Quantitative approaches to similarity in cognitive linguistics 2: the phonology of idioms

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

Units in Cognitive Grammar

- The central concept in Cognitive Linguistics is the unit, which in Langacker's Cognitive Grammar is defined as follows (1987:57):

  a structure that a speaker has mastered quite thoroughly, to the extent that he can employ it in largely automatic fashion, without having to focus his attention specifically on its individual parts for their arrangement […] he has no need to reflect on how to put it together.

- units can exhibit different degrees of complexity
  - morphemes or monomorphemic words
  - polymorphemic words
  - fully-fixed multi-word expressions
  - partially filled multi-word expressions
  - syntactic / argument structure constructions

Units in Cognitive Grammar and (relations between/within) their poles

- In Cognitive Grammar, symbolic units are
  - conventionalized associations of a
  - phonological pole and a
  - semantic pole
- relations between these parts of a unit can be looked at in two ways
  - relations between the semantic and the phonological pole: between-pole relations, which are *not* my topic (cf. arbitrariness, motivation, iconicity, …)
  - relations within one pole: within-pole relations
    - there has been much work on semantic within-pole relations
    - there has been less work on phonological within-pole relations
    - but sometimes these surface in surprisingly clear ways: there are a lot of alliterations in idioms with *to run* (cf. Gries 2006)

What I am going to talk about today …

- Two different case studies at two different levels of specificity of units (more constructions below)
  - V-NP idioms (fully lexically filled)
    - *kick the bucket*
    - *run the risk*
    - *lose one's cool*, …
      (these can sometimes be further modified)
  - the *way*-construction (partially lexically filled)
    - *make your way to the stage*
    - *find your way to the hall*
    - *fight his way through the crowd*, …
- I will
  - test how much these constructions also exhibit alliteration effects
  - compare those against different random baselines
  - compare those again non-conventionalized counterparts
  - correlate those with collocational/collostructional attraction

Where to count alliterations

- The data consist of all V-NP idioms listed in the Collins Cobuild Dictionary of Idioms (1995) where
  - the V is a full lexical verb (not an auxiliary)
  - the NP is the direct object of the V
  - the V takes no further complements/adjuncts
  - the idiom occurs at least once per 2m words in the corpus on which the dictionary is based
- some examples
  - *spill the beans*
  - *gain some ground*
  - *get the boot*
  - *lend a hand*
  - *bite the bullet*, …

  (Thanks to Stefanie Wulff for her making these data available to me!)

How to count alliterations

- For each of the V-NP_{DirObj} idioms, I noted
  - the initial segment of the verb
  - the initial segment of the head noun of the NP_{DirObj}
    - *build bridges* → b b
    - *lose face* → l f …
  - if the NP_{DirObj} also involves additional content words, I also noted the initial segments of these words
    - *fight a losing battle* → f l f b l b
    - *keep a straight face* → k s k f s f …
  - the pronunciations of the above words were taken from the CELEX database (Baayen et al. 1995)
- then I computed the percentage of alliterations
- but: we need (a) baseline(s)

How many alliterations to expect: one type of baselines

- There are several ways of obtaining such expected baseline frequencies: on the basis of word-initial phonemes
  - without regard to type and token frequencies
    - this baseline is based on the number of phonemes that words in the CELEX database begin with
  - by taking into consideration their frequencies in differently frequent word types
    - this baseline is based on the probabilities that phonemes are the first phonemes in word types in the CELEX database
  - by taking into consideration their frequencies in word tokens
    - this baseline is based on the probabilities that phonemes are the first phonemes in word tokens in the CELEX database

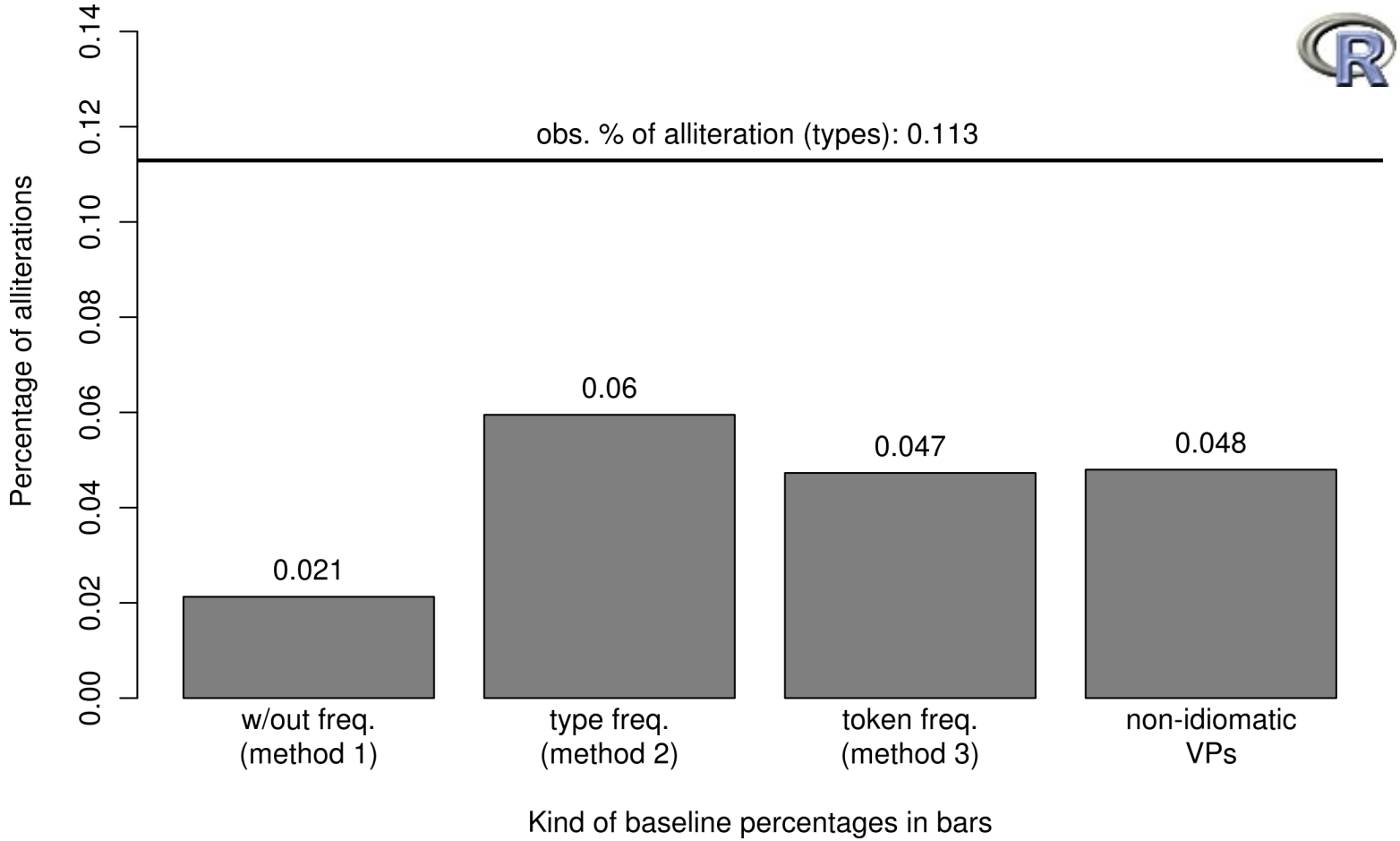
How many alliterations to expect: another type of baseline

- As a kind of control group, I
  - randomly sampled two transitive clauses from each of the 170 ICE-GB files whose names begin with S1[AB]
  - counted alliterations in the same way as before
    - first phoneme of the verb
    - first phoneme of the head noun of the direct object
    - first phonemes of additional content words

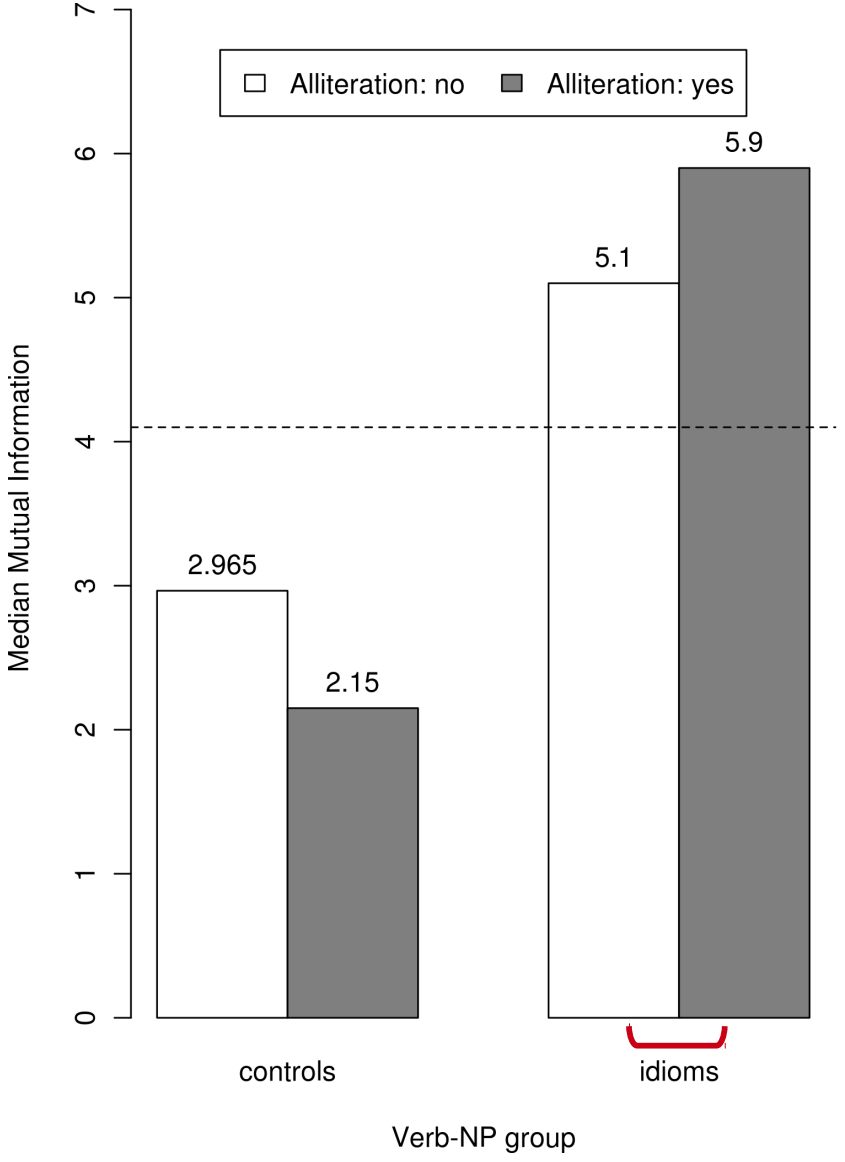
Do the verb and the head noun 'like' each other?

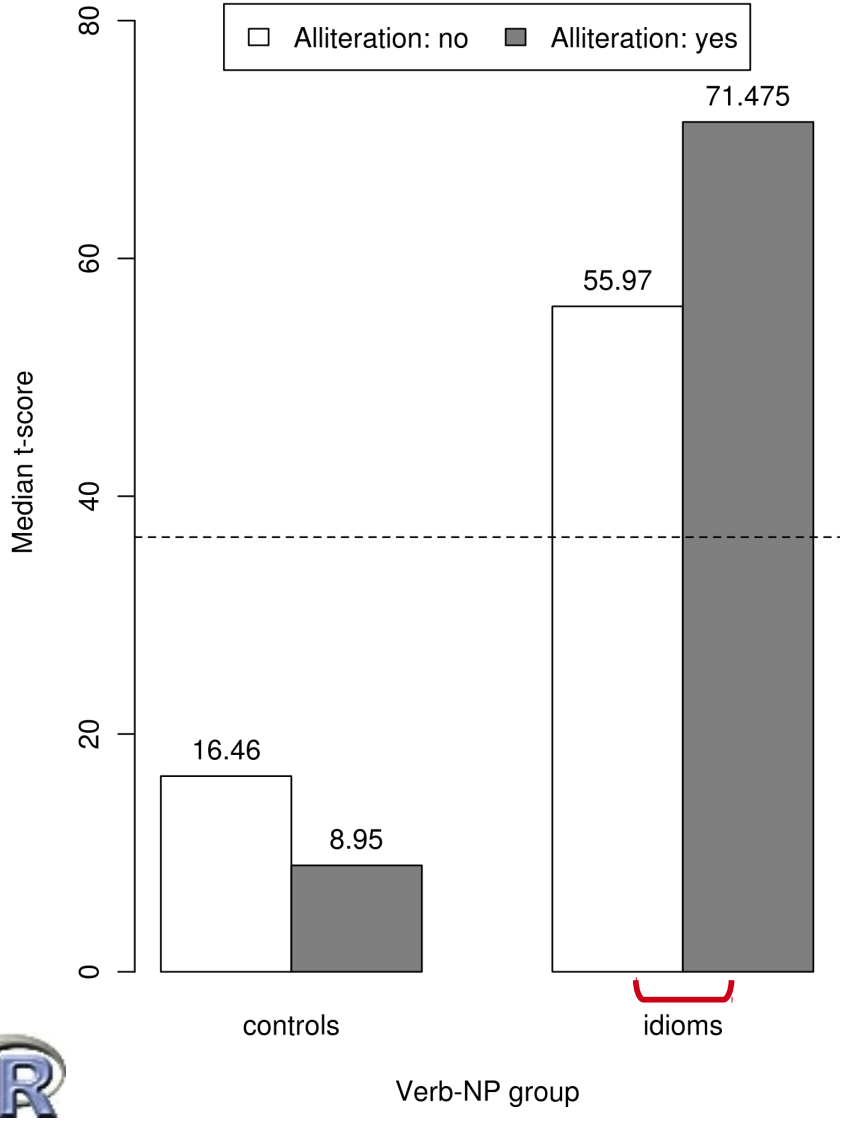
- I retrieved
  - the frequency of each verb
    - from the V-NP idioms
    - from the V-NP controls
  - the frequency of each head noun
    - from the V-NP idioms
    - from the V-NP controls
  - their co-occurrence frequency in all sentences from the BNC World
- I computed two measures of collocational attraction
  - *MI*
  - *t*
- statistical 'design':
  - Collocational strength ~ V-NP group (*idiom* vs. *control*) * Alliteration (*yes* vs. *no*)

Results concerning observed and expected proportions of alliterations



Results concerning collocations





Where to count alliterations

- The data consist of *way*-constructions from the BNC World
  - $SUBJ_{theme}$ V_{move} POSS way $[_{PP}$ P NP/S$]_{path/goal}$
- the constructions were retrieved by a manually-cleaned concordance of the sequence of a possessive pronoun immediately followed by *way*
- overall number of *way*-constructions: 5831
- some examples
  - *The British Task Force made its way across the Atlantic*
  - *The water found its way into the volcanic vent*

How to count alliterations

- For each of the constructions, I noted
  - the initial phoneme of the verb in the verb slot
    - *banged her way* → b
    - *wound your way* → w
  - the pronunciations of the verbs were taken from the CELEX database
- then I computed the percentage of alliterations
  - for types
  - for tokens
- but: we need (a) baseline(s)

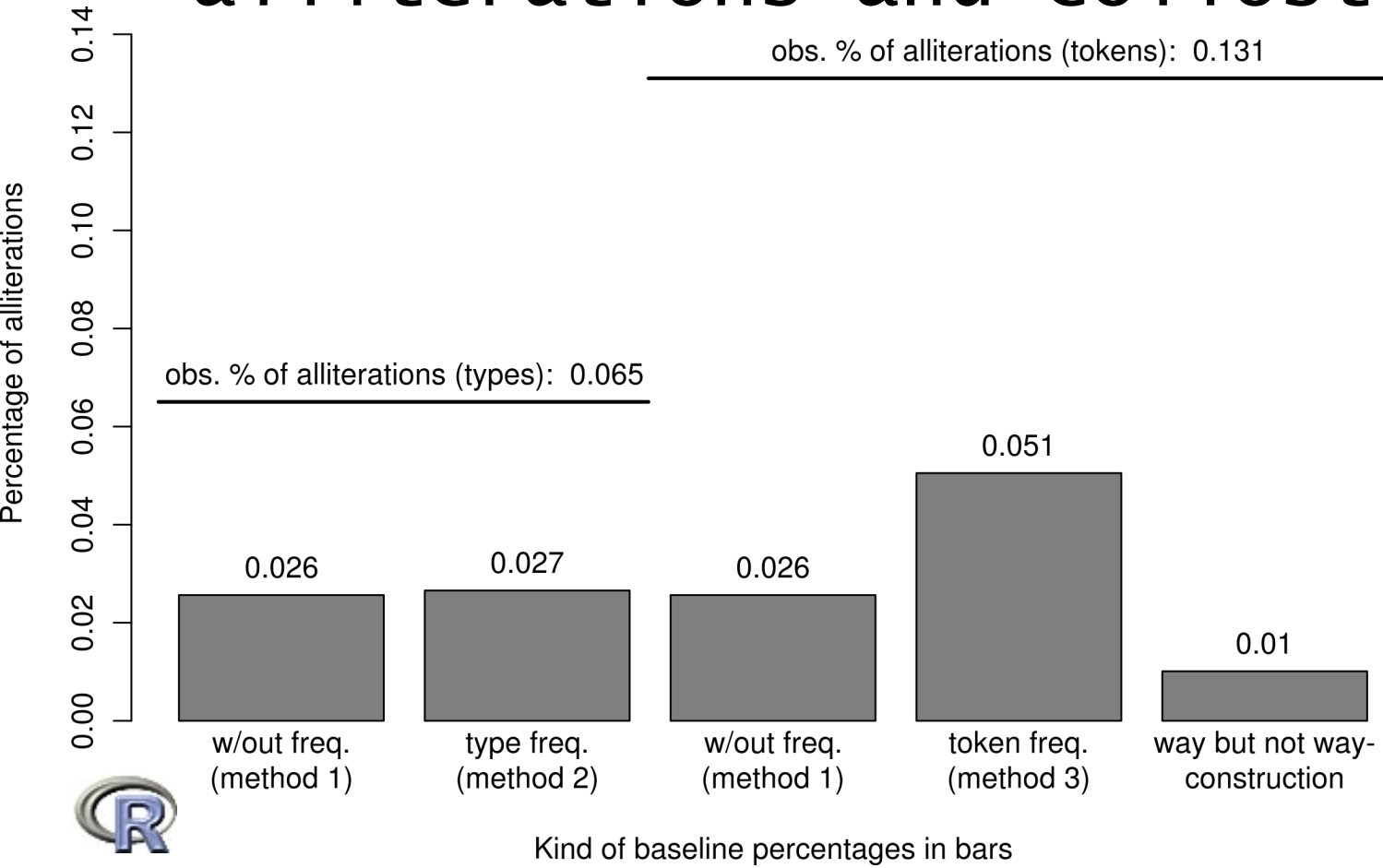
How many alliterations to expect: different kinds of baselines

- There are several ways of obtaining such expected baseline frequencies: on the basis of word-initial phonemes
  - without regard to type and token frequencies
    - this baseline is based on the number of phonemes that words in the CELEX database begin with
  - by taking into consideration their frequencies in differently frequent word types
    - this baseline is based on the probabilities that phonemes are the first phonemes in word types in the CELEX database
  - by taking into consideration their frequencies in word tokens
    - this baseline is based on the probabilities that phonemes are the first phonemes in word tokens in the CELEX database
- as a kind of control group, I retrieved all instances of the direct object *way* in transitive clauses in the ICE-GB and checked them for alliterations

Do the verb and *way* 'like' each other?

- I retrieved
  - the frequency of each verb lemma in the *way*-construction
  - the frequency of the *way*-construction (known from above)
- I computed a collexeme analysis, which quantifies the attraction/repulsion of each verb to the *way*-construction
  - $(-)\log_{10} p_{\text{Fisher-Yates exact test}}$ (cf. Stefanowitsch & Gries 2003)
  - ΔP (cf. Ellis & Ferreira-Junior 2009)
- statistical design
  - Collostruction strength ~ Alliteration (*yes* vs. *no*) (w/ each measure)

Results concerning proportions of alliterations and collostructions



	Measure of attraction -log p_{FYE}	ΔP
alliteration w/ verb: yes	1.69	0.00964
alliteration w/ verb: no	1.16	0.00347
p-value from $U_{one\text{-}tailed}$-test	0.1115	0.0669

Interim summary and questions for discussion

- The results are unambiguous
  - there are strong alliteration effects
  - these differ significantly from baselines regardless of how observed/expected frequencies are computed
  - these differ significantly from non-conventionalized but otherwise analogous structures
  - these are weakly but suggestively correlated with measures of collocational/collostructional attraction, which they appear to reinforce
- now …
  - does this phenomenon serve some purpose? if so, which?
  - how does it come about?
  - why is this effect observable in the form of alliterations?

Towards conventionalization of units involving alliteration

- One possible account
  - at some point, (a) speaker(s) used/created an expression
  - because of the alliteration, this was
    - 'fun' to use
    - easy to memorize
    - therefore used often enough to become (more) entrenched
  - this process is not unlike that undergone by, say, new and creative subtractive word-formations (e.g., *chunnel*, *foolosopher*, …)
- how could this process be accounted for?
  - with the growing recognition of the relevance of similarity/analogy for language learning and processing
  - chunking
  - phonological constituents (cf. Langacker 1997)

Similarity, chunking, and phonological constituency

- Similarity: we know of many other similarity effects
  - similarity of novel utterances to previous utterances is correlated with the novel utterances' acceptability (Bybee 2010:59)
  - a structure *S* primes towards *S* later more if the structures and how their slots are filled are more similar (Gries 2005, Szmrecsanyi 2005, Snider 2009)
  - similarity on various levels facilitates the emergence and perseverance of new subtractive word-formations (Gries 2004, 2006)
- why alliterations?
  - word beginnings facilitate recognition more than word endings (cf. Noteboom 1981, Bergen 2004 on phonaesthemes)
  - artificial-language learners can identify words with non-adjacent syllable dependencies better when those exhibit alliterations (Onnis et al. 2005)

Similarity, chunking, and phonological constituency

- Langacker (1997) distinguishes
  - semantic/conceptual constituents, based on links connecting elements fulfilling valence requirements of, or elaborating another element
    - "Another kind of conceptual group is the semantic pole of a complex lexical item […] It is well known that idioms are often phonologically discontinuous […], hence not symbolized by a classical phonological constituent" (Langacker 1997:15)
  - phonological constituents, based on temporal contiguity, rhythmic cohesiveness, "stress, pitch level, and even similarity in segmental content"
    - add to this the fact that "entries sharing phonetic and semantic features are highly interconnected depending upon the degree of similarity" (Bybee 2010:62f.)
- hypothesis: similarity (word beginnings = salient) → recognition of a phonological constituent, → higher degree of interconnectedness, → chunking → constructionalization

What's next …

- Possible next steps
  - increase the data base by looking at more types and tokens of the same conventionalized constructions
  - increase the data base by exploring other conventionalized constructions and/or proverbs
  - explore how much the obtained similarity effects are dependent on the constructions' slots not being too flexible (*into*-causative)
  - adopt a more comprehensive/flexible view of similarity (as in studies of blends and complex clippings; cf. Gries 2004, 2006)
  - adopt a more sophisticated quantitative methodology

What's next …

- With regard to similarity: what is the scope of this similarity effect 1? (cf. Bybee 2010 on *strung* verbs)
  - it could be identity of first phonemes (*run the risk*)
  - it could be similarity of first phonemes (*gimme a break*)
  - it could be identify of onsets (*fly the flag*)
  - it could be similarity of onsets (*gain some ground*)
  - it could be similarity of whole words (*get the boot*)
    - the pronunciation with syllabification and stress
    - the pronunciation without syllabification and stress
    - the segmental structure
    - the syllabic length and the phonemic length
    - articulatory similarity [ðə kæt ɪz aʊdə ðə bæg] [meɪk hedweɪ]
  - all of the above will be studied here
- what is the scope of this similarity effect 2? Is this more widespread and/or predictive? E.g., we know similarity in completely fixed proverbs/sayings is very high … And what does it mean/reflect?
- addition: the *into*-causative: [$_{VP}$V [$_{NP}$ Pat …] *into* V*ing*]

What's next …

- With regard to quantitative methodology: how can such data be studied best?
- obviously, one still needs some control group, and I will combine the non-*way*-constructions clauses and the non-idiomatic V-NP_{DirObj} examples
- for most above approaches to similarity, mere comparisons of obs-vs.-exp percentages won't work
- additional problem: the data usually violate all the assumptions that 'the usual tests' require
- strategy adopted here: robust statistics, which
  - can handle outliers, skewed distributions, heterogeneous variances etc. better than traditional statistics
  - do not lose as much information as the traditional alternatives of medians, *U*-tests, etc.

Measuring similarity 1: first phoneme and onset identity

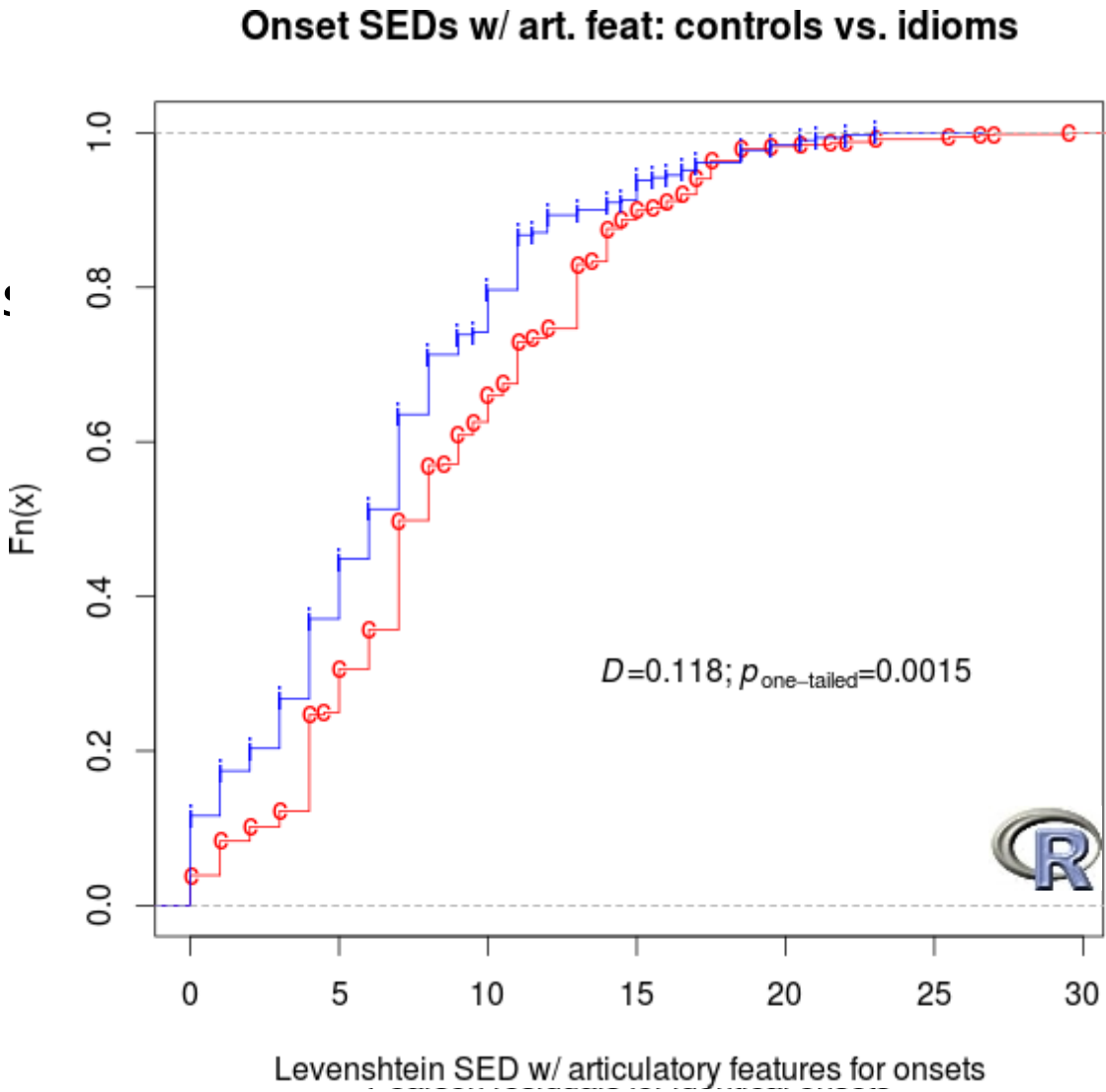
- The first phonemes of all four pattern types (controls and the three patterns) were taken from the CELEX database
- checking for identity of first phonemes
  - within each pattern, I cross-tabulated first phonemes of $word_1$ with first phonemes of $word_2$
  - I computed the Pearson residuals for the main diagonal
  - I compared the distribution of Pearson residuals
    - of the idioms to the controls
    - of *way*-construction to the controls
    - of *into*-causatives to the controls
  - I explored the distributions statistically
    - using Kolmogorov-Smirnov tests (with a correction for ties)
    - using a robust alternative to the *t*-test (Yuen 1974)
- then, the same was done for the onsets

Measuring similarity 1: first phoneme and onset similarity

- The first phonemes of all four pattern types (controls and the three patterns) were taken from the CELEX database
- checking for similarity of first phonemes
  - within each pattern, I computed for both first phonemes the version of the Levenshtein string edit distance that also considers articulatory features (Heeringa 2004)
  - I explored the distributions …
    - of the idioms to the controls
    - of *way*-constructions to the controls
    - of *into*-causatives to the controls
  - … statistically
    - using Kolmogorov-Smirnov tests (with a correction for ties)
    - using a robust alternative to the *t*-test (Yuen 1974)
- then, the same was done for the onsets

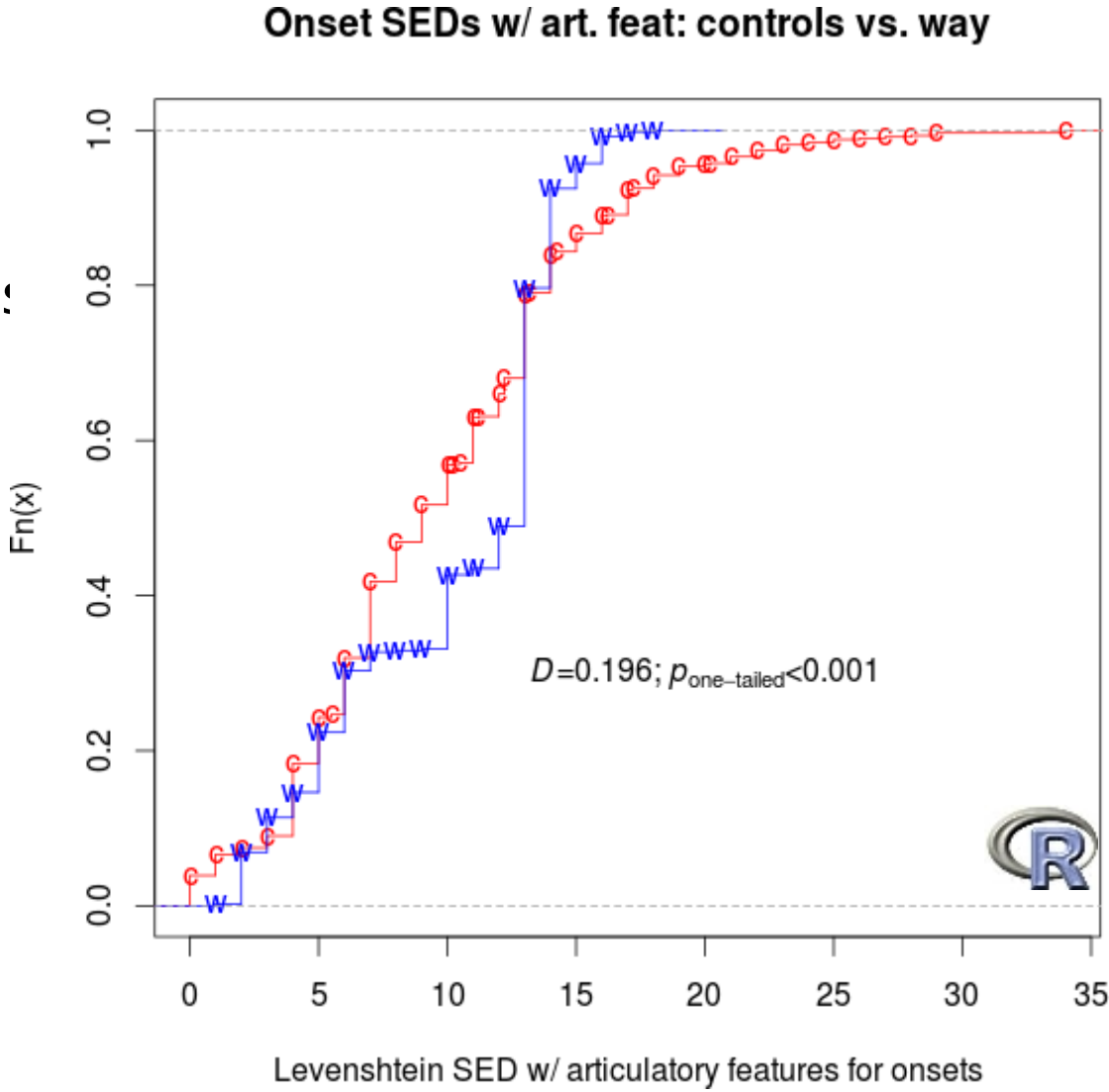
Measuring similarity 1a: controls vs. V-NP$_{DirObj}$ idioms

- Controls vs. idioms
  - identity of 1st phonemes
    - controls $<_{sim}$ idioms
      for ks and yuen
  - similarity of 1st phonemes
    - controls $<_{sim}$ idioms
      for ks and yuen
  - identity of onsets
    - controls $=_{sim}$ idioms
      for ks and yuen
  - similarity of onsets
    - controls $<_{sim}$ idioms
      only for ks



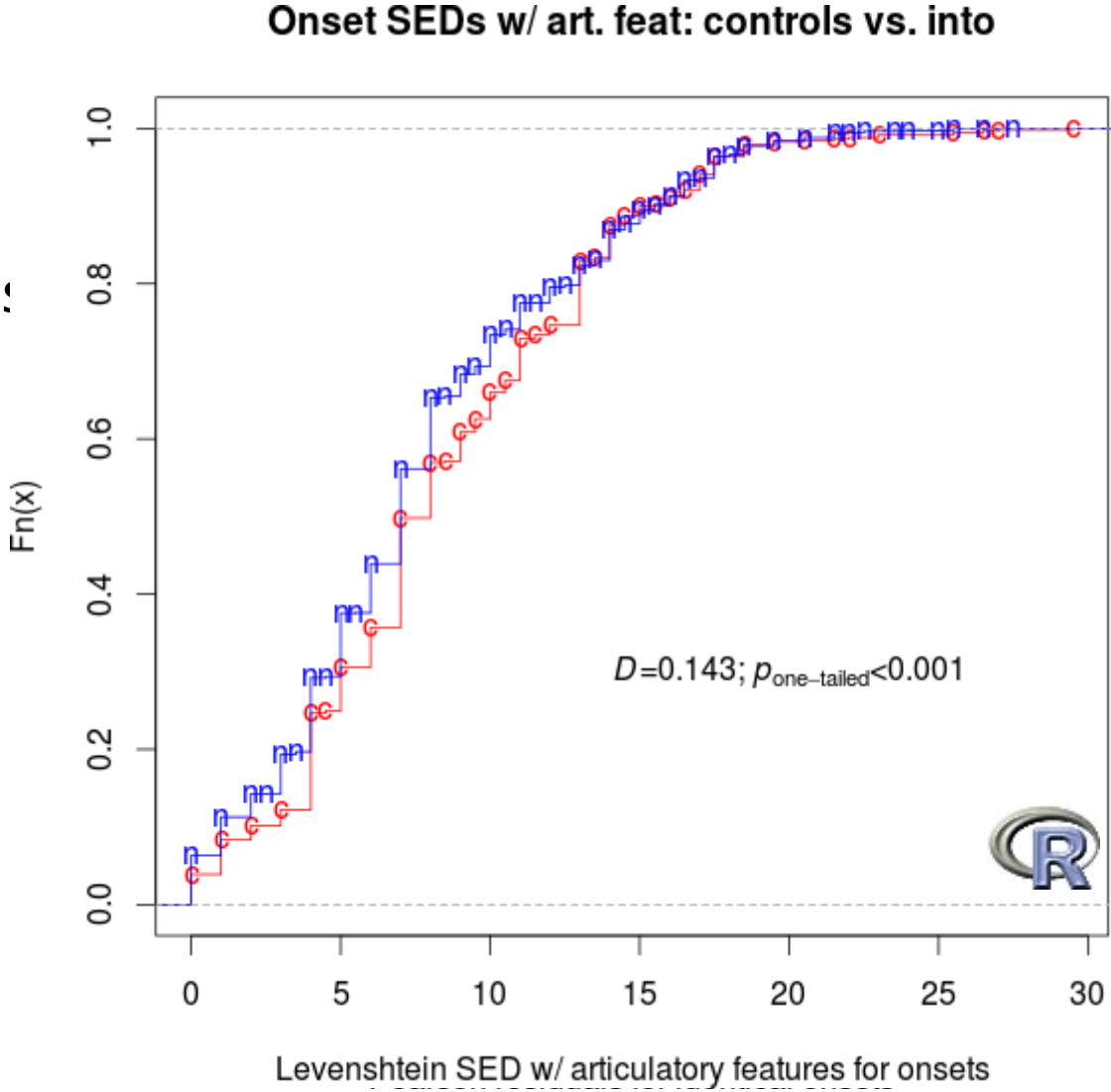
Measuring similarity 1b: controls vs. *way*-constructions

- Controls vs. *way*
  - identity of 1st phonemes
    - controls $<_{sim}$ *way*
      for 1-step M-est.ed CIs
  - similarity of 1st phonemes
    - controls $<_{sim}$ *way*
      for ks and yuen
  - identity of onsets
    - controls $<_{sim}$ *way*
      for 1-step M-est.ed CIs
  - similarity of onsets
    - controls $>_{sim}$ *way*
      for ks and yuen



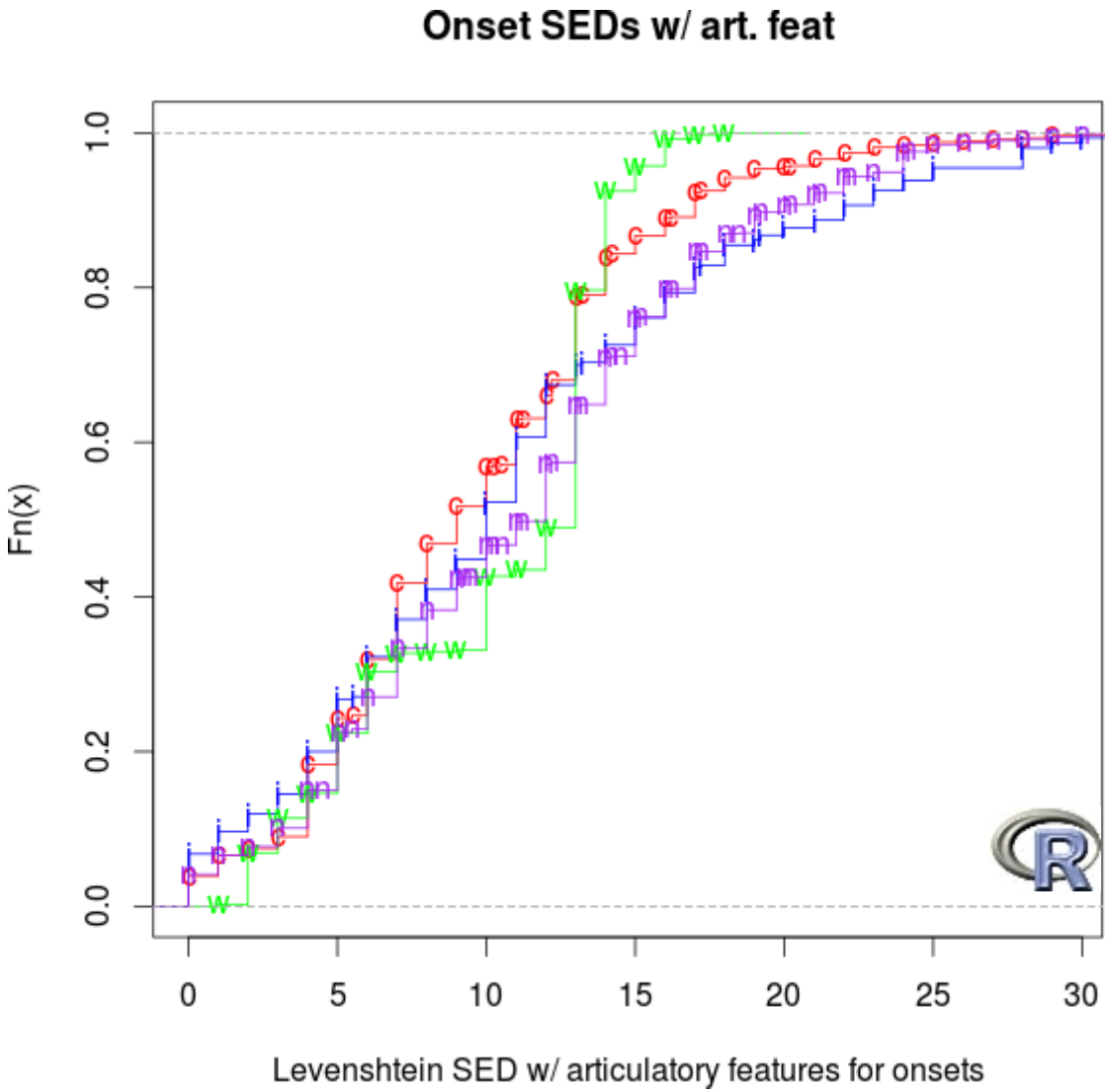
Measuring similarity 1c: controls vs. *into*-causatives

- Controls vs. *into*
  - identity of 1st phonemes
    - controls $=_{sim}$ *into*
      for ks and yuen
  - similarity of 1st phonemes
    - controls $<_{sim}$ *into*
      for ks and yuen
  - identity of onsets
    - controls $=_{sim}$ *into*
      for ks and yuen
  - similarity of onsets
    - controls $>_{sim}$ *into*
      for ks and yuen



Now all together …

- The ecdf plots are hard to interpret although some tendencies emerge
- the first phonemes show a clear pattern
  - idioms>*way*≈*into*>control
    for lm and t1wayv2
- the onsets are a mess
  - control=idioms=*into*=*way*
    for lm and t1wayv2
- overall interim conclusion: there are significant differences between (most of) the three patterns' first phonemes and onsets and those of the controls



Measuring similarity 2

- For the relevant words of all four patterns, I extracted from the CELEX database
  - the full transcription rI-'mEm-b@R
  - the phonemic transcription rImEmb@R
  - the segmental structure CVCVCCVC
  - the stress pattern u S u
  - the syllabic length 3
  - the phonemic length 8
- the first four above were compared for their similarity using normalized Levenshtein SED
- the last two above were compared for their similarity in terms of their difference (1-2)
- I explored the distributions statistically using
  - robust alternatives to one-way ANOVAs and
  - robust alternatives to confid. intervals (Wilcox 2012)

Measuring similarity 2

- Results for full transcriptions
  - idioms > *way* > control > *into*
- results for phonemic transcriptions
  - *into* = idiom = control > *way*
- results for segmental structure
  - idioms > *into* = control > *way*
- results for stress pattern
  - idioms = *way* > control > *into*
- results for syllabic length
  - idioms > *way* = control > *into*
- results for phonemic length
  - idioms > control > *way* = *into*
- the results are not unequivocal …
- but if they are compared to chance orderings, the following have a $p_{\text{one-tailed}}$-value of 0.05: full transcription, stress pattern, syllabic length

Summary and concluding remarks

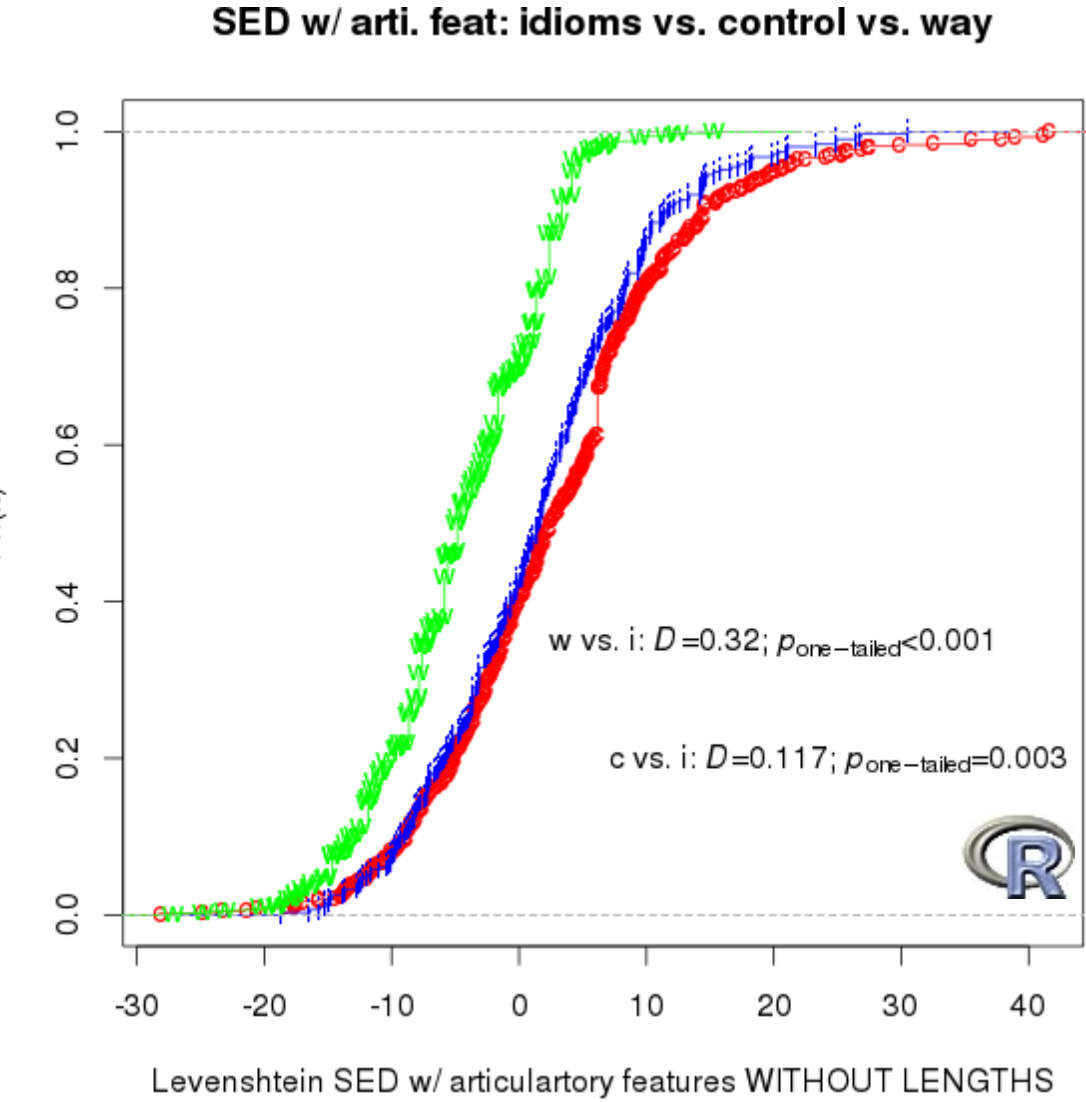
- The results are unambiguous: significant differences between the three patterns and the controls for
  - esp. identity/similarity of first phonemes
  - some degree of identity/similarity of onsets
  - full transcriptions, stress patterns, syllabic lengths
- thus, within-pole similarity of units across morphology and syntax is greater than
  - expected by chance
  - in non-conventionalized but otherwise similar structures
- unlike priming etc., the present effect is very local: within-VP, within multi-word words/units

Summary and concluding remarks

- More interestingly,
  - we know that similarity facilitates the formation/retention of completely-fixed units such as sayings or binomials: *going great guns*, *the cat is out of the bag*, … (e.g., Malkiel 1959, thx2 VGM)
  - now we see that this also holds for successively less morphological and more flexible/syntactic units and …
  - … it seems as if the degree of (alliterative) similarity is correlated with the flexibility of the unit
    - completely-fixed sayings/proverbs >= idiomatic V-NPs (which allow modification) > *way* (where one slot is flexible) > *into* > (where two slots are flexible) > controls
  - … seems to make sense from an exemplar-based perspective given the above and the fact that "entries sharing phonetic and semantic features are highly interconnected depending upon the degree of similarity" (Bybee 2010:62f.)
  - does this mean that similarity constrains (probably very weakly) the productivity of how slots are filled?

A word of caution: measuring similarity with phonemes with artic. features

- For the phonemic transcr
  words from of idioms, co
  way-construction, I compu
  version that also consid
- the differences are quit
- … but what do they reall
- lm(SED ~ length1*length2
- what if we partial the l
  - ks and yuen yield signifi
- but do we even want leng
- is it possible to confla
  into one (vector) measur
- which could then be expe
  in other applications (b
  definition of neighborho



Thank you!

http://tinyurl.com/stgries