Quantitative approaches to similarity in cognitive linguistics 1: the phonology of blends

Stefan Th. Gries
Department of Linguistics
University of California, Santa Barbara
http://tinyurl.com/stgries

Blending is a complex process …

- Blending is one of the most perplexing word-formation processes, given that
  - it is not as rule-governed as derivational processes
  - it is not as productive as most derivational processes
  - it is more creative than most derivational processes
  - it involves conscious effort and word play on the part of the coiner, which often
    - results in 'violations' of more rigid morphological rules
    - includes the 'integration' of a many kinds of information that are not central to linguistic study (e.g., the interplay between orthography and punctuation)
  - it nevertheless exhibits superficial similarity to other intentional word-formation processes (e.g., compounding, (complex) clipping, abbreviations, acronyms)
  - it has an unplanned counterpart in the form of speech-error blends
  - …

… and what that may mean for their analysis

- Given the interaction of all these characteristics, it comes as no surprise that some have adopted a somewhat pessimistic stance towards blends
  - "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)
- and it's true: blends involve a mind-boggling degree of complexity, and the kind of (near-)categorical rules and processes we often find elsewhere in morphology are hard to come by …

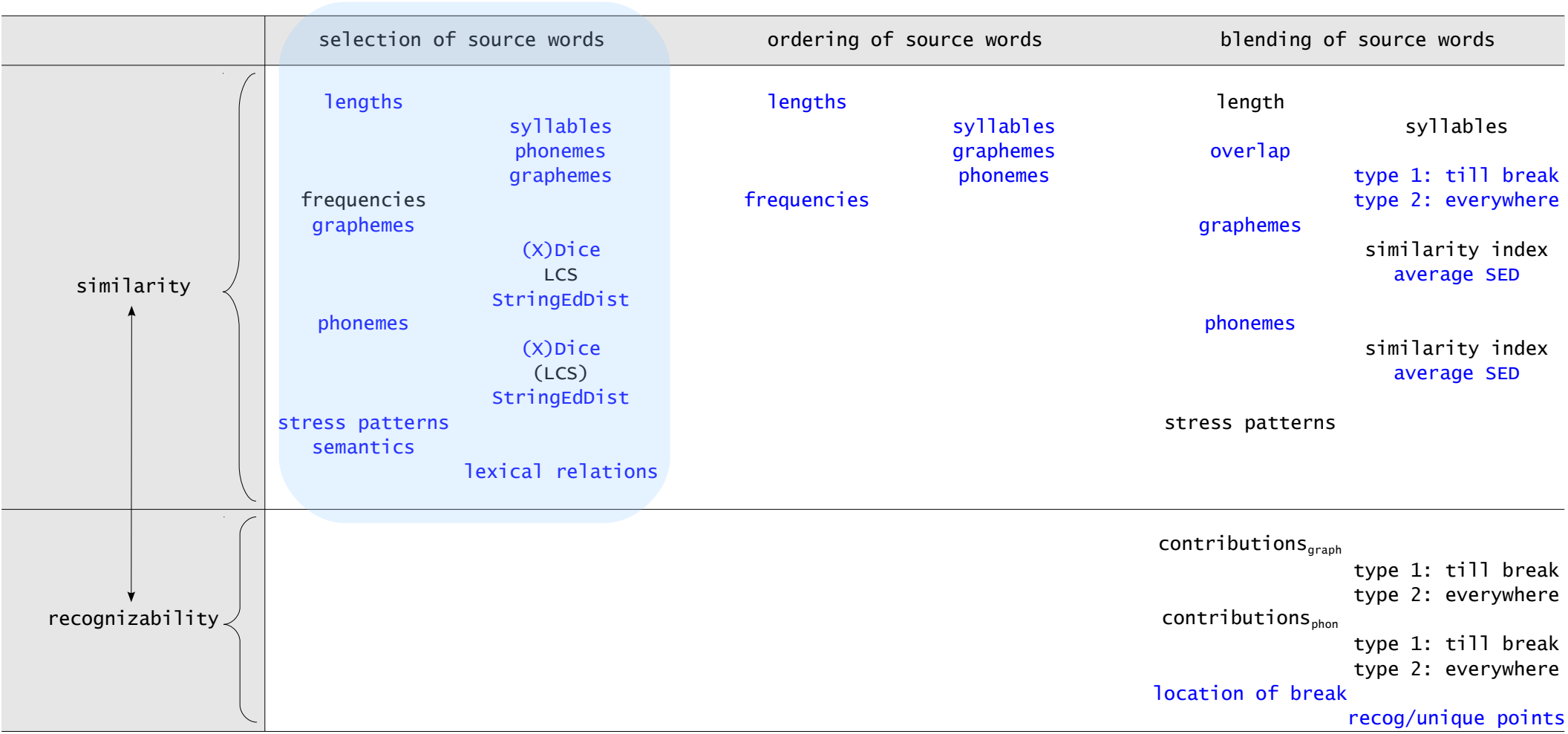
The complexity of blends, and what that may mean for their analysis

- On the other hand, just because blends do not exhibit many, if any, categorical rules does not mean that blends are unpredictable
- in fact, most, if not all, linguistic phenomena are not categorical in nature, but probabilistic and multifactorial – and so are blends
- we should therefore adopt a probabilistic approach to the analysis of blend (structure), but we need
  - larger samples than those studied in some of the classic studies (314 in Pound (1914), 132 in Cannon (1986), …)
  - statistical methods that can handle probabilistic distributions better than intuition/hunches alone
- definition of blend: fusion of 2(+) words where a part of sw1 is combined with a part of sw2, where at least one sw is shortened and/or the fusion may involve overlap of sw1 and sw2

Overview of this talk

- In what follows, I will present several case studies regarding the three 'temporal stages' in blending
  - the selection of two source words
  - the ordering of the source words
  - the blending of the source words
- the case studies involve (non-standard) elements from many different levels of linguistic analysis
  - graphemes and phonemes
  - graphemic and phonemic *n*-grams
  - syllables and their constituents
  - words, their lengths, frequencies (and semantics)
- important methodological considerations
  - intentional blends require comparisons to other formations (other intentional formations, error blends)
  - intentional blends must be tested against baselines
  - successive fine-tuning of methods
- database: 2329 formations, 151,103 data points

Where we are now …



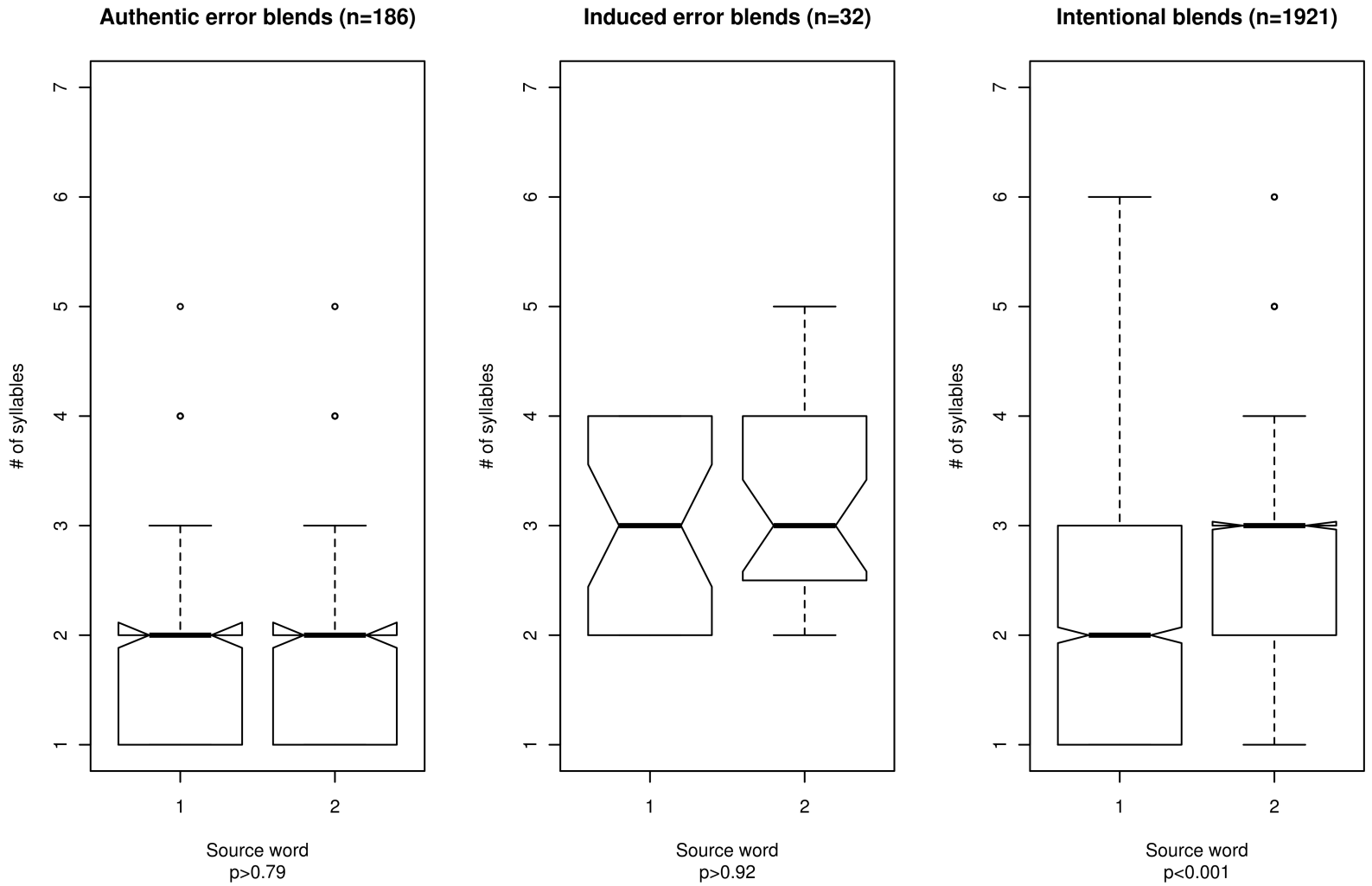
Characteristics of the source words that are chosen to be blended

- Earlier studies have shown or argued that the source words users select to blend are similar to each other
  - this seems to hold for speech-error blends'
    - phonological characteristics (MacKay 1987, Kubozono 1990)
    - syntactic characteristics (MacKay 1987, Berg 1998)
    - semantic characteristics (Levelt 1989, Berg 1998),
  - and for intentional blends (Kubozono 1990, Kelly 1998)
- but there are many different ways words can be similar to each other
  - different ways in which words are similar to each other
    - length (using different units)
    - frequency/dispersion
    - phonemic/graphemic material
    - stress patterns
    - semantics
  - different places where words are similar to each other
  - different ways to measure all these similarities

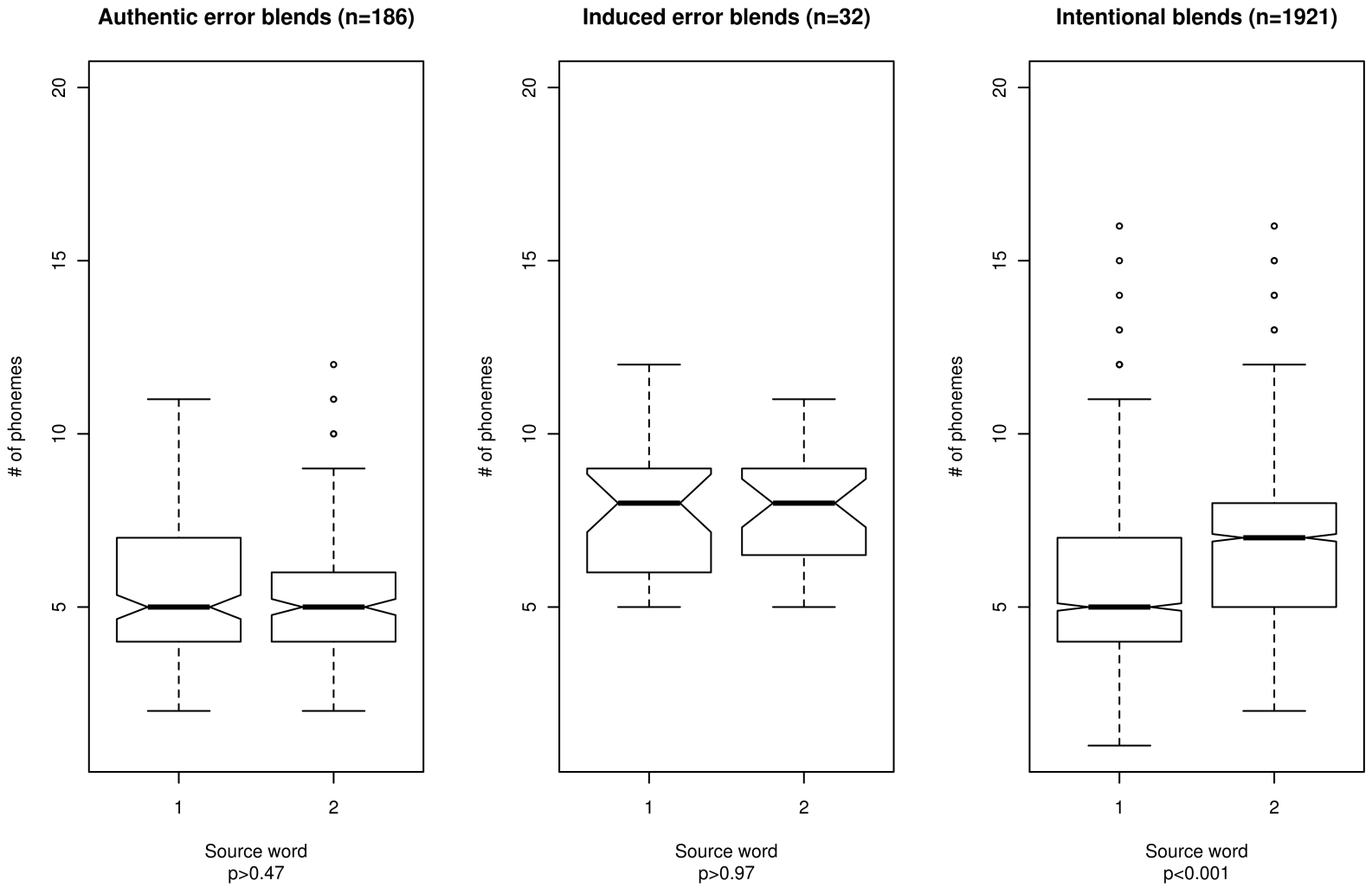
Comparing the lengths of source words

- For the source words of forms in my database, I determined their lengths, i.e.
  - their lengths in syllables
  - their lengths in phonemes
  - their lengths in graphemes
- in addition, for each form, I know its formation type, i.e. whether it's
  - an authentic error blend
  - an induced error blend
  - an intentional blend
  - some other kind of formation (e.g., a complex clipping)
- then, the lengths of source words were compared with the types of blends (excluding, for now, other formations)

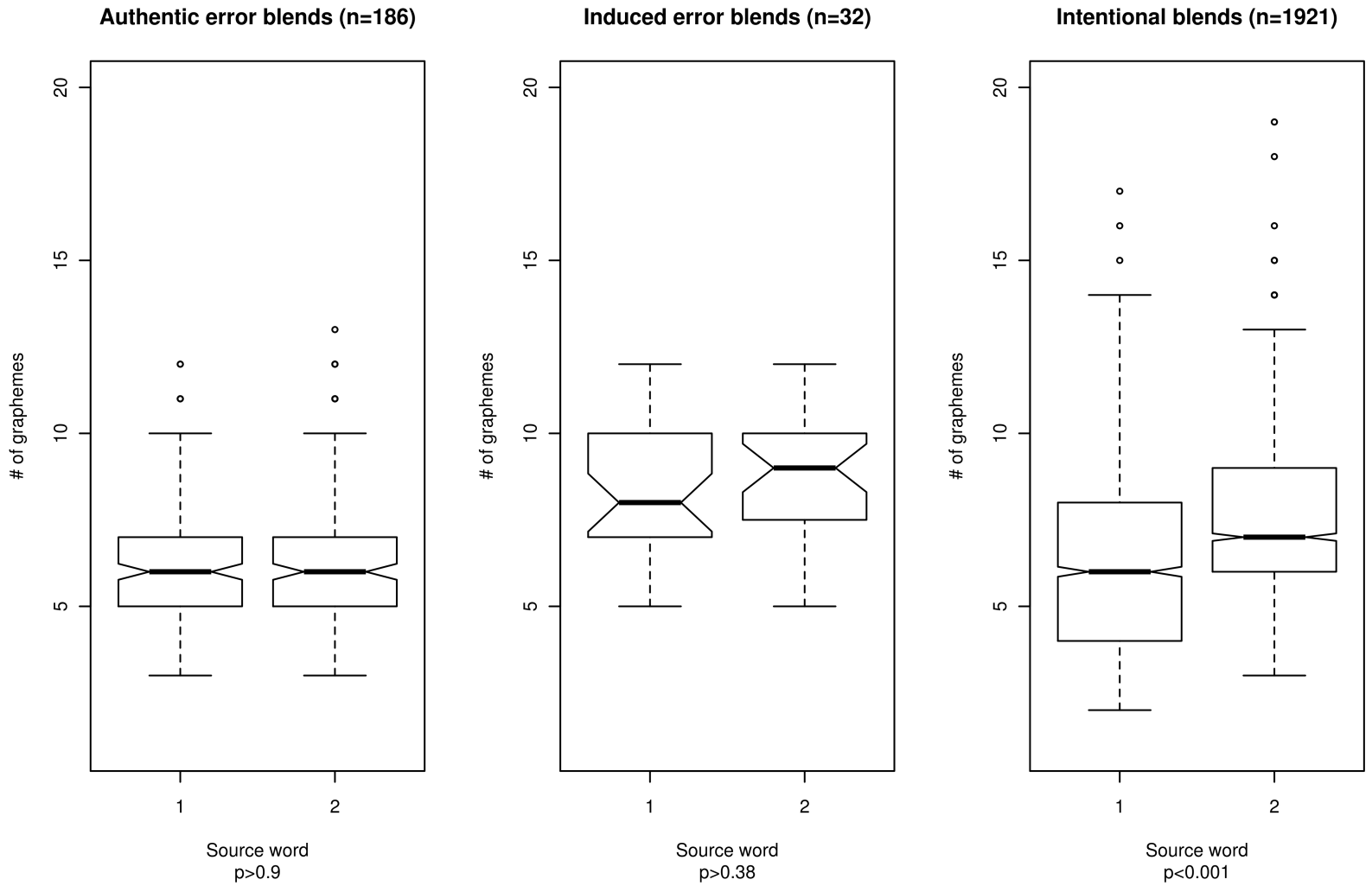
Blend type × source word → syllabic length



Blend type × source word → phonemic length



Blend type × source word → graphemic length

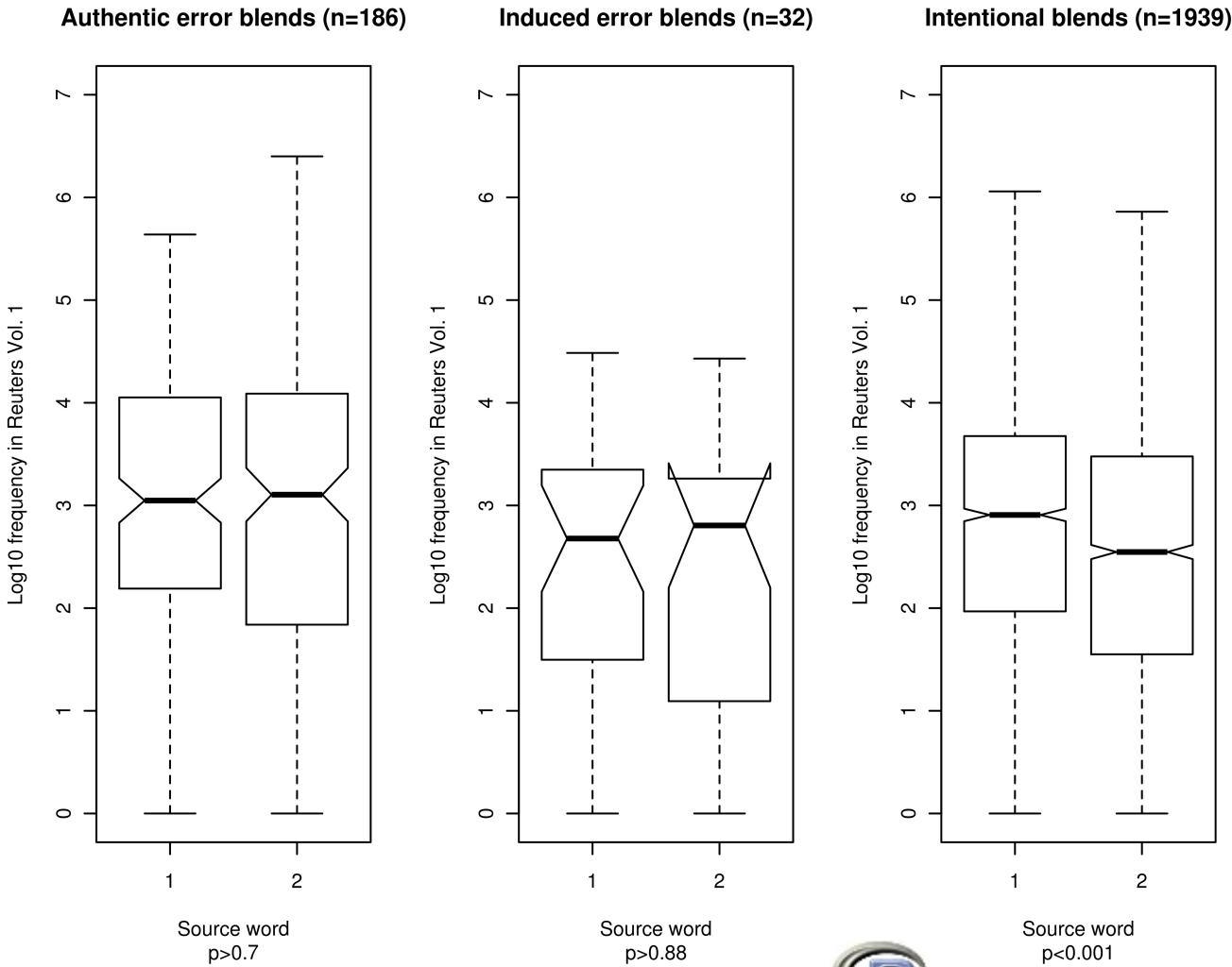


Interim summary

- The source words of
  - error blends are short and not different from each other
  - induced error blends are much longer but not different from each other
  - intentional blends are differently long
    - sw1 is short
    - sw2 is significantly longer
- this shows two things
  - the source words of error blends are different in kind from the source words of intentional blends
  - findings from induced error blends may have to be taken with a grain of salt since their source words are very different from blends from authentic settings
- note: these are not pairwise comparisons (yet)

Blend type × source word → frequency

- I computed the frequencies of all source words of all blends in the Reuters corpus and compared
  - error blends (ns)
  - induced errors (ns)
  - intentional blends: sw1>sw2 (***)
- (same for ranges)



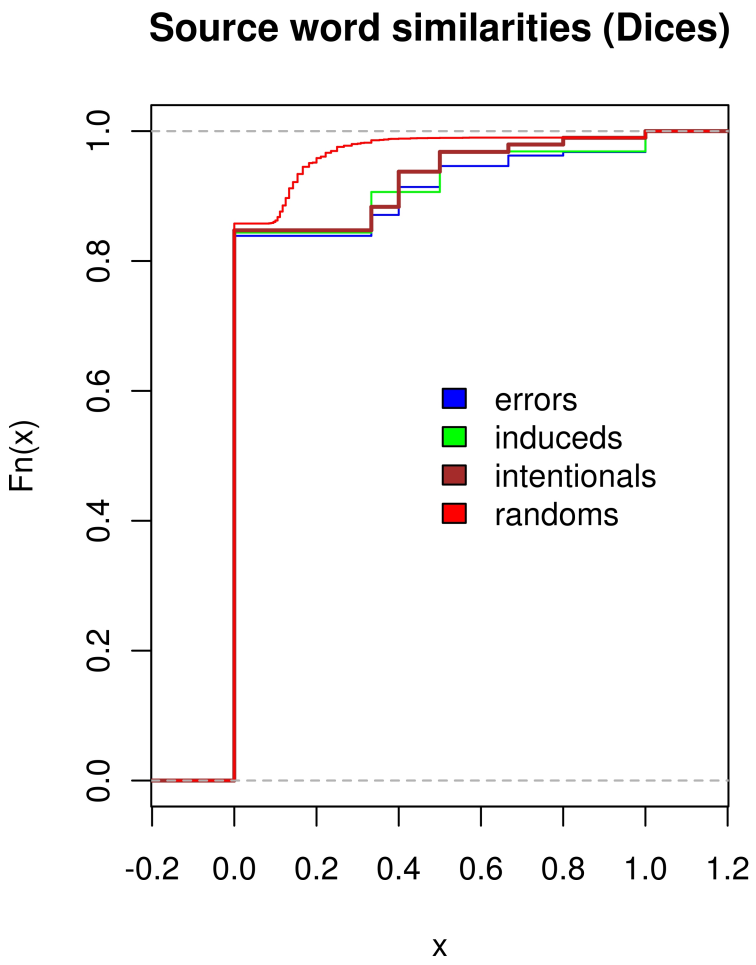
Comparing the elements of source words

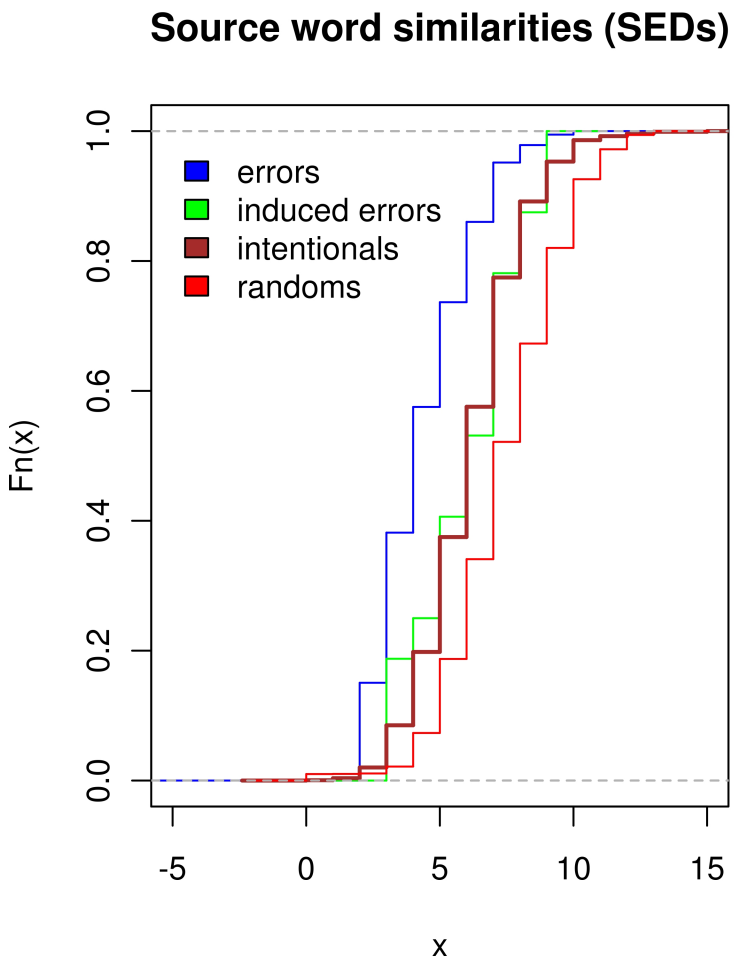
- How similar are source words of blends to each other when we compare
  - error (authentic and induced) to intentional blends
  - attested blends to randomly chosen words
- how do we measure similarity?
  - Dice: the percentage of (any type of) shared bigrams
    - *channel*: *ch ha an nn ne el*
    - *tunnel*: *tu un nn ne el*

    $6/11 = 0.55$
  - string edit distance (Levenshtein)
    - *channel* → *tunnel*
      - delete the *c*
      - replace the *h* by a *t*
      - replace the *a* by a *u*

      $= 3$
- data: (phonemic descriptions of) source words from
  - 186 error blends
  - 32 induced error blends
  - 1939 intentional blends (excl. complex clippings)

Blend type × similarity type (vs. random baselines)





Interim summary

- Focusing on string edit distances only,
  - the source words of all types of blends are more similar to each other than expected from chance alone
  - the source words of induced errors are not significantly more similar to each other than the source words of intentional blends
    - one the one hand, that makes sense: both types of source words were chosen intentionally
    - on the other hand, it again shows that findings from induced error blends may not lend themselves to generalizations as easily as was expected
- note also
  - the source words of error blends are globally similar to each, but
  - the source words of intentional blends are most similar to each other around the cut-off point
    we'll return to that below

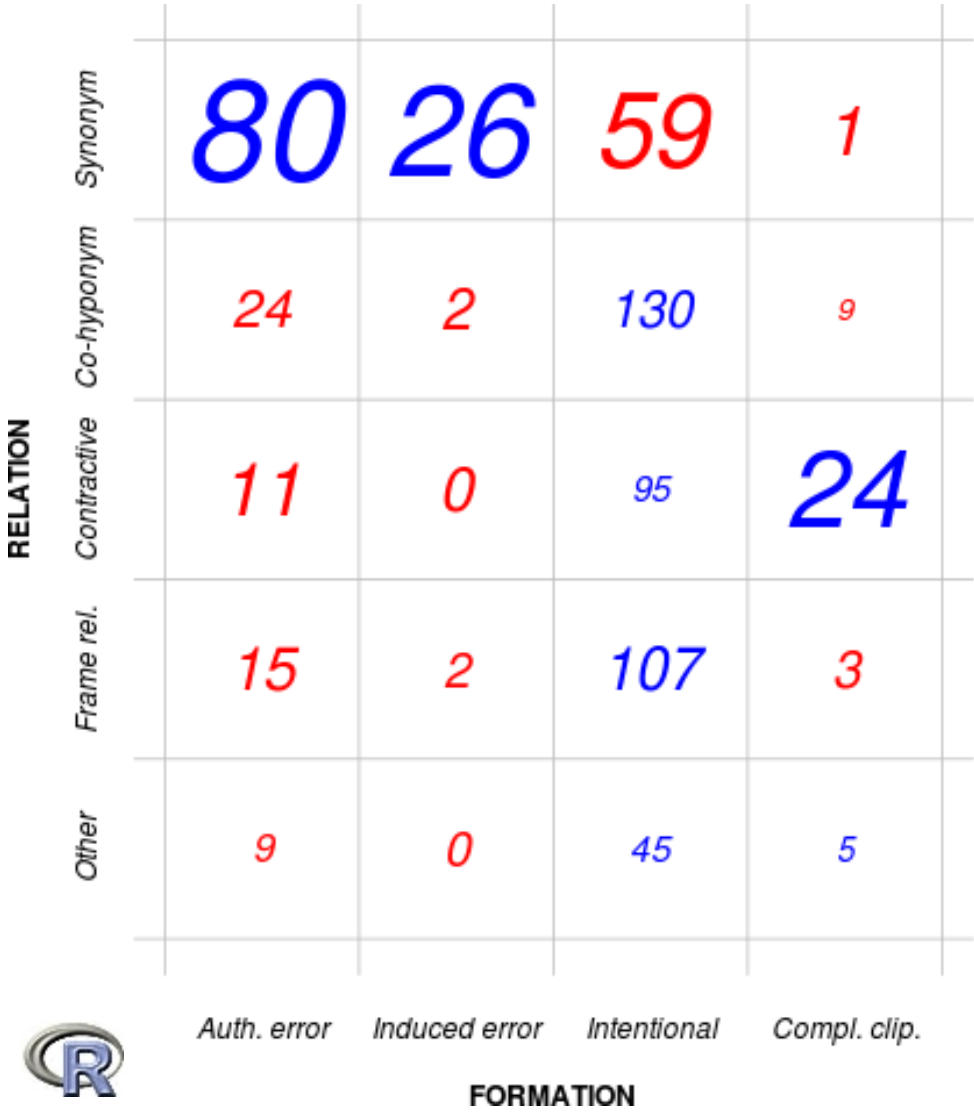
Comparing the stress patterns of source words

- How similar are source words of blends to each other when we compare their stress patterns (within groups of identically long source words)?
- 2,139 formations were coded for
  - the syllabic lengths of both source words
  - the stress patterns of the words, i.e., whether each syllable was stressed or unstressed (only primary stress)
- for example,
  - *webinar*
    - *web*: 1: s and *seminar*: 3: suu
  - *jokelore*
    - *joke*: 1: s and *folklore*: 2: su
  - *transponder*
    - *transmission*: 3: usu and *responder*: 3: usu
- method: cross-tabulation plot (Gries 2009)

	s	su	us	suu	usu	uus	suuu	usuu	uusu	uuus	suuuu	usuuu	uusuu	uuusu	uusuuu	uuusuu	uuuusu
s	148	189	35	159	81	10	25	43	44	0	2	0	13	6	1	1	0
su	72	164	22	100	60	11	22	21	48	1	0	1	5	4	1	4	2
us	10	12	12	17	9	2	0	4	5	1	0	0	1	1	0	0	0
suu	60	88	12	67	31	1	9	15	28	2	0	1	5	6	0	0	1
usu	23	33	3	13	29	3	9	6	9	0	0	2	2	1	0	0	0
uus	2	6	1	3	1	0	1	1	0	0	0	0	0	1	0	0	0
suuu	9	17	2	12	4	1	6	4	4	0	0	0	2	1	0	0	0
usuu	16	17	3	11	9	0	1	13	11	1	0	1	0	0	0	0	0
uusu	14	20	1	10	20	0	2	5	14	0	0	0	2	2	0	0	2
uuus	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
suuuu	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
usuuu	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
uusuu	5	9	0	4	3	1	1	3	2	0	0	0	1	0	0	0	0
uuusu	4	3	0	1	4	1	0	0	8	0	0	0	0	0	0	0	0
uusuuu	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
uuusuu	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
uuuusu	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0

Blend type × the source words' semantic relation

- What about the semantic relation between the source words? 647 forms were analyzed (preliminarily, this is work-in-progress!)
- four semantic relations account for 90% of the data
- three very clear classes emerge (X^2=211.07, df=12, p<0.001, V=0.33)
  - error blends: synonyms
  - intentional blends: all categories are frequent
  - complex clippings: contractives

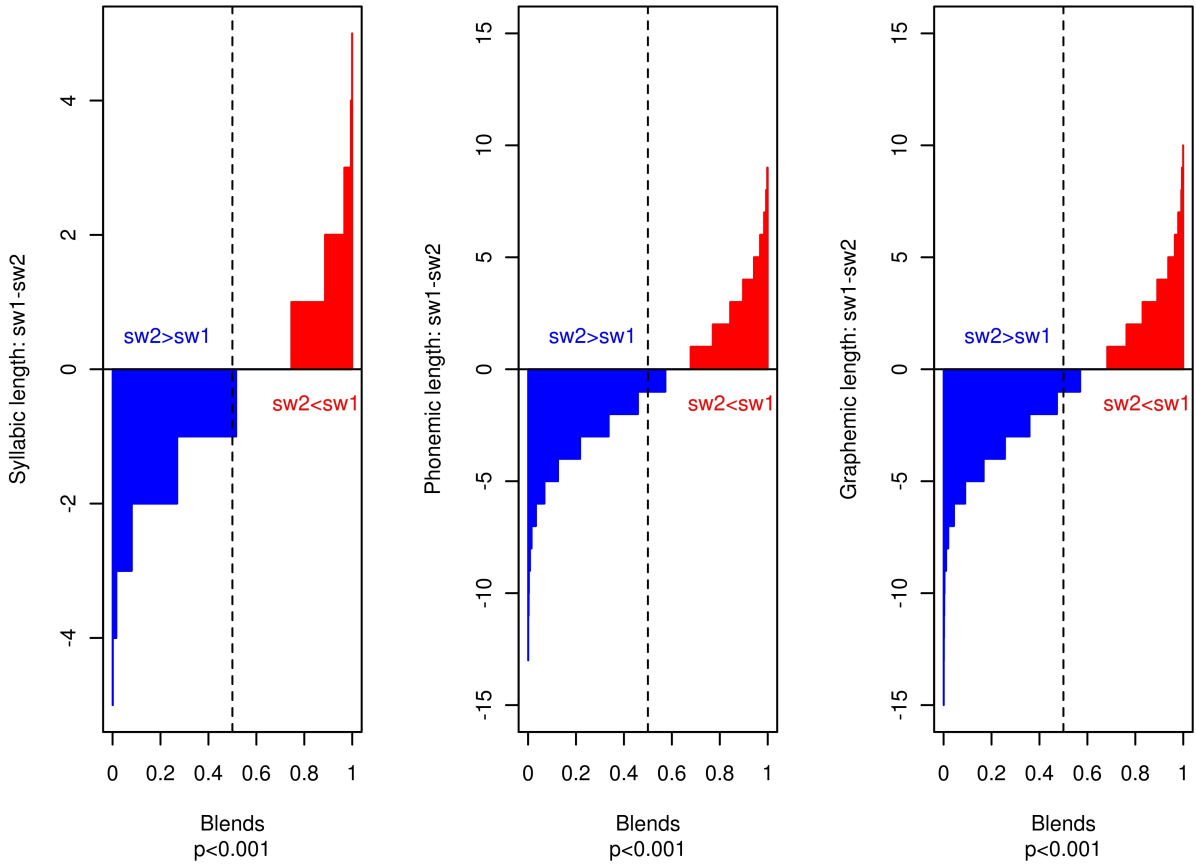


Where we are now …

	selection of source words	ordering of source words	blending of source words
similarity	lengths syllables phonemes graphemes frequencies graphemes (X)Dice LCS StringEdDist phonemes (X)Dice (LCS) StringEdDist stress patterns semantics lexical relations	lengths syllables graphemes phonemes frequencies	length syllables overlap type 1: till break type 2: everywhere graphemes similarity index average SED phonemes similarity index average SED stress patterns
recognizability			$contributions_{graph}$ type 1: till break type 2: everywhere $contributions_{phon}$ type 1: till break type 2: everywhere location of break recog/unique points

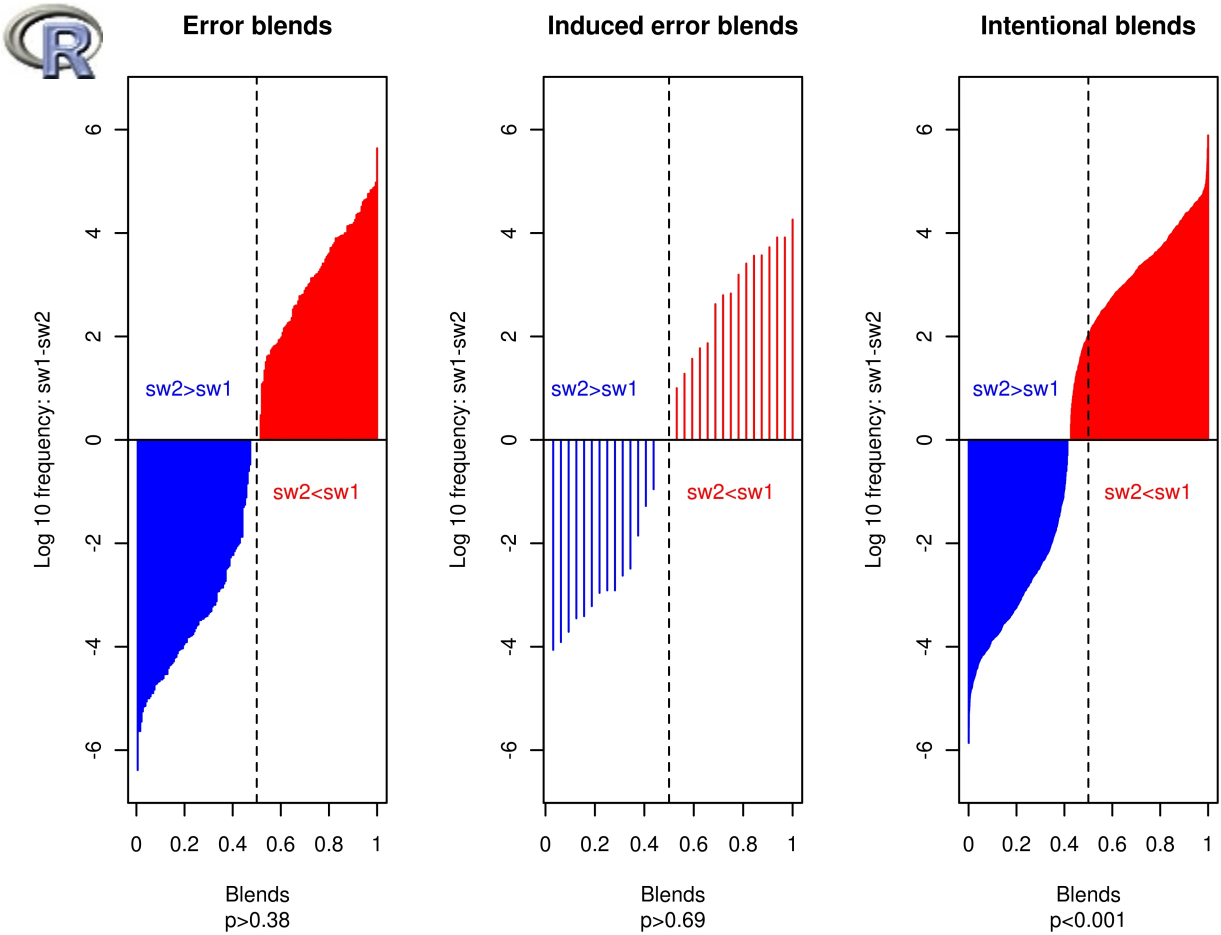
Blend type × source word → lengths (pairwise)

- The above data have shown that the two source words are similar in general, but also exhibit some differences, in particular with intentional blends
- but the above didn't discuss differences in a by-blend nature, i.e., pairwise differences
- however, everything stays the same
  - error blends: no signi-ficant difference
  - intentional blends: sw1<sw2



Comparing the frequencies of source words

- The above data have shown that the two source words are equally frequent with error blends, but not with intentional blends
- but the above didn't discuss differences in a by-blend nature, i.e., pairwise differences
- however, everything stays the same
  - error blends (both): no significant difference
  - intentional blends: sw1>sw2
- (same for range)



Where we are now …

	selection of source words	ordering of source words	blending of source words
similarity	lengths syllables phonemes graphemes frequencies graphemes (X)Dice LCS StringEdDist phonemes (X)Dice (LCS) StringEdDist stress patterns semantics lexical relations	lengths syllables graphemes phonemes frequencies	length syllables overlap type 1: till break type 2: everywhere graphemes similarity index average SED phonemes similarity index average SED stress patterns
recognizability			$contributions_{graph}$ type 1: till break type 2: everywhere $contributions_{phon}$ type 1: till break type 2: everywhere location of break recog/unique points

The similarity of the source words to the blend

- In previous work, I argued that blends are coined under the influence of two opposing (!) factors
  - recognizability
    - I don't think recognizability is ill-defined
    - and I don't think it's the "key" thing
  - similarity
- why do these counteract each other (to some degree)?
  - the two source words would be most recognizable when nearly all of their material was present in the blend:
    Chevrolet ∞ *Cadillac* → *Chevrolecadillac*
    Caddillac ∞ *Chevrolet* → *Cadillachevrolet*
  - but this is not fun anymore: while both source words are perfectly recognizable, the blend is not similar to either source word anymore
- ideally, we would have a way to quantify the degree to which a blend strikes a balance between these two forces …

A first operationalization

- In previous work, I argued for a similarity index $SI_{G/P}$, to be computed as follows

$$\left(\frac{\text{number of units of sw}_1\text{ into blend}}{\text{number of units of sw}_1} \times \frac{\text{number of units of blend out of sw}_1}{\text{number of units of blend}}\right)$$

$$\left(\frac{\text{number of units of sw}_2\text{ into blend}}{\text{number of units of sw}_2} \times \frac{\text{number of units of blend out of sw}_2}{\text{number of units of blend}}\right)$$

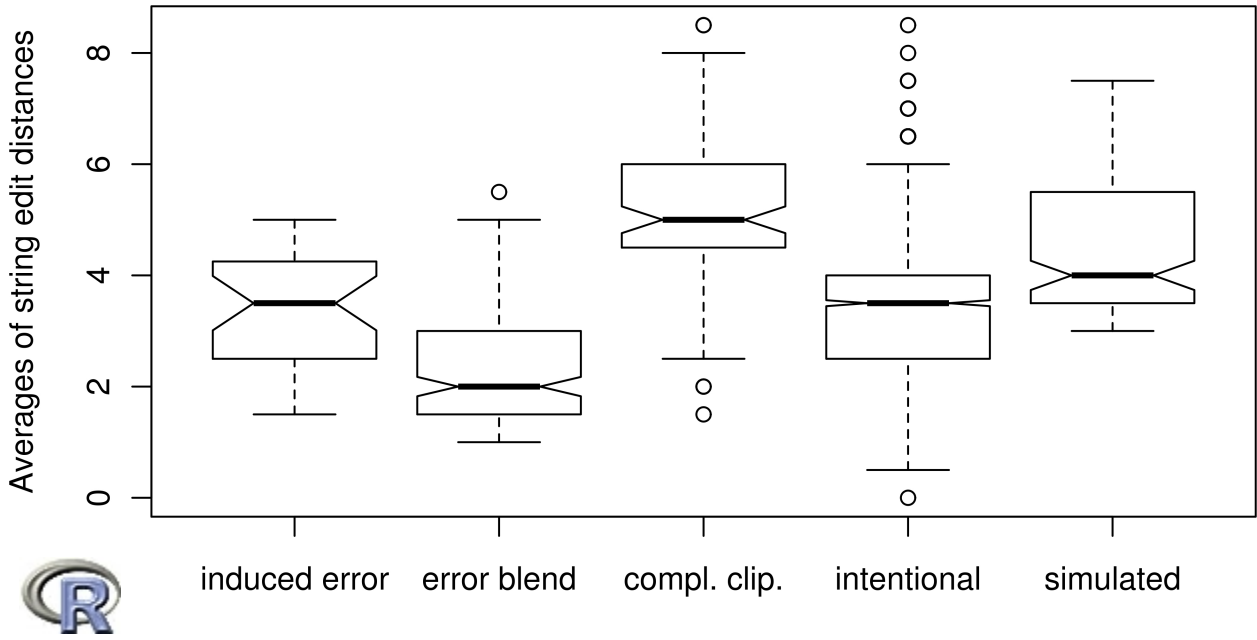
- that is, $^6/_7$ letters of *channel* make up $^6/_7$ letters of *chunnel*, and $^5/_6$ letters of *tunnel* make up $^5/_7$ letters of *chunnel*; SI_G *chunnel* = 0.665
- this seemed like a good idea at the time
  - compare to SI_G *brunch* = 0.304 and SI_G *breakfunch* = 0.36
- but …
  - SI_G *Chevrolac* = 0.414 and SI_G *Cadillet* = 0.344
  - SI_G *Chevrolecaddillac* = 0.472 and SI_G *Cadillachevrolet* = 0.531 ??!!

A second, better operationalization

- Obviously, the proposed SI captures some of what's going on in terms of recognizability …
- … but it doesn't penalize blends enough which are too dissimilar to their source words
- better measure: the average of the Levenshtein string edit distances (ASED) between both source words and the blend
- for example
  - ASED for *chunnel* = 1.5
  - ASED for *Chevrolac* = 3.5
  - ASED for *Cadillet* = 3.5
  - ASED for *Chevrolecadillac* = 8
  - ASED for *Cadillachevrolet* = 7.5
- ASED is also preferable since we can then use the same type of measure for both comparisons: source word to source word *and* source words to blends

The similarity of source words to blends

- Data
  - 218 error blends (186 authentic, 32 induced)
  - 1940 intentional blends and 97 complex clippings
  - 144 simulated blends
- the ASEDs show a few interesting things
  - error blends are most similar to their source words
  - blends from intentional source words are less similar
  - complex clippings are *very* different from their sources, in fact, more different than even average simulated blends



Can we say where source words get split up?

- We have seen a variety of results regarding the choice, ordering, and blending of words
  - sw1 is similar to sw2
  - sw1 is shorter than sw2
  - sw1 contributes less of itself than sw2
- but can we also determine what governs the exact choice of a cut-off point?
- useful theories in this context
  - activation-based models (e.g., McClelland & Rumelhart)
  - computational search models (e.g., Forster)
  - the cohort model (Marslen-Wilson)

Recognition points: what they are

- Uniqueness point (UP): the point P at which a word W can be uniquely identified from a candidate set of words
- Recognition point (RP) the empirical estimate of UP
  - the point P within a word W at which a majority of speakers (e.g., 85%) can recognize W when presented with parts of W with (e.g., 80%) confidence
- RPs exhibit a word-frequency effect of tokens: more frequent words are recognized faster (by ≈20%) than their closest competitors

Recognition points: how to get them

- RPs are usually obtained experimentally, but a corpus-based approach is also conceivable
- for example
  - in the British National Corpus, the string "islamiciza" has only one possible continuation: "tion"
  - in the CELEX database, the string [ɪslɑmɪsaɪzeɪ] has only one possible continuation: [ʃən]
- a corpus-based approach can take into consideration the size and the information distribution of the candidate set
  - what is the type frequency of the candidate set?
  - what are the token frequencies of these types?
- but how would one approach cut-off points of blends in relation to RPs?

Log 10 type frequency
0
1
2
3
4
a
ag
agi
agit
agita
agitat
agitati
agitatio
agitation
Parts of 'agitation'

Log 10 token frequency
0
1
2
3
4
5
6
a
ag
agi
agit
agita
agitat
agitati
agitatio
agitation
Parts of 'agitation'

Problems with this approach

- The database is too large to allow for a manual identification of cut-off points on the basis of such plots (≈ 9000 plots)
- the manual identification of such cut-off points is often far from straightforward because many graphs don't exhibit such neat dents
- this method looks only at the type frequency of a particular word beginning – it doesn't include the token frequency / information distribution

distribution	unit	a	ab
uninformative	types	250:250:250:250	25:25:25:25
	token	1000	100
informative	token	1000	100
	types	975:22:2:1	95:3:1:1

A better approach: selection points (SP)

- A better approach: we choose the first point after a part of a word W at which W is the most frequent word with that part …

Part of *agitation*	types w part	rank of *agitation*
a	4347	595
ag	137	24
agi	12	1
agit	8	1
…	…	…

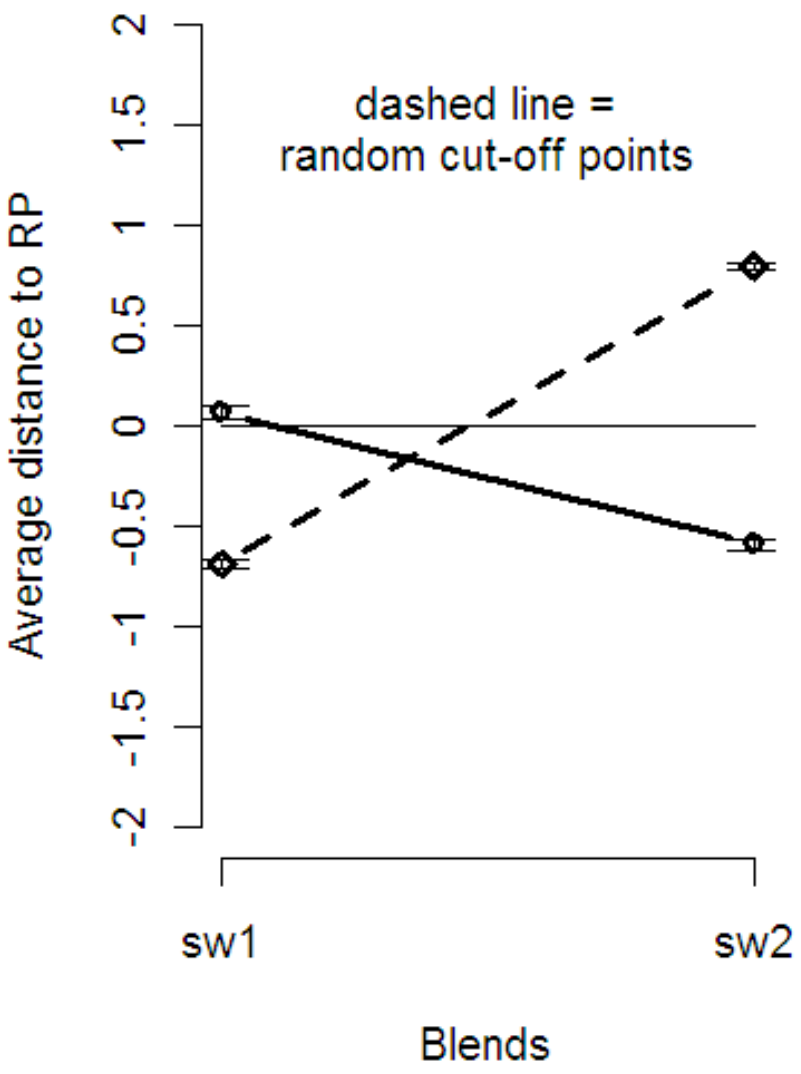
- this approach is very conservative (as yet)
  - it requires exact matches
  - it uses only the minimum value
- but how do we test whether whatever we find is a significant result?
- we need a baseline against which differences between cut-off points and SPs can be evaluated

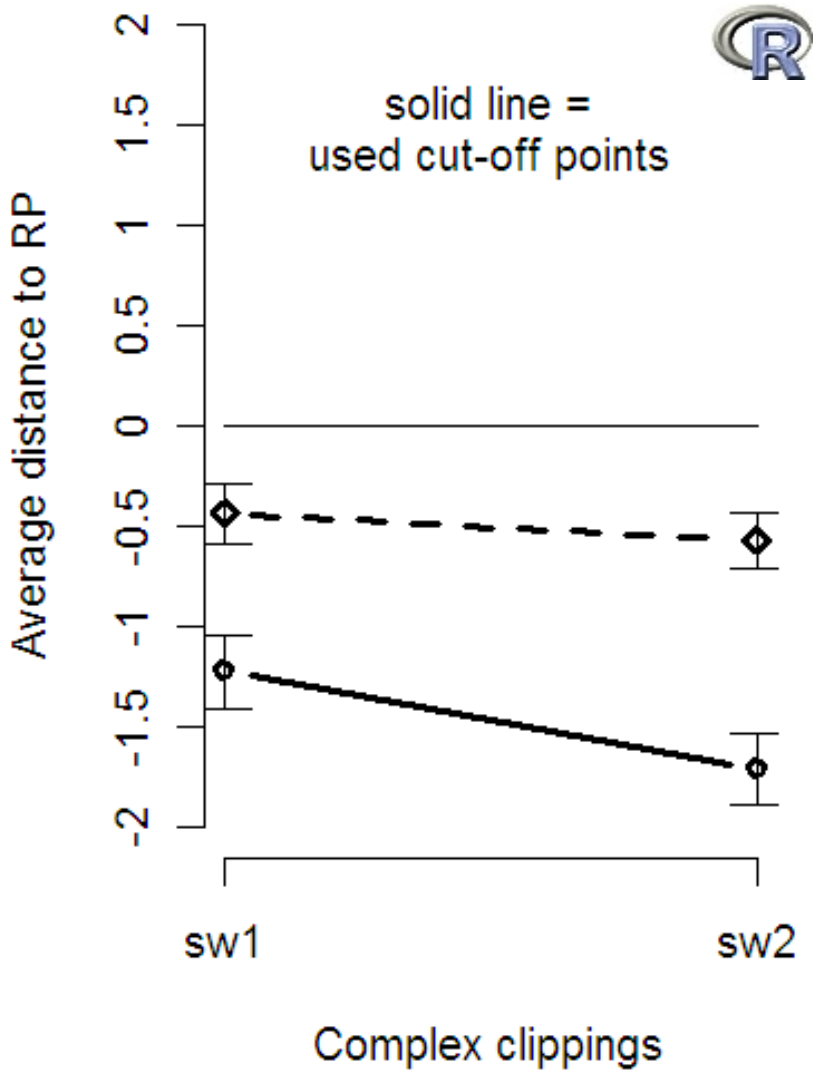
A better approach: the baseline

letters of source word	*a*	*g*	*i*	*t*	*a*	*t*	*i*	*o*	*n*
$rank_{source\ word}$ of all types	595	24	1	1	1	1	1	1	1
distance to ideal cut	-2	-1	0	1	2	3	4	5	6
$median_{distances}$ to ideal					↑				
chosen cut-off point				↑					

- The coiner has chosen a cut-off point that
  - is closer to the hypothesized ideal SP than a random average cut-off point
  - cuts off the first source word only after the SP
- similar comparisons can be made
  - for the first source word of all blends
  - for the second source word of all blends (in the opposite direction)
  - for the source words of all complex clippings
  - on the basis of phonemes and graphemes

Formation type × source word × data type





Formation type × source word × data type

- For blends,
  - the chosen cut-off point is closer to the hypothesized ideal cut-off point, the SP, than the average cut-off point expected by chance
    - sw1: split up nearly exactly at the SP
    - sw2: split up 0.5 elements before the SP
- for complex clippings,
  - the chosen cut-off point is much further away (earlier) than the average cut-off point expected by chance
- the processes exhibit *clear* differences with regard to how their source words are split up
- why are the results less than perfect?
  - contextual effects: context (of all types!) facilitates the recognition so sw2 can be split up earlier and still be recognized
  - morphemic contributions of neo-classical compounds may distort the picture (*tele-*, *-thon*, …)

How do intentional blends happen? A not quite serious interim summary

- Selection: a speaker chooses two source words which
  - are similar in terms of their lengths, syllables, stress patterns, phonemes, and graphemes (especially in middle)
  - are often in a close semantic relationship (relatedness)
  - fit what's to be said (funnily)
- ordering: a speaker
  - leaves the source words in the modifier-head order they come in some phrase, or
  - establishes such a structure or other, and/or
  - puts the shorter and more frequent word first
- blending: a speaker
  - cuts the two source words up close to their RPs
  - fuses them, using more of sw2, so as to
    - maximize overlap in the middle fusion section
    - maximize phonemic/graphemic similarity mappings elsewhere
    - create a blend that's more similar to sw2 / the blend's head
- intentional blends are very different from both errors and complex clippings

Now, why study morphological processes involving conscious effort(s)?

- There are probably many who would consider the study of such messy, limited and non-productive, creative, and conscious processes not particularly revealing
- after all, are we usually not interested in the unconscious working of the linguistic system?
- yes, but
  - as mentioned before, it's not like most of language is neatly categorical anyway
  - even freak processes like blending have to tap into the same linguistic system and are subject to many of its constraints
  - blending can tell us about how conscious/intentional processes apply to, or interact with, the subconscious part of the system
  - sometimes, what is a conscious and fun freak process at the point of time x can affect processes at a much later point of time

Similar fun/psycholinguistic factors at work: lexicalization 1

- In some recent work, I studied
  - 211 lexically fully-specified V-NP idioms
    - *kick the bucket*
    - *run the risk*
    - *lose one's cool*

    (some of these can be modified)
  - 5831 lexically partially specified *way*-constructions
    - *… make your way to the stage …*
    - *… find your way to the hall …*
    - *… fight his way through the crowd …*
- strangely enough, both of these types of constructions exhibit an interesting phonological patterning, and just like blending, …
- this patterning is compatible with explanations involving fun and psycholinguistic models of language production and comprehension

Similar fun/psycholinguistic factors at work: lexicalization 2

- Results: the V-NP idioms and the *way*-constructions involved alliterations highly significantly more often than expected by chance: there is a preference to have multi-constituent symbolic units that exhibit phonological similarity
- why is that? this patterning may be
  - priming …
  - phonological constituents …
  - lexicalized word play …
- similar questions to what we look at in blends, and similar methods, and even the kinds of necessary improvements are similar …

Future work

- Necessary improvements: we need
  - larger collections of blends (duh)
  - more comprehensive description (duh again)
  - better methods: more comprehensive/flexible measures of word similarity to detect similarity on different levels of precision
    - re segments, *channel* is identical to *tunnel*
    - re phonemes, speakers may vary between [ə] and [ɪ], which makes the analysis of *impostinator* trickier
    - re articulatory features, the [tʃ] in *channel* is not identical to the [t] in *tunnel*, but it is similar
    - this will benefit
      - the comparison of sw1 to sw2
      - the identification of the contributions of sw1 and sw2
      - the identification of (ideal) SPs
      - the comparison of sw1 and sw2 to the blend
    - first easy steps: Levenshtein distances with mapping (e.g., to handle segment (clusters)) and/or Damerau weighting

Future work

- Necessary improvements: we need
  - better methods: better measures of frequency/dispersion to determine whether the simplistic frequency differences are robust
  - more psycholinguistics, both in terms of methods (e.g., neighboorhood density) and as an explanatory approach
  - more experimentation to determine whether speakers coin the car brand *Chevrolac* or *Cadillet* or …
  - a more flexible approach to the taxonomy/classification of word-formation processes (maybe a prototype approach of the type argued for by López Rúa)
  - measures of semantic similarity, etc., etc.
- but one thing is already safe to say: blends are far from unpredictable and their characteristics
  - are identifiable using larger databases, reference corpora, statistical techniques (+ baseline comparisons)
  - are firmly grounded in cognitive, but ultimately psycholinguistic and probabilistic mechanisms

Thank you!

http://tinyurl.com/stgries