Cognitive determinants of subtractive word formation: A corpus-based perspective*

STEFAN TH. GRIES

Abstract

This paper investigates mechanisms underlying the coining of intentional morphological blends and complex clippings. In one case study, I investigate the degree to which (a corpus-based definition of) psycholinguistic recognition points play a role in these subtractive word-formation processes. Also, I am concerned with the issue whether a separation of these two categories, which has been embraced by some but not all morphologists, is supported. Given the role that similarity plays in subtractive word-formation processes, a second case study investigates the degree to which the source words of blends and complex clippings are similar to each other and, again, whether the empirical findings warrant this distinction in the first place.

Keywords: *blends; complex clippings; subtractive word-formation; uniqueness points; recognition points; recognizability; similarity; corpora.*

1. Introduction

One of the most creative word-formation processes—where *creative* is used in the sense of 'defying characterization by means of hard-and-fast productive rules'—is blending, i.e., the intentional subtractive word-formation process exemplified by a few examples in (1), where parenthesized letters are those that enter into the blend.

(1) a. (br)eakfast × l(unch) → brunch
b. (mot)or × h(otel) → motel
c. (fanta)stic × f(abulous) → fantabulous
d. (fraud) × (auditor) → frauditor

While these examples are probably all too well-known, they mask the fact that blending is a process which has so far not been defined in such

Cognitive Linguistics 17–4 (2006), 535–558
DOI 10.1515/COG.2006.017
0936–5907/06/0017–0535

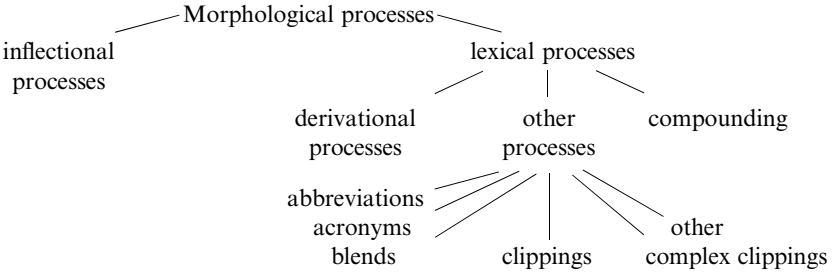

Figure 1. *Simplistic schema of a frequent classification of morphological processes*

a way as to properly set it apart from a variety of other subtractive processes which are superficially similar on one or more dimensions. Part of the reason for this lack of a widely accepted definition is probably the fact that subtractive word-formation processes are among the most understudied word-formation processes. In fact, for some scholars and in a variety of textbooks, they are not even part of regular derivational word formation proper because they are conscious processes that defy characterization by hard-and-fast productive morphological rules. More specifically, a very simplistic classification openly embraced or at least implied in much morphological work is the one represented in Figure 1 (cf. Algeo 1978 for a refined multidimensional classification).

A definition of blending to which many scholars could probably agree is the one in (2).

(2) Blending as a word-formation process involves coining a new word out of already existing source words such that, *typically*,

– two words (rather than three or more) are merged;
– one or both of the words undergoes shortening in the merger, which may be graphemic or segmental;
– if no shortening occurs, the words exhibit partial overlap, which may be graphemic or segmental (cf. (1d)).

While the above examples in (1) fit the definition in (2) quite well, the issue becomes much more complicated when looking at more varied cases. Is a word *W* a blend even if

– the merging is 'nonlinear / recursive'? (cf., e.g., *transmigrate* × *modify* → *transmogrify*)
– *W* contains material that is not from one of the source words or that has been changed in the process of blending? (cf., e.g., *quick* × *concrete* → *quikrete* and *deliciously* × *delightfully* → *delishfully*)

- *W* looks like a neo-classical compound? (cf., e.g., *movie* × *marathon* → *moviethon*)
- the process contracts syntagmatically adjacent words? (cf., e.g., *permanent* × *agriculture* → *permaculture*)

Also, note that the above definition in (2) leaves open the location of the splinters in their source words: Is a word *W* a blend only if the beginning of the first source word (henceforth sw_1) is merged with the end of the second source word (henceforth sw_2) as above in (1) or also if the beginning of sw_1 is merged with the beginning of source word sw_2, i.e., what has sometimes been referred to as a complex clipping (cf., e.g., *system* × *administrator* → *sysadmin*)?

If one turns to previous work on blends, three points become immediately obvious. First, there is a large body of mostly classificatory work, trying to answer questions such as 'how does one distinguish blends from other similar word-formation processes such as the one outlined above?'. This would be exactly the kind of study needed to determine the categorial status of blends, and much of the work falling under this heading raises important issues and/or proposes interesting criteria on the basis of which word-formation processes can be distinguished (cf. Algeo 1977, Algeo 1978 for a paradigm example, and López Rúa 2002, 2004 for a prototype-inspired approach largely using, but apparently unaware of, Algeo's criteria). However interesting these studies are, one of their shortcomings is that, polemically speaking, they basically attempt to squeeze blends etc. into an a priori established set of categories on the basis of some criteria without ever determining to what degree the criteria invoked are warranted when subjected to empirical scrutiny.

Second, in contrast to much classificatory work, there is only a handful of studies which attempt to tackle the issue of how blends are actually formed.[1] These studies adopt a preliminary definition of blends much as the one I proposed above and investigate, for example, the order of the source words in blends, the choice of the location of the cut-off points, the lengths of the source words' splinters constituting the blend, the role similarity plays on different levels of analysis, etc.; cf. Kubozono (1990), Berg (1998), Kelly (1998), Kaunisto (2000), and for work from a more cognitive perspective, cf. Lehrer (1996), Kemmer (2003), Gries (2004a, b, c). Especially the work by Gries has been concerned with the fact that blend coiners choose source words for a blend that (i) communicate what is to be communicated and that (ii) are more similar to each other graphemically, segmentally, and phonologically than one would expect on the basis of chance. Also, not only do blend coiners choose similar words to blend, they also blend them in such a way as to render the blend

similar enough for both source words to be recognized again since otherwise the wit of many blends could not be appreciated in the first place. However, a shortcoming of this work by Gries is that he investigated the notion of recognizability only with respect to the amount of material of each source word that is still part of the blend even though other approaches are potentially more useful and revealing.

Third, some scholars at least are puzzled by these two issues—the complexity of how blends are actually formed and the issue of how to come up with a viable morphological taxonomy—to such a degree that their conclusions about the degree of patterning observable at all are rather pessimistic. Bauer (1983) and Cannon (1986) admit it most openly:

> 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)

> we find no discernible relationship between phonology [...] and a viable blend. [...] This fact helps to make blends one of the most unpredictable categories of word-formation. (Cannon 1986: 744)

In the present study, I will address the two shortcomings mentioned above: (i) the fact that recognizability may correlate with more than just the amount of graphemic/segmental material of the source words and the blend and (ii) the fact that classificatory approaches to subtractive word-formation processes often do not motivate the choice of parameters, which is why a bottom-up test of proposed distinctions may be useful and in fact called for.

Section 2 will investigate the degree to which the psycholinguistic notion of recognition or uniqueness points plays a role in the off-line formation of blends. One point to be looked at is that coiners of subtractive word formations must ensure that their creation's component parts can be recognized again. However, the secure way of doing this—simply including (nearly) the whole word—is not available since blends and complex clippings would then not exhibit the wit for which they are frequently put to use (esp. in advertising) because (i) no cunning word play would be involved and (ii) the blend would not be similar to both its source words anymore. If, for example, the automobile brands Chevrolet and Cadillac were to merge, I dare say nobody requested to symbolize that in a witty blend would suggest *Chevrolet* × *Cadillac* → *Chevroladillac*. Thus, I will investigate whether word coiners make use of the so-called recognition point of the source words involved in order to ensure that, first, their new creation is not too long and thus not very witty (as it would be if

both words would hardly be shortened and just stuck together) and, second, not too short to be recognized in the first place (as it would be if too little of the source words is still present in the blend, as in *Chevrolet* × *Cadillac* → *Chac*). As a matter of fact, I would assume that in, say advertising and brand name development, even more factors play a role, including for example the desire to make the word formation not too similar to competing product names, which can be important to minimize the risks of trademark infringements or customers mixing up names of medications etc. In a way, all this can be phrased in a parlance that is very familiar to cognitive linguists such that blend coining is a very intricate process requiring coiners to deliberate how to strike an optimal balance between many different competing motivations; the process will therefore often involve experimenting with, and fine-tuning of, different formations until the 'right' formation has been identified and is, thus, clearly an offline process.

Another point to be tested in the very same section is whether the notion of recognition points also allows us to distinguish between two different subtractive word-formation processes, namely blends and complex clippings. This comparison would be interesting because scholars are divided as to whether these are actually two different classes, which is why an empirical study may contribute to our knowledge of the theoretical status of the two processes.

Given the above and the results to be discussed in Section 2, Section 3 will then turn to the role of similarity in subtractive word formation. While earlier work by Gries has shown that similarity plays a role for the formation of blends (cf. especially Gries 2004b), it has remained unclear whether this also applies to other subtractive word formations. I will investigate whether blends and complex clippings behave differently when looking at the role similarity plays in their formation and the objective is, again, to determine whether different subtractive word-formation processes can be distinguished on the basis of data rather than preconceptions about what their defining characteristics are.

2. A corpus-based approach to recognition/uniqueness points

2.1 *Methods*

In this section, I will investigate the role of recognition/uniqueness points on blend-formation. However, I must first clarify the corpus-based operationalization in quite some detail to make explicit what method was chosen on which grounds. The uniqueness point *UP* of a word *W* is the point at which W can be uniquely identified from a set of candidate words. The

recognition point *RP* of a word *W* is the empirical estimate of *W*'s *UP*. More specifically, *RP* is the point at which a majority of speakers (e.g., 85%) can recognize *W* with a high probability (e.g., 80%) when presented with parts of *W*. It will be important below to know that *RP*s exhibit a word-frequency effect of tokens: more frequent words are recognized faster (by approx. 20%) than their closest competitors (cf. Marslen-Wilson 1987: 91f.).

*RP*s have been determined both experimentally—for example using gating tasks, phoneme monitoring, shadowing, lexical decision tasks, or word vs. nonword detection—as well as on the basis of (usually electronic) dictionaries or on the basis of natural language corpora. As is obvious from the title of this paper, I will approach *RP*s in the latter fashion, i.e., on a strictly corpus-linguistic basis. To give two examples for how *RP*s maybe approximated very simplistically in corpora:

– in the British National Corpus World Edition, the letter sequence *islamiciza* narrows possible continuations down to the unique possibility *islamicization*;
– in the CELEX database for English (Baayen et al. 1995), the phoneme sequence [ɜbənaɪzeɪ] narrows possible continuations down to the unique possibility [ɜbənaɪzeɪʃən].[2]

One attractive feature of using a corpus-based approach to *RP*s is that this approach makes it possible to not only identify the *RP* as such, but one can easily also identify also the number of types of all candidate sets as well as the frequency distribution of each candidate set. For example, when the target word is *islamicization*, a corpus-based frequency list of words allows for identifying all words starting with *i* and their frequencies of occurrence, all words starting with *is* and their frequencies of occurrence, etc. up to *islamiciza*., where only *islamicization* and its frequency are left. However, how would one approach cut-off points of blends and other subtractive word-formation processes?

Let us approach this question using the example of (*agit*)*ation* × (*prop*)*aganda* → *agitprop*. In other words, how would we operationalize the RP of *agitation*, i.e. the point where subjects may be (most) likely to guess from the part they are exposed to that sw_1 that entered into the complex clipping is *agitation*? One easily conceivable possibility would be to, first, determine for each beginning of *agitation* the number of types and/or tokens that start with this beginning (cf. Table 1).

In a second step, one could then plot the type and token frequencies along the parts of *agitation* to determine the point where the cost of adding another letter (of course, the logic also applies to phonemes) does not result in an appropriate further reduction of the search space. This

Table 1. *Type and token frequencies of words beginning with beginnings of* agitation *(based on the CELEX database)*

Part of sw_1	Types starting with part of sw_1	Tokens starting with part of sw_1	Examples
a	4,347	2,840,567	*a*, *able*, *adore*, *agree*, ...
ag	137	45,320	*agave*, *age*, ...
agi	12	347	*agile*, *agitator*, ...
agit	8	267	...
agita	8	267	...
agitat	8	267	...
agitati	3	125	*agitation*(*s*), *agitating*
agitatio	2	118	*agitation*(*s*)
agitation	2	118	*agitation*(*s*)

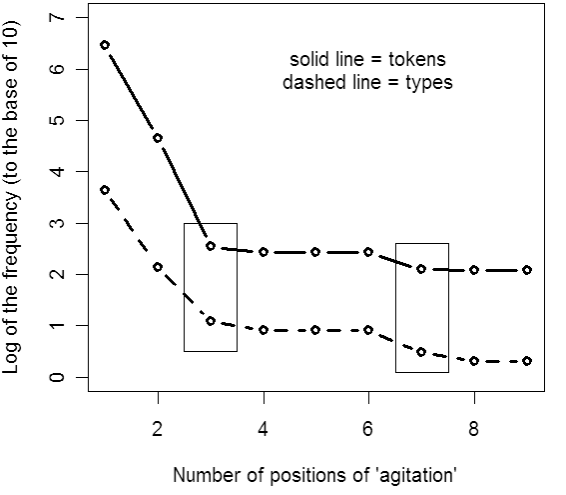

Figure 2. *A 'scree plot' representation for type and token frequencies of parts of* agitation *(based on the CELEX database)*

approach, basically an adaptation of the scree plot technique used in factor analysis, is represented in Figure 2.

Figure 2 suggests two approximations to the *RP* that are marked with the rectangles. One is at the third letter, i.e., at *agi*, while the other is at the seventh letter, i.e., at *agitati*. On both occasions, the search space is reduced markedly but the next letter will not make guessing the word much easier.

However attractive this approach may seem at first, it has a few problems associated with it. A practical problem is that, the larger the

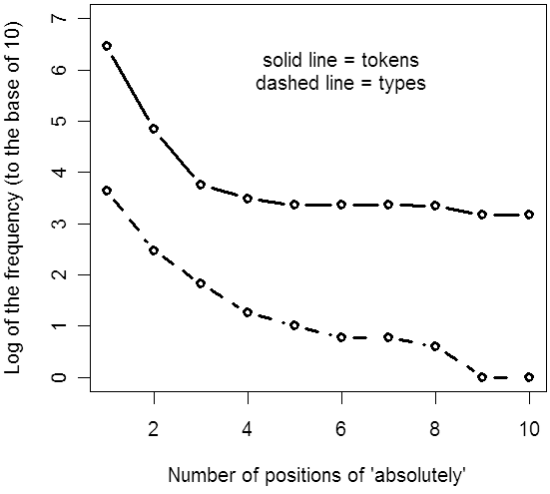

Figure 3. *A scree plot representation for type and token frequencies of parts of* absolutely *(based on the CELEX database)*

database, the more graphs one would have to inspect. Even worse, for each blend one would have to look at four graphs: (i) sw_1 using letters (as above), (ii) sw_2 using letters, (iii) sw_1 using phonemes, and (iv) sw_2 using phonemes. With more than a few hundred cases this becomes infeasible quickly. Another practical problem is that not all cases can be decided straightforwardly, as is obvious from the analogous representation for *absolutely* (as used in, say, *absolutely* × *positively* → *absotively*), where Figure 3 shows that no similarly obvious *RP* emerges.

However, while these shortcomings might be overcome using some ingenious statistical technique,[3] there is one shortcoming that can not. The point is that this method only looks at the number of word types or tokens which are possible given a particular part of a source word—it does not take into consideration the frequency distributions of these candidate types or tokens. Imagine a word starting with the letter sequence *abs*. Let us also assume that upon giving the first two letters, *ab*, there are 100 word tokens in our corpus that start with *ab*. Now, the method exemplified above would result in a data point (2, 2): we are looking at the second letter, and $\log_{10} 100$ is two, too. However, the method does not take into consideration the frequency distribution of the 100 tokens. Let us assume just for the sake of the argument that the 100 tokens in fact instantiate just four types. There are now two extreme possibilities for how the distribution may look like, which are represented in Figure 4.

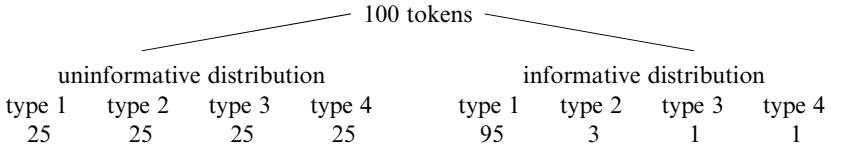

Figure 4. *Hypothetical distributions of four types across 100 tokens*

Obviously, the left distribution is extremely uninformative: *ab* is not a good clue because the four types from the remaining candidate set are all equally likely, which is also reflected in the entropy value for this distribution: $H = 2$. The right distribution, however, is very informative: the likelihood that type 1 is the target word is overwhelmingly high and entropy is correspondingly low: $H \approx 0.35$. However, the method outlined above cannot distinguish between these two distributions. Thus, what is needed is a way of identifying distributions which makes it easy to identify a source word that does not only depend on the number of types or tokens in the candidate set. In this paper, I will adopt the following method to approximate the *RP* of a word *W*. For each part of a source word *W* of a subtractive word-formation process (i.e., for *a*, *ag*, *agi*, *agit*, . . . , *agitatio*, *agitation*)

– count the number of types in the corpus that begin with this part;
– count the number of tokens in the corpus that begin with this part;
– determine the number of types that begin with this part that have higher token frequencies than the target word;
– locate the first position of the minimum of these frequencies.

Let us clarify this procedure on the basis of the example from Table 1 above; consider Table 2.

Table 2. *Type and token frequencies of words beginning with beginnings of* agitation

Part of sw_1	Types starting with part of sw_1	Tokens starting with part of sw_1	Frequency rank of *agitation*
a	4,347	2,840,567	595
ag	137	45,320	24
agi	12	347	1
agit	8	267	1
agita	8	267	1
agitat	8	267	1
agitati	3	125	1
agitatio	2	118	1
agitation	2	118	1

The three left columns are the same as before, the key change is the rightmost column. It provides the frequency rank of the target word, *agitation*, of all types that start with the part given in the leftmost column. In other words, Table 2 is to be interpreted as follows. There are 2,840,567 tokens in the CELEX database starting with *a*. These are made up of 4,347 types. Of these 4,347 types, 594 (= 595 − 1 for *agitation* itself) are more frequent than *agitation*, which is why *a* is not a good clue to *agitation*. The second row reveals that there are 45,320 tokens in the CELEX database starting with *ag*. These are made up of 137 types. Of these 137 types, 23 are more frequent than *agitation*. Now finally, there are 347 tokens in the CELEX database, which are made up of 12 types, and of these 12 types, *agitation* is the most frequent one. Thus, this is the first position of the overall minimum, and, thus, it is here where the part of the leftmost column becomes the most likely clue for *agitation* for the first time. While it is this point within *agitation* that is singled out by the proposed method, this point is probably still a little early for a psycholinguistic RP proper, which is why I will refer to *agi* as the *SP* (for *selection point*) of *agitation* (following a suggestion by R. Harald Baayen).

One final step is necessary. We have now seen how *RP*s can be approximated on a corpus-linguistic basis using *SP*s, but the final questions that remain are (i) how to determine whether coiners of blends or complex clippings care about *SP*s when they choose a cut-off point and (ii) how to test whatever result we get for significance. What is needed is a random baseline or, even better, an index that measures the deviation of all possible cut-off points from the actually chosen cut-off point.

My answer to these challenges can be understood easiest with reference to Figure 5. I compute for each source word the *SP* as above (circled in Figure 5) and for each position at which the coiner of a subtractive word formation I compute the distance in letters/segments from the *SP*, which are given in the last row of Figure 5. From this we can compute the average random distance to the cut-off point at the *SP*, namely the mean of the set of all distances $\{-2, -1, 0, 1, 2, 3, 4, 5, 6\}$, which is +2. The final step then consists in comparing the actual cut-off point at distance +1

letters of source word to be recognized	a	g	i	t	a	t	i	o	n
frequency rank of source word	595	24	(1)	1	1	1	1	1	1
distance to the ideal cut-off point	−2	−1	0	+1	+2	+3	+4	+5	+6
				↑	↑				
				actual cut-off point	'randomly'-chosen point				

Figure 5. *Distances of cut-off points from the* SP

with this random one, and in this case we find that whoever coined *agitprop* chose a cut-off point for *agitation* whose distance to the *SP* is smaller than the average random distance from the *SP*.

Note that while the overall method is now clear, we have so far only been concerned with the letters of sw_1 in a complex clipping. The same has to be done for the letters of sw_2. In addition and to be on the safe side, the same has to be done for the phonemes of both source words. Finally, once we look at blends, the direction of analysis for sw_2 must be reversed because a defining characteristic of the blends investigated here is that these are word formations where sw_2 contributes its end to the blend rather than its beginning. This means that for the analysis of the sw_2 in blends, the signs in the row 'distance to the ideal cut-off point' were reversed such that, as before in Figure 5, positive distances indicate that the coiner has included more material into the blend / complex clipping than necessary while negative distances indicate that the coiner has included less material than necessary; cf. Figure 6 for an example.

The data to be investigated here are from my own collection of 2,200 subtractive word formations compiled from a variety of sources including dictionaries, scholarly works on blends, the 'Among the new Words' column of the journal American Speech and many more. Of these, 1,740 were included into the present study, namely

- 1,672 cases where the beginning of the sw_1 was merged with the end of sw_2 (i.e., cases that are uncontroversially referred to as blends);
- 68 cases where the beginning of sw_1 was merged with the beginning of sw_2 (i.e., cases that are often referred to as complex clippings).[4]

The corpus that was used as a reference for the type and token frequencies is a slightly modified version of the CELEX database. While larger corpora are certainly available, CELEX has the advantage that the pronunciations of most—but not all; cf. n. 4—source words of my blends are included and could be extracted and evaluated using a variety of niPerl

letters of source word to be recognized	f	a	b	u	l	o	u	s
frequency rank of source word	2152	(90)	65	5	3	1	1	1
distance to the ideal cut-off point	+2	+1	0	−1	−2	−3	−4	−5

↑ actual cut-off point

↑ 'randomly' chosen point

Figure 6. *Distances of cut-off points from the* SP

(cf. Sutton 2005) and R (cf. R Development Core Team 2005) scripts that I wrote. On the other hand, much psycholinguistic work has actually be conducted on the basis of much smaller corpora. For example, Wurm and Ross's (2001) study on Conditional Root Uniqueness Points uses frequency data from the Brown corpus, a corpus about 5.6% the size of the one used for CELEX, which is why the size of CELEX together with its transcription data makes it a more than suitable alternative.

On the basis of this data set, a repeated-measures ANOVA was computed. The explanatory variable was the distance to the ideal cut-off point, the within-items explanatory variables were MEDIUM (letters vs. phonemes), SOURCEWORD (sw_1 vs. sw_2), and the CUT-OFFPOINT (chosen vs. random average), and the between-items explanatory variable was FORMATION (blend vs. complex clipping).

2.2 *Results and discussion*

There are highly significant effects of both FORMATION and CUT-OFFPOINT (but not of MEDIUM). However, these two effects are of course only interesting when investigated in conjunction since only then can one distinguish to what degree blends and complex clippings differ with respect to chosen and random cut-off points. One of the truly relevant terms is, therefore, the interaction, FORMATION × CUT-OFFPOINT. While this interaction turns out to be highly significant ($F_{1,1738} = 30.96$; $p < 0.001$), it is still qualified by an even higher-order interaction, namely FORMATION × CUT-OFFPOINT × SOURCEWORD ($F_{1,1738} = 104.84$; $p < 0.001$). As usual, the nature of such a higher-order interaction is best understood on the basis of a graphical display as that in Figure 7.

For both sw_1 and sw_2 of blends, we find that the absolute average distances of the actually chosen cut-off points to the *SP* are smaller than those of the random cut-off points to the *SP*. More precisely, the average cut-off point for sw_1 is nearly exactly on the cut-off point,[5] but the average cut-off point for sw_2 is half an element, i.e., half a letter or phoneme too early.[6] For complex clippings, on the other hand, a completely different picture emerges: The absolute average distances of the actually chosen cut-off points to the *SP* are approximately three times as high as the ones expected by chance, reflecting the fact that coiners of complex clippings cut off both source words way before the *SP*.

These results provide considerable support for the role of *SP*s in the formation of blends but not complex clippings, which in turn provides bottom-up support for distinguishing the two word-formation processes within a general classification of morphological processes.[7] But what are we to make of the fact that, while blend coiners observe the *SP* nearly

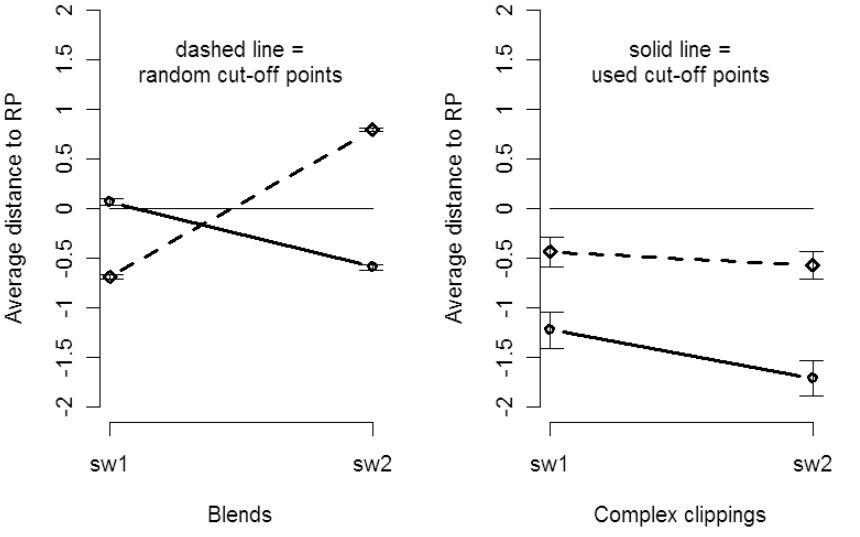

Figure 7. *Mean distances (±1 s.e.) to SP in* FORMATION × CUT-OFFPOINT × SOURCEWORD

perfectly for sw_1, they cut off sw_2 approximately half a segment too early? Several explanations are conceivable, some of which immediately lead to suggestions for how to refine analyses for future work.

First, the fact that blend coiners do not consider it necessary to take the *SP* as seriously for sw_2 as they do for sw_1 could be due to contextual effects: Gating tasks show that context speeds up recognition by more than 40% (cf. Grosjean 1980) so that sw_2 may simply need less material for identification: By the time the reader/hearer arrives at linguistic material that does not seem to belong to sw_1 anymore, he has already all general contextual clues and the clues triggered by sw_1 at his disposal for identifying sw_2. Especially this latter kind of contextual information, that about sw_1, may be particularly relevant in two ways. On the one hand, sw_1 will provide semantic clues that facilitate the access of sw_2. On the other hand, recall the fact that blends often involve a fusion of words involving some degree of overlap in the middle (cf. the definition of blends in (2) above as well as the examples (1b–d)). Now, overlapping letters or segments would of course make the retrieval of source words easier in general, and especially so in the case of blends since sw_2 is not encountered in a way that words are normally, namely from left to right. One look at the data indicates that blends make much more use of overlapping than complex clippings; cf. Table 3 giving the observed frequencies (and expected frequencies in parentheses).

Table 3. *The correlation between (lack of) overlap and blends and complex clippings*

	blends	complex clippings	totals
overlap	962 (939.8)	16 (38.2)	978
no overlap	710 (732.2)	52 (29.8)	762
totals	1,672	68	1,740

As a chi-square test shows, overlap is much underrepresented in complex clippings ($\chi^2 = 30.7$; $df = 1$; $p < 0.001$; odds ratio ≈ 4.4). Thus, the mere fact that blends utilize overlap more strongly may obviate the need for adhering to the *SP* for sw_2. Add to this the facts that Gries (esp. Gries 2004b) has shown that sw_1 and sw_2 exhibit similarity on many different levels of analysis and that it is exactly sw_2 that has a significant tendency to determine the overall length of the blend and its phonological structure. Thus, the blend as such already provides many clues for the identification of sw_2. For all these reasons, the retrieval of sw_2 is facilitated in so many respects that the *RP* is perhaps just not that essential anymore.

Second, the database investigated included as blends all items fitting the characterization given in (2) above. The reason for that was to be able to start from an as objective and replicable database as possible. However, as a result of this, the database includes a variety of what is sometimes referred to as neo-classical compounds, i.e. formations one part of which is a part of a Latin/Greek term; examples include several formations with, say, *tele-* or *-thon* (an example used above was *movie-thon*). Since these splinters have arguably taken on morpheme status in the meantime, such formations may distort the picture: obviously, if a splinter does not have morpheme status, a reader/hearer needs to recover its source word to retrieve its meaning and is dependent on graphemic/phonological clues, but if a splinter has already been morphemicized, then no particular graphemic/phonological clues are necessary anymore since no source word needs to be accessed anymore. While it is not immediately obvious how an objective distinction between blends, complex clippings, and neo-classical compounds would ultimately look like, a replication of this study on the basis of such a more precise characterization may well show that the set of blends excluding neo-classical compounds does in fact behave even more as expected with respect to *SP*s and may even support having a separate class for neo-classical compounds to begin with.

Third, recall that the identification of the SPs has been made extremely rigorously (the source word had to be *exactly* the most frequent word)—maybe a probabilistic, similarity-based approach utilizing articulatory

features or a wider range of highly ranked words would allow for a better match between the actual choices of cut-off points and *SP*s. In addition, both reviewers pointed out quite correctly that what is at issue here is not so much the exact match, but some overall match of the part of the source word with the morphological family of the target word. For the example in the introductory example of this section, this may mean that it would be enough to recognize that *agit* 'activates' the family made up by *agitate*, *agitates*, *agitated*, *agitating*, *agitatation*, and maybe other forms such as *agitatively* and/or even prefixed/suffixed versions of these words. I agree that this may very well be the case, but at present at least I don't see how this proposal would be tackled in practice. On the one hand, the man power required to largely manually classify 2,000+ word formations into morphological families awaits funding from a larger grant. On the other hand, one would still have to devise a statistic that unifies the different frequencies of all the words of a morphological family, ideally also incorporating into that statistic information about the combined type-token ratio etc. For these two reasons, I will have to postpone such an analysis until later.

Fourth, the analysis of the sw_2 was made on the assumption that speakers try to match the parts of sw_2 against candidate words that end with these parts. However, maybe speakers are not that precise in their formation of blends and just try to match source word parts against candidate words that *contain* these parts.

Finally, it goes without saying that determinants other than *SP*s may strongly influence the formation of blends and complex clippings; the syllabic and/or prosodic structure and the location of the points of breaking up the source words and fusing them together are cases in point (cf. Kubozono 1990 for some discussion). In fact, even a cursory glance at a few dozen examples will support this intuition very strongly and the current data base is actually exhaustively annotated for the precise phonological structure of the blends and complex clippings as well as their source words and their points of break up and fusion to allow for precisely such analyses. However, considerations of space do not allow to discuss this matter here, which I will therefore address in future work.

In sum, this section has shown that (i) the formation of blends is in fact substantially correlated with corpus-derived *SP*s and that (ii) a superficially very similar word-formation process, complex clipping, does not exhibit such a tendency. Interestingly, this finding also indicates that the intentional creation of blends at least suggests that their coiners make use of a general mechanism involved in the comprehension of words when they form a neologism, as if trying to anticipate comprehenders' strategies. Possible explanations for the findings as well as extensions and

amendments were discussed; several of these involved the issue of how similar the source words are to each other and to the blend and the following section will investigate whether blends and complex clippings differ with respect to the degree of similarity between their source words.

3. A quantitative approach to word similarity

The fact that blends usually involve an element of word play or punning is a triviality and has been noted long ago; instances such as *fool* × *philosopher* → *foolosopher* are well-known cases in point. However, there are few studies that have actually investigated the role of similarity empirically. One of these is Gries (2004b), who looks at the different levels at which similarity is found (i) between the two source words of a blend and (ii) between the two source words on the one hand and the resulting blend on the other hand. While Gries finds significant effects of similarity on the graphemic, segmental, and phonological level, he restricts his attention to blends as defined above, leaving aside the issue of whether other subtractive word-formation processes may in fact exhibit similar effects. In this section, I will examine the questions of (i) whether the source words entering into blends are more similar to each other than would be expected by chance and (ii) whether this is also true of the source words entering into complex clippings. If the two subtractive word-formation processes exhibited different preferences regarding the similarity of their source words, this would provide additional support to distinguishing them in morphological classification.

3.1 *Methods*

The database used for this case study is the same as that for the previous one. As a first step, however, the notion of similarity has to be operationalized. In this case study, I will use three different kinds of similarity measures in order to make sure that no single measure introduces unwanted bias into the analysis. Consider as an example the blend *channel* × *tunnel* → *chunnel*. The three kinds of measures I will use are the following:

– four bigram-based measures: Dice, XDice, and weighted versions of Dice and XDice (cf. Brew, Chris and McKelvie 1996): Bigram-based measures are built on the assumption that two words are similar to each other to the extent that they share bigrams, i.e., sequences of two adjacent letters (or phonemes). In the case of *channel* and *tunnel*, for example, *channel* consists of the six bigrams {*ch*, *ha*, *an*, *nn*, *ne*, *el*} and *tunnel* consists of the five bigrams {*tu*, *un*, *nn*, *ne*, *el*}, so the

number of shared bigrams of the two words is three: *nn*, *ne*, and *el*. Dice divides the number of shared bigrams multiplied with two by the number of all bigrams; thus Dice for *channel* and *tunnel* is $^{6}/_{11} = 0.545$. XDice does the same, but includes extended bigrams, too, i.e., for *tunnel* {*tn*, *un*, *ne*, *nl*}, resulting in $^{10}/_{20} = 0.5$. Finally, the two weighted versions increase the count of matching bigrams at the beginning and end of the words by one to account for the fact that word beginnings and endings are psycholinguistically important points. In the case of *channel* and *tunnel*, weighted Dice becomes $^{6+1 \text{ (the matching } \{el\})}/_{11+2} = 0.538$.[8]

- four longest-common-subsequence (LCS) based measures: LCS_1, LCS_2, Relative LCS_1, and Relative LCS_2: LCS-based measures utilize the number of shared letters occurring in both words in the same order (but not necessarily contiguously). The longest common subsequence was determined by means of an R script I wrote. It splits up the both source words into segments (letters or phonemes) and then goes through, say, sw_1 segment by segment to determine whether one segment of sw_1 is also identical to one segment of sw_2. If this is the case—for example, the first *n* in *channel* and *tunnel*—then the position of this segment in sw_1 and sw_2 is stored (4 for *channel* and 3 for *tunnel*) so that the search for the next matching segment(s) only uses the segments after the match. In the case of *channel* and *tunnel*, all the following segments can be aligned. LCS_1 and LCS_2 divide this number by the length of sw_1 and sw_2 respectively, i.e. $^{4}/_{7}$ is LCS_1 (for *channel*) and $^{4}/_{6}$ is LCS_2 (for *tunnel*). Relative LCS_1 divides this number by the length of the longer word only, and Relative LCS_2 divides it by the summed length of both source words.
- four edit-distance (ED) based measures: ED_1, ED_2, mean ED, minimal ED: ED-based measures simply count the number of single-element operations necessary to convert one word into another. For *channel* and *tunnel*, the number of operations is three: delete the *c*, replace *h* by *t*, replace *a* by *u*. ED_1 and ED_2 are the edit distance from sw_1 to sw_2 and vice versa; mean ED is the mean of ED_1 and ED_2, and minimal ED is the minimum of ED_1 and ED_2; the edit distance used is the Levenshtein string edit distance as implemented in the function *agrep* in R.[9]

Again, these computations were performed for all 1,740 blends and complex clippings in my database once on the basis of letters and once on the basis of phonemes. In order to also obtain a random baseline against which the data from the authentic word formations can be compared, I wrote an R script which picked 1,000 words at random from the

CELEX database and computed all $4 \cdot 3 = 12$ measures from above on all 499,500 pairs that can be generated from 1,000 words.

The most complex statistical design that could be implemented to determine the role similarity plays for blends and complex clippings would be a doubly repeated measures MANOVA: The response variables would be all values of the measures of similarity. The within-items explanatory variables would be MEDIUM (letters vs. phonemes) and the 12 different measures introduced above. The between-items explanatory variable would be the kind of word formation FORMATION (blend vs. complex clipping vs. randomly chosen words). However, since the main interest is in fact just on FORMATION (source words of blends vs. source words of complex clipping vs. random word pairs) whereas the interactions are largely irrelevant, several steps were undertaken to simplify the procedure. First, the four different measures in each of the three groups turned out to be correlated highly enough to just pool them.[10] Second, given the different directionality of the similarity measures (cf. above n. 9), I just did separate ANOVAs for the bigram-based measures, the LCS-based measures, and the ED-based measures.

3.2 *Results and discussion*

The most relevant results are summarized in the panels of Figure 8 (means and 99% confidence intervals of the interaction between MEDIUM, FORMATION, and the kind of similarity measure).

First, for the bigram-based measure and the LCS-based measure all effects are highly significant, which is little surprising given the large number of elements involved and which is why measures of effect size are more useful here. The strongest effect is in fact FORMATION: As is indicated in the two left panels, the similarity of the source words of blends is consistently higher than that of complex clippings and randomly chosen word pairs. In fact, complex clippings are nearly perfectly in the middle between blends and randomly-generated word pairs. However, the confidence interval of complex clippings in writing is in fact very close to the mean of randomly generated word pairs, which is as small as one would expect from a random baseline. Interestingly, there is also a significant interaction between the FORMATION and MEDIUM: While it is obvious that the average similarity of the randomly-chosen word pairs is higher in writing (given the smaller inventory of letters compared to sounds), the average similarity of the authentic word formations is much higher in speaking.

Second, for the ED-based measure again all differences are highly significant. Again, the effect of FORMATION is strongest such that blends

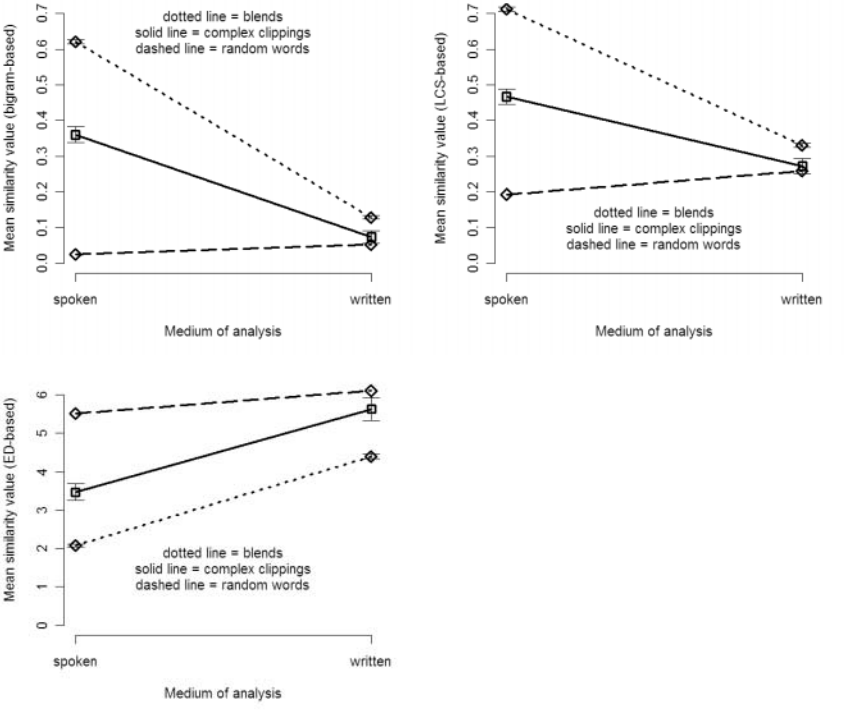

Figure 8. *Mean similarity values of the different kinds of similarity measure for* MEDIUM × FORMATION

exhibit the highest degrees of similarity and complex clippings being in the middle between blends and random word pairs. The interaction between FORMATION and MEDIUM, though significant, does not yield any interesting findings other than perhaps a tendency that with the authentic word formations again the spoken medium shows a stronger similarity effect.

There are still some minor problems with this approach. One is the problem of how to acknowledge the fact that pronunciation varies across speakers. For example, there is a blend *cinema* × *universal* → *cineversal*. The point is that in order to be as objective as possible, I coded the pronunications of these words on the basis of the CELEX database. For this example, this meant coding *cinema* as /sɪnəmə/ and *universal* as /junɪvɜsəl/ and all phonemic measures would have to consider the vowel following the [n] as different. However, it may well be possible that the coiner of this blend actually pronounces *cinema* as /sɪnɪmə/ or *universal* as /junəvɜsəl/, rendering the words more similar to a phonemic measure than their transcriptions in the CELEX database. Similarly, even if the

blend coiner pronounces the words as they are transcribed in CELEX, it may still be that he exploits the articulatory similarity between /ɪ/ and /ə/. Both of these issues could be handled by developing similarity measures that (i) are based on articulatory features and (ii) handle such cases probabilistically or by exhaustively permuting possible pronunciations. However, both these refinements are terribly complex since they presuppose an agreed-upon weighting of articulatory features, which does not appear to exist (cf. Kondrak 2002, 2003 for work in this direction, which can unfortunately not be applied to the present data). Also, the objectivity of the analysis will be difficult to uphold: the pronunciations I consider possible may not be those other scholars would consider possible.

Lastly, it is again obvious that additional determinants, especially phonological ones, will exert a considerable influence on the formation of blends and complex clippings. As in Section 2, the fact that these factors were not included here is not mean to downplay their indubitable relevance. The interested reader is referred to Kubozono (1990) and Gries (2004b) for some discussion of phonological determinants of subtractive word-formation.

Be that as it may, these suggestions for refinement must not diminish the fact that this case study has gone further in the analysis of graphemic and phonemic similarity of subtractive word-formation processes than any previous one, and the conclusion from this case study is clear: This section has shown that the two word-formation processes differ strongly in terms of the amount of similarity exhibited by the source words they involve. The absolute size of this effect is particularly strong in speaking, but in writing complex clippings behave nearly like random word pairs. On the basis of these findings, a theoretical distinction between the two kinds of word-formation processes is clearly supported.

4. Conclusion

The previous two sections have shown that subtractive word formation is by far not as arbitrary as has often been assumed. In addition to some previous findings (most notably Kubozono, Berg, Kelly, and Gries) concerning source words' lengths, frequencies, contributions to new coinages, similarity etc., it has now become clear that source words of blends are chosen such that they are much more similar to each other than random words and, and this is the crucial point, that this is *not* necessarily true of complex clippings. Also, source words of blends are merged such that their cut-off points are closer to their psycholinguistic *RP* than those of the source words of complex clippings or randomly chosen cut-off points.

These findings have some implications for the study of 'freaky' word-formation processes. On the one hand, it shows that these processes do allow for empirical analysis in a variety of ways, but in ways that many morphologists dealing with them have not yet begun to think: Using multifactorial statistics and randomly generated baselines etc. in the analysis of large collections of examples is not part of much work in subtractive word formation, but I hope that the previous sections demonstrate the potential of such approaches (however much individual refinement of the details may still be necessary).

On the other hand, the previous sections have also shown that especially these 'freaky' processes invite cognitively inspired approaches, i.e., approaches in which core cognitive-psychological notions such as type and token frequencies, similarity, recognition domains etc. play a vital role. This appears to be true on two levels. First, on the level of classification, where, e.g., López Rúa (2002) has proposed considering subtractive word-formation processes as constituting a radial category. Second, on the level of empirical findings simply because the present findings are completely compatible with psycholinguistic models of lexical access such as the cohort model and/or usage-based models incorporating similarity and activation/entrenchment as their governing principles. In fact, I even argue that such studies can even contribute to the development of a morphological classification: It was shown that bottom-up data can support or falsify the distinction between blends and complex clippings that has been proposed in a top-down manner. Also, maybe a refinement of the *RP* analysis along the lines suggested at the end of Section 2 will make it possible to devise a more objective data-driven way to recognize neoclassical compounds.

In sum, as a framework governed by the cognitive commitment, Cognitive Linguistics provides the ideal tools to investigate and explain many mechanisms governing subtractive word-formation and may in turn again benefit from the concepts utilized in such studies and from such explanations.

Received 12 November 2005
Revision received 11 June 2006

University of California,
Santa Barbara

Notes

* I wish to thank Daniel Stahl from King's College, London, UK, for his advice on statistical matters. Also, I thank one anonymous reviewer and in particular the second reviewer, R. Harald Baayen, for a large number of useful pointers and suggestions. The usual disclaimers apply. Author's email address: ⟨stgries@linguistics.ucsb.edu⟩.

1. I leave aside studies that looked at error blends, i.e., unintentional blends resulting from slips of the tongue; cf. MacKay (1973), Berg (1998), Laubstein (1999), and Gries (2004a, c) for more detailed investigations of error blends and how error blends differ from intentional blends.
2. I used the word form files of the CELEX database, i.e. ⟨EFW.CD⟩. The original version of this file was, however, slightly changed before all the analyses reported here were done. First, it was preprocessed to homogenize the coding of the blends and complex clippings and the transcription in the CELEX database (which, for example, involves homogenizing syllabic /m/, /n/, /l/ and /R/). Second, the few source words of blends that were not already part of the database were added. Thus, the figures presented here differ slightly from the original version of the database. However, since these changes were applied to the database as a whole, this does not bias the data in any direction.
3. One reviewer pointed out that this problem could be easily overcome and that "[n]o 'ingenious statistical technique' is required." I agree that a threshold value for the downstep could be easily defined and programmed. However, one problem is that there seem to be no well-established criteria for this threshold (contrary to, say, the similar problem of scree plots in factor analysis, where one could at least resort to the objective and well-motivated threshold of *Eigenvalue* > 1). Thus, one would either have to define that arbitrarily, running the risk of biasing the data in unpredictable ways, or use several different thresholds to attempt to determine a reasonable threshold value in a *post hoc* manner. I therefore am not so confident that this would be as straightforward a way out of this problem as it may seem.
4. That is, for this study I left out cases where the formation contains material not present in either source word or where one source word contributes more than one part of it to a formation.
5. Note that the extremely small standard error (0.031) rules out the theoretically conceivable possibility that approximately 50% of the cut-off points are at $+3$ and the rest at -3. While this situation would result in an average similarly close to 0 as the actually observed one, this would have inflated the standard error extremely, contrary to what is observed here.
6. Since the four-way interaction is insignificant ($F < 1$; $p = 0.375$), there is no difference between letters and phonemes.
7. As R. Harald Baayen points out, the fact that the random cut-off points of sw_2 of blends is comparatively higher may be due to the fact that the sw_2 of blends is on average significantly longer than sw_1 of blends according to a Wilcoxon test ($V = 345{,}522$, $p < 0.0001$; cf. also Gries 2004b for more fine-grained exposition). However, this does still not explain why sw_2 is cut off too early to be discussed below.
8. Readers may wonder whether bigram-based measures actually reflect word similarity well and/or whether something as content-free as bigrams are the right way of operationalization. However, there is some empirical evidence that even information such as letter bigrams at least constitute knowledge that is accessible on some level of the (connections of the) linguistic system; cf. Underwood (1971).
9. Note that the measures have different directionality: For the bigram-based measures and the LCS-based measures, high and low values mean high and low similarity respectively (because high values means a lot of material is shared in the right order) whereas for the ED-based measure high and low values mean low and high similarity respectively (because high values mean many editing operations are necessary).
10. The average intercorrelations of the three groups of measures for each medium (after

	bigram-based	LCS-based	ED-based
letters	0.973	0.957	0.883
phonemes	0.978	0.97	0.88

Table (i). *Intercorrelations between different kinds of measures of word similarity*

Fisher Z transformation with subtracting the constant 0.001 to handle cases where $r = 1$ and re-transformation) are given in Table (i).

References

Algeo, John
1977 Blends, a structural and systemic view. *American Speech* 52(1), 47–64.
1978 The taxonomy of word making. *Word* 29(1), 122–131.

Baayen, R. Harald, Richard Piepenbrock, and Leon Gulikers (eds.)
1995 *The CELEX Lexical Database (CD-ROM)*. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

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

Berg, Thomas
1998 *Linguistic Structure and Change: An Explanation from Language Processing*. Oxford: Oxford University Press.

Brew, Chris and David McKelvie
1996 Word-pair extraction for lexicography. *Proceedings of the Second International Conference on New Methods in Language Processing*, 45–55.

British National Corpus Consortium
2000 *The British National Corpus World Edition*.

Cannon, Garland
1986 Blends in English word formation. *Linguistics* 24(4), 725–753.

Ellis, Nick C.
2002 Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition* 24, 143–188.

Grosjean, François.
1980 Spoken word recognition processes and the gating paradigm. *Perception and Psychophysics* 28(4), 267–283.

Gries, Stefan Th.
2004a Shouldn't it be *breakfunch*? A quantitative analysis of the structure of blends. *Linguistics* 42(3), 639–667.
2004b Isn't that fantabulous? How similarity motivates intentional morphological blends in English. In Achard, Michel and Suzanne Kemmer (eds.), *Language, Culture, and Mind*. Stanford, CA: CSLI, 415–428.
2004c Some characteristics of English morphological blends. In Andronis, Mary, Erin Debenport, Anne Pycha, and Keiko Yoshimura (eds.), *Papers from the 38th Regional Meeting of the Chicago Linguistics Society*, Vol. II: *The Panels*. Chicago, IL: Chicago Linguistics Society, 201–216.

Kaunisto, Mark
2000 Relations and proportions in the formation of blend words. Paper presented at the International Quantitative Linguistics Conference.
Kelly, Michael H.
1998 To 'brunch' or to 'brench': Some aspects of blend structure. *Linguistics* 36(3), 579–590.
Kemmer, Suzanne
2003 Schemas and lexical blends. In Berg, Thomas et al. (eds.), *Motivation in Language: From Case Grammar to Cognitive Linguistics. A Festschrift for Günter Radden*. Amsterdam/Philadelphia: John Benjamins, 69–97.
Kondrak, Grzegorz
2002 Algorithms for language reconstruction. Unpublished Ph.D. thesis, University of Toronto.
2003 Phonetic alignment and similarity. *Computers and the Humanities* 37(3), 273–291.
Kubozono, Haruo
1990 Phonological constraints on blending in English as a case for phonology-morphology interface. *Yearbook of Morphology* 3, 1–20.
Laubstein, Ann Stuart
1999 Lemmas and lexemes: The evidence from blends. *Brain and Language* 68(1), 135–143.
Lehrer, Adrienne
1996 Identifying and interpreting blends: An experimental approach. *Cognitive Linguistics* 7(4), 359–390.
López Rúa, Paula
2002 On the structure of acronyms and neighbouring categories: A prototype-based account. *English Language and Linguistics* 6(1), 31–60.
2004 The categorial continuum of English blends. *English Studies* 85(1), 63–76.
MacKay, Donald G.
1973 Complexity in output systems: Evidence from behavioral hybrids. *American Journal of Psychology* 86(4), 785–806.
Marslen-Wilson, William
1987 Functional parallelism in spoken word-recognition. *Cognition* 25(1), 71–102.
Plag, Ingo
2003 *Word Formation in English*. Cambridge: Cambridge University Press.
R Development Core Team
2005 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical. Computing, Vienna, Austria. http://www.R-project.org.
Sutton, Blair
2005 *Numen Inest Perl 5.8.7.2*. Version of July 17, 2005. http://sourceforge.net/projects/niperl/.
Underwood, Benton J.
1971 Recognition memory. In Kendler, Howard H. and Janet T. Spence (eds.), *Essays in Neo-Behaviorism*. New York: Appleton-Century-Crofts, 313–335.