to be done to determine which other results, if any, were amplified due to the nature of the observational data.

One other conclusion to be drawn from this study certainly for us is a recognition of how difficult some of these analyses are even just from a methodological and statistical perspective. Even the controlled experimental data required not only an inordinate amount of transcription and error-checking, but also a data management/processing and statistical approach that go beyond much of mainstream types of analysis (of blends, but maybe also in much of linguistics in general). While it is possible to get some results from simple cross-tabulation and chi-square tests (as in Gries [2012] or Arndt-Lappe & Plag [2013]), once one wants to go beyond this and adopt the kind of analyses common in other contemporary corpus- and psycholinguistic studies, things become complicated very quickly. For instance, our case study of Hypothesis 3 first generated complete null results until we noticed that the source word similarities must be included as a control variable – only then did we see the more reasonable results reported here. Given the multitude of results that still await similar kinds of validation and the large number of factors that affect blend formation or at least need to be controlled for, blend researchers certainly have their work cut out for them.

BIBLIOGRAPHY

ARNDT-LAPPE Sabine & PLAG Ingo, 2013, "The role of prosodic structure in the formation of English blends", *English Language and Linguistics* 17(3), 537-563.

BAAYEN Harald R., PIEPENBROCK Richard & GULIKERS Leon, 1995, *The CELEX lexical database* (CD-ROM). University of Pennsylvania, Philadelphia, PA: Linguistic Data Consortium.

BAT-EL Outi & COHEN Evan-Gary, 2006, "Stress in English blends: a constraint-based approach", *in* RENNER Vincent, MANIEZ François & ARNAUD Pierre J. L. (Eds.), *Cross-disciplinary perspectives on lexical blending*, Berlin: Mouton de Gruyter, 193-212.

CANNON Garland, 1986, "Blends in English word-formation", *Linguistics* 24(4), 725-753.

GRIES Stefan Th., 2004a, "Shouldn't it be *breakfunch*? A quantitative analysis of the structure of blends", *Linguistics* 42(3), 639-667.

GRIES Stefan Th., 2004b, "Isn't that fantabulous? How similarity motivates intentional morphological blends in English", *in* ACHARD Michel & KEMMER Suzanne (Eds.), *Language, culture, and mind*, Stanford, CA: CSLI, 415-428.

GRIES, Stefan Th., 2004c, "Some characteristics of English morphological blends", *in* ANDRONIS Mary A., DEBENPORT Erin, PYCHA Anne, & YOSHIMURA Keiko (Eds.), *Papers from the 38th Regional Meeting of the Chicago Linguistics Society: Vol. II. The Panels*, Chicago, IL: Chicago Linguistics Society, 201-216.

GRIES, Stefan Th., 2006, "Cognitive determinants of subtractive word-formation processes: a corpus-based perspective", *Cognitive Linguistics* 17(4), 535-558.

GRIES Stefan Th., 2012, "Quantitative corpus data on blend formation: psycho- and cognitive-linguistic perspectives", *in* RENNER Vincent, MANIEZ François & ARNAUD Pierre J. L. (Eds.), *Cross-disciplinary perspectives on lexical blending*, Berlin & New York: Mouton de Gruyter, 145-167.

KAUNISTO Mark, 2000, "Relations and proportions in the formation of blend words", Paper presented at the Fourth Conference of the International Quantitative Linguistics Association (Qualico), Prague.

KELLY Michael H., 1998, "To *brunch* or to *brench*: some aspects of blend structure", *Linguistics* 36(3), 579-590.

KUBOZONO Haruo, 1990, "Phonological constraints on blending in English as a case for phonology-morphology interface", *Yearbook of Morphology* 3, 1-20.

RENNER Vincent, MANIEZ François & ARNAUD Pierre J.L. (Eds.), 2012, *Cross-disciplinary perspectives on lexical blending*, Berlin & New York: Mouton de Gruyter.

APPENDIXES

Appendix A. Prompts for the blending task

(1) You are a marketing agent for a fruit snacks company that has just come out with a series of new fruit snacks that combines flavors of two different fruits. Your job is to entice people to buy the products by creating clever product names that combine the two fruits together. Keep the order of the fruits in the name the same as you are given.

Example: *peach* × *apple* → *papple*

(2) You are an agricultural scientist trying to patent new types of vegetables containing genetic material from two different types of vegetables. Unfortunately, competitors are also trying to patent the same combinations. You must come up with creative names for your new vegetables to ensure that the patented names are unique. Keep the order of the vegetables in the new names the same as you are given.

Example: *lettuce* × *radish* → *lettish*

(3) You are a dog breeder trying to get famous by coming up with the most popular new breed. You've decided to breed several different types of dogs together. For each of the following pairs, come up with a catchy name for the new type of breed by blending the names of the two types of dogs together. Keep the type of dogs in the order they are given.

Example: *beagle* × *husky* → *busky*

(4) For each pair of words you're given assume you're in a new merger meeting between two automobile companies. You're a marketing agent whose job is to blend the

names of the car brands together in order to come up with a clever new car brand name. Keep the names in the order that you're given.

Example: *Chevrolet* × *Cadillac* → *Chevradillac*

ABSTRACTS

In this paper, we discuss the results of a blend production experiment and how it relates to previous research that was nearly exclusively based on observational data. Specifically, we study three different findings from published research, namely that (i) the shorter source word contributes more of itself to the blend than the longer source word, (ii) source word2 determines blend stress (more than source word1), and (iii) blending maximizes similarity between source words and blends. Using statistical techniques so far not employed in research on blends, we show that most findings from observational data regarding the three hypotheses studied are supported, but also occasionally tampered down.

Cet article analyse les résultats d'une étude expérimentale de productions d'amalgames et la façon dont ils diffèrent ou non de ceux d'études antérieures fondées sur des données d'observation. Plus précisément, nous analysons trois conclusions tirées de recherches déjà publiées, à savoir : (i) le mot source le plus court contribue pour une part plus significative à l'amalgame que le mot source plus long, (ii) le mot source2 détermine l'accentuation de l'amalgame (plus que le mot source 1), et (iii) le processus d'amalgamation tire au maximum partie de la similarité entre les mots sources et les amalgames produits. Nous avons eu recours à des techniques statistiques non employées jusqu'à présent pour l'étude du processus d'amalgamation, afin de démontrer que la plupart des conclusions tirées des données d'observation quant aux trois hypothèses ci-dessus sont confirmées, mais doivent également parfois être modulées.

INDEX

Keywords: experimental blend production, observational blend collection, source word lengths, similarity, regression modeling
Mots-clés: production expérimentale d'amalgames, collecte d'amalgames d'observation, longueurs des mots sources, similarité, modèle de régression

AUTHORS

STEFANIE WULFF

University of Florida & UiT The Arctic University of Norway
swulff@ufl.edu

STEFAN TH. GRIES

UC Santa Barbara & JLU Giessen
stgries@linguistics.ucsb.edu