AMLaP - September 4, 2014

Lexical clustering in efficient language design

Isabelle Dautriche
Ecole Normale Supérieure

Kyle Mahowald
MIT

Edward Gibson
MIT

Anne Christophe
Ecole Normale Supérieure

Steven T. Piantadosi
University of Rochester

What we know about the lexicon

The relationship between word forms and their meanings is **arbitrary** (e.g., Saussure, 1916)

cat	pusa	kottur
kakis	kedi	meo
gato	macka	chat
billi	Q'az	Katze

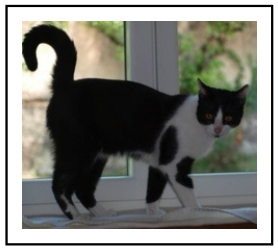
As a result, there are many free parameters left for how we choose our word forms

The set of word forms may be designed for:

- **communicative efficiency**
- **ease of processing and learnability**

What we know about the lexicon

The relationship between word forms and their meanings is **arbitrary** (e.g., Saussure, 1916)



cat	pusa	kottur
kakis	kedi	meo
gato	macka	chat
billi	Q'az	Katze

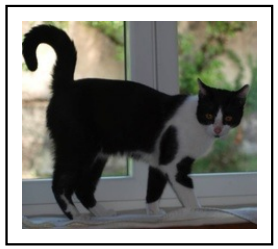
As a result, there are many free parameters left for how we choose our word forms

The set of word forms may be designed for:

- **communicative efficiency**
- **ease of processing and learnability**

What we know about the lexicon

The relationship between word forms and their meanings is **arbitrary** (e.g., Saussure, 1916)



cat	pusa	kottur
kakis	kedi	meo
gato	macka	chat
billi	Q'az	Katze

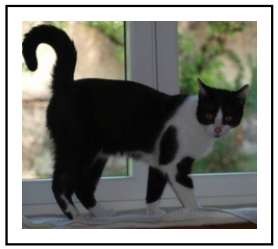
As a result, there are many free parameters left for how we choose our word forms

The set of word forms may be designed for:

- **communicative efficiency**
- **ease of processing and learnability**

What we know about the lexicon

The relationship between word forms and their meanings is **arbitrary** (e.g., Saussure, 1916)



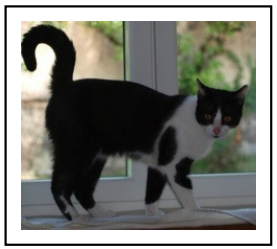
cat	pusa	kottur
kakis	kedi	meo
gato	macka	chat
billi	Q'az	Katze

As a result, there are many free parameters left for how we choose our word forms

The set of word forms may be designed for:

- **communicative efficiency**
- **ease of processing and learnability**

Communicative efficiency in the lexicon



FEP

FEB

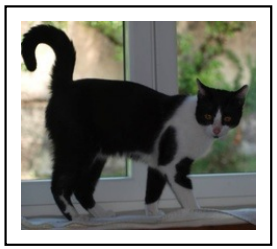
- phonetic inventories are more likely to be distinctive (Flemming, 2002)
- potentially confusable words are pronounced more slowly and more carefully (Gahl et al., 2012)

Prediction for the lexicon: Word forms should be maximally **dissimilar** from each other (within the bounds of phonotactics)

$\Rightarrow$ A pressure for **dispersion**

Communicative efficiency in the lexicon

Highly confusable!



FEP

FEB

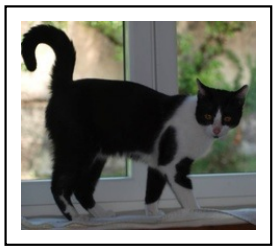
- phonetic inventories are more likely to be distinctive (Flemming, 2002)
- potentially confusable words are pronounced more slowly and more carefully (Gahl et al., 2012)

Prediction for the lexicon: Word forms should be maximally **dissimilar** from each other (within the bounds of phonotactics)

$\Rightarrow$ A pressure for **dispersion**

Communicative efficiency in the lexicon

Highly confusable!



FEP

FEB

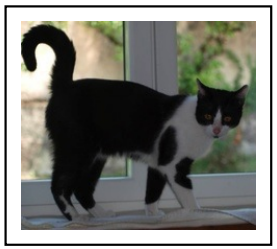
- phonetic inventories are more likely to be distinctive (Flemming, 2002)
- potentially confusable words are pronounced more slowly and more carefully (Gahl et al., 2012)

Prediction for the lexicon: Word forms should be maximally **dissimilar** from each other (within the bounds of phonotactics)

$\Rightarrow$ A pressure for **dispersion**

Communicative efficiency in the lexicon

Highly confusable!



FEP

FEB

- phonetic inventories are more likely to be distinctive (Flemming, 2002)
- potentially confusable words are pronounced more slowly and more carefully (Gahl et al., 2012)

Prediction for the lexicon: Word forms should be maximally **dissimilar** from each other (within the bounds of phonotactics)

$\Rightarrow$ A pressure for **dispersion**

Processing and learnability in the lexicon

Highly confusable!

FEP

FEB

- word form regularity helps grouping words into categories (Monaghan et al., 2011)
- adults find easier to learn and remember words living in high-density neighborhood (Storkel et al., 2006; Vitevitch et al., 2012)
- systematicity of form-meaning mappings in the lexicon (Monaghan et al., 2014; us)

Prediction for the lexicon: Word forms should be maximally **similar** to each other (more than expected by phonotactics)

$\Rightarrow$ A pressure for **clumpiness**

Processing and learnability in the lexicon

Highly confusable!

FEP

FEB

But also easier to learn and remember!

- word form regularity helps grouping words into categories (Monaghan et al., 2011)
- adults find easier to learn and remember words living in high-density neighborhood (Storkel et al., 2006; Vitevitch et al., 2012)
- systematicity of form-meaning mappings in the lexicon (Monaghan et al., 2014; us)

Prediction for the lexicon: Word forms should be maximally **similar** to each other (more than expected by phonotactics)

$\Rightarrow$ A pressure for **clumpiness**

Processing and learnability in the lexicon

Highly confusable!

FEP

FEB

But also easier to learn and remember!

- word form regularity helps grouping words into categories (Monaghan et al., 2011)
- adults find easier to learn and remember words living in high-density neighborhood (Storkel et al., 2006; Vitevitch et al., 2012)
- systematicity of form-meaning mappings in the lexicon (Monaghan et al., 2014; us)

Prediction for the lexicon: Word forms should be maximally **similar** to each other (more than expected by phonotactics)

$\Rightarrow$ A pressure for **clumpiness**

Processing and learnability in the lexicon

Highly confusable!

FEP

FEB

But also easier to learn and remember!

- word form regularity helps grouping words into categories (Monaghan et al., 2011)
- adults find easier to learn and remember words living in high-density neighborhood (Storkel et al., 2006; Vitevitch et al., 2012)
- systematicity of form-meaning mappings in the lexicon (Monaghan et al., 2014; us)

Prediction for the lexicon: Word forms should be maximally **similar** to each other (more than expected by phonotactics)

$\Rightarrow$ A pressure for **clumpiness**

OUTLINE

Two competing pressures:

- A pressure for **dispersion** (communicative efficiency..)
- A pressure for **clumpiness** (systematicity, learnability, memory..)

The question

Is there evidence for dispersion or clumpiness of word forms in the lexicon?

Two approaches:

- Large-scale analysis of lexicons
- Comparison of natural languages to plausible "null" lexicons

OUTLINE

Two competing pressures:

- A pressure for **dispersion** (communicative efficiency..)
- A pressure for **clumpiness** (systematicity, learnability, memory..)

The question

Is there evidence for dispersion or clumpiness of word forms in the lexicon?

Two approaches:

- Large-scale analysis of lexicons
- Comparison of natural languages to plausible "null" lexicons

Large-scale analysis

If the lexicon is optimized in any direction, **frequent word forms** should be more optimized simply because they are more used.

Can a pressure for dispersion or clumpiness be found in the frequency patterns of word use? (see Frauenfelder et al. 1993).

Predictions:

- **Dispersion**: the most frequent words should be phonologically more unique
- **Clumpiness**: the most frequent words should share more sound sequences

Method:

- orthographic lexicons from 115 languages extracted from Wikipedia (top 20k most frequent word-tokens)
- for each word-token, calculate:
  - number of neighbors (cat/car and cat/cats)
  - frequency
  - orthographic probability as a proxy for phonotactic probability (using a trigram model)

Large-scale analysis

If the lexicon is optimized in any direction, **frequent word forms** should be more optimized simply because they are more used.

Can a pressure for dispersion or clumpiness be found in the frequency patterns of word use? (see Frauenfelder et al. 1993).

Predictions:

- **Dispersion**: the most frequent words should be phonologically more unique
- **Clumpiness**: the most frequent words should share more sound sequences

Method:

- orthographic lexicons from 115 languages extracted from Wikipedia (top 20k most frequent word-tokens)
- for each word-token, calculate:
  - number of neighbors (cat/can and cat/cats)
  - frequency
  - orthographic probability as a proxy for phonotactic probability (using a trigram model)

Large-scale analysis

If the lexicon is optimized in any direction, **frequent word forms** should be more optimized simply because they are more used.

Can a pressure for dispersion or clumpiness be found in the frequency patterns of word use? (see Frauenfelder et al. 1993).

Predictions:

- **Dispersion**: the most frequent words should be phonologically more unique
- **Clumpiness**: the most frequent words should share more sound sequences

Method:

- orthographic lexicons from 115 languages extracted from Wikipedia (top 20k most frequent word-tokens)
- for each word-token, calculate:
  - number of neighbors (cat/car and cat/cats)
  - frequency
  - orthographic probability as a proxy for phonotactic probability (using a 3-gram model)

Large-scale analysis

If the lexicon is optimized in any direction, **frequent word forms** should be more optimized simply because they are more used.

Can a pressure for dispersion or clumpiness be found in the frequency patterns of word use? (see Frauenfelder et al. 1993).

Predictions:

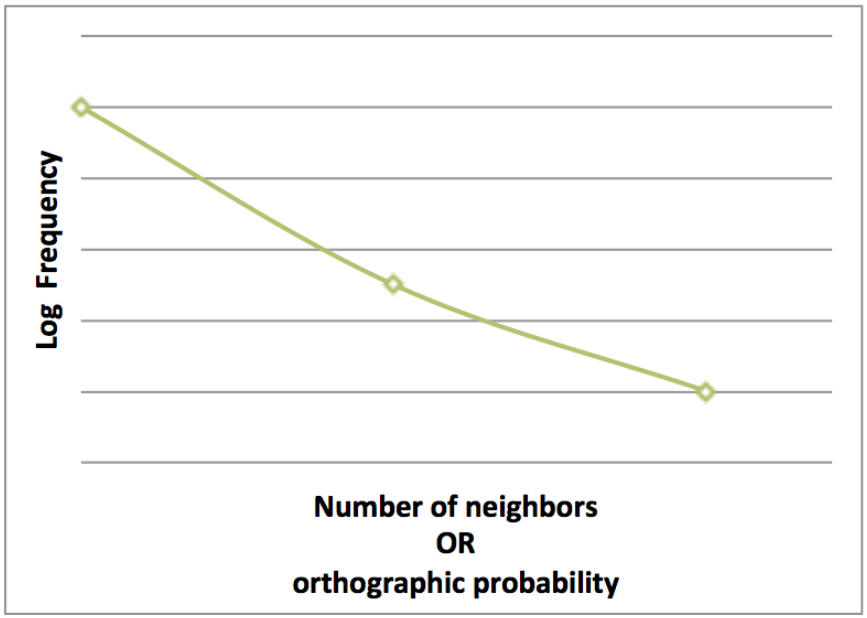
- **Dispersion**: the most frequent words should be phonologically more unique
- **Clumpiness**: the most frequent words should share more sound sequences

Method:

- orthographic lexicons from 115 languages extracted from Wikipedia (top 20k most frequent word-tokens)
- for each word-token, calculate:
  - number of neighbors (cat/car and cat/cats)
  - frequency
  - orthographic probability as a proxy for phonotactic probability (using a 3-gram model)

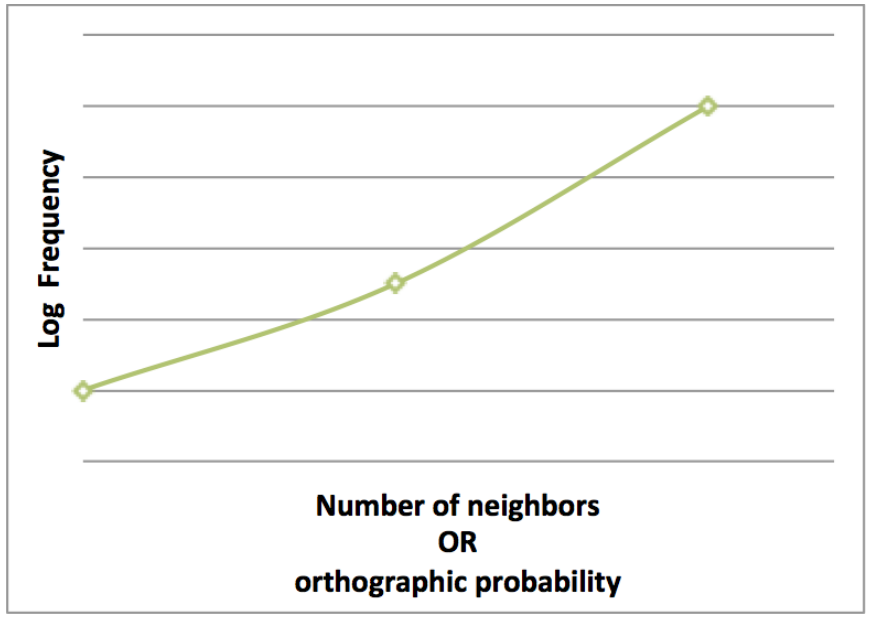
Possible results

Dispersion
Frequent words have fewer neighbors
(and lower orthographic probability)

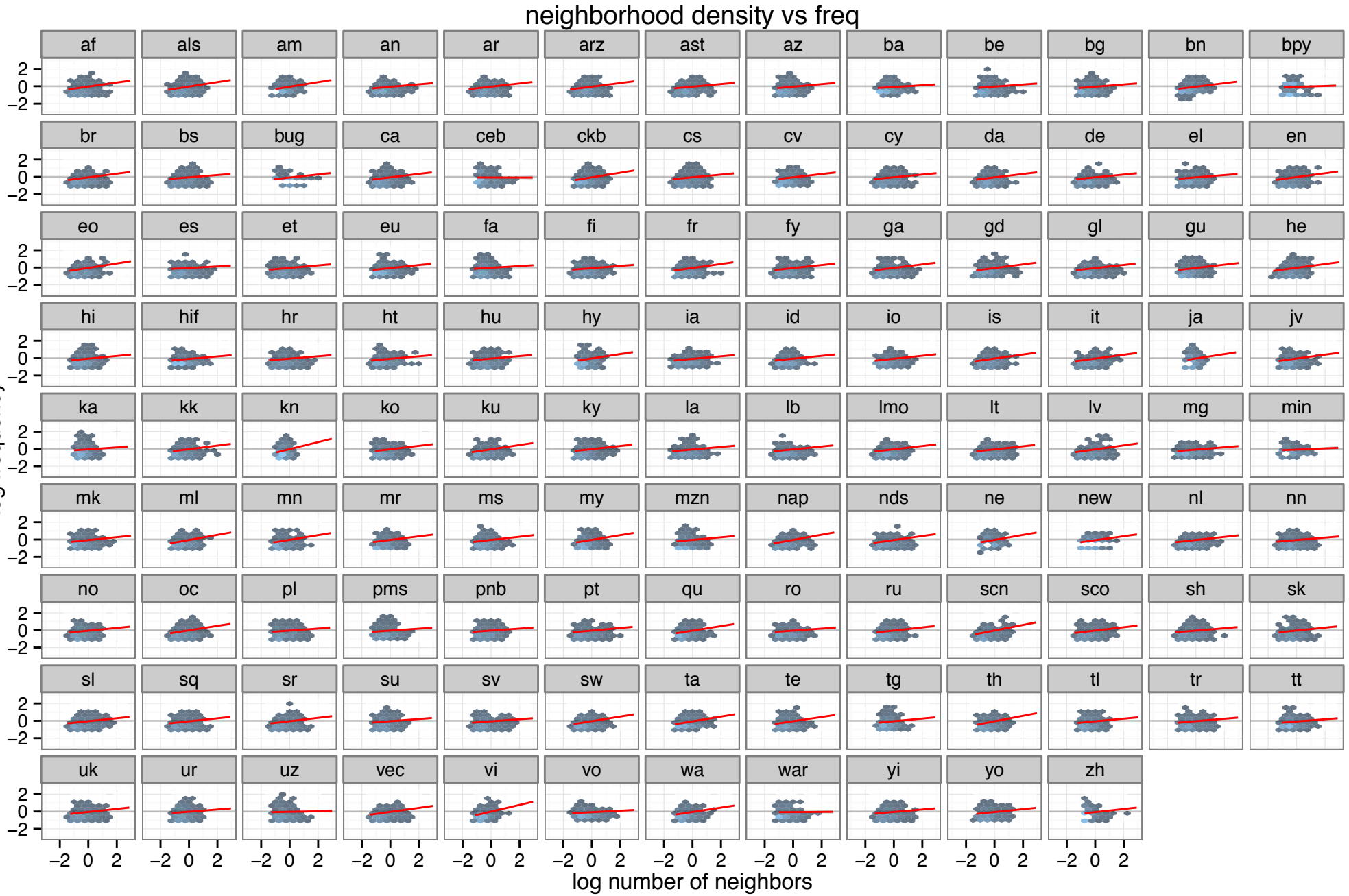


Negative correlation

Clumpiness
Frequent words have more neighbors
(and higher orthographic probability)



Positive correlation

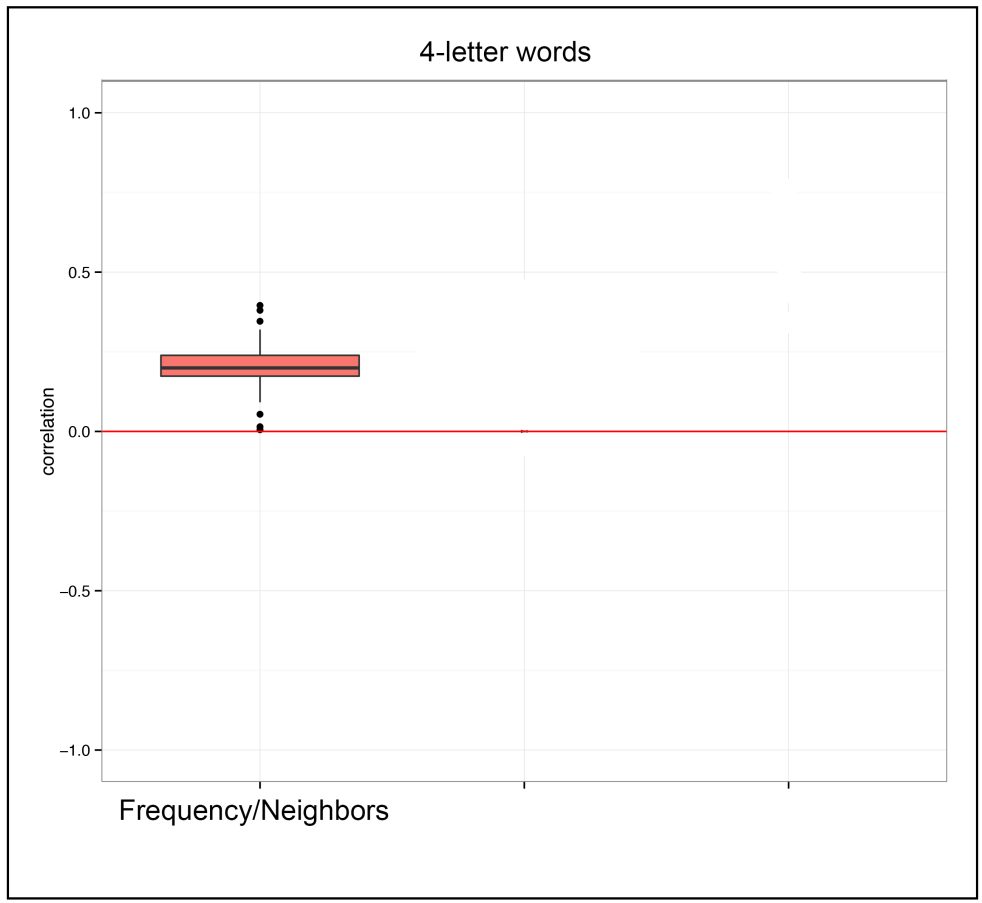

neighborhood density vs freq
af als am an ar arz ast az ba be bg bn bpy
br bs bug ca ceb ckb cs cv cy da de el en
eo es et eu fa fi fr fy ga gd gl gu he
hi hif hr ht hu hy ia id io is it ja jv
ka kk kn ko ku ky la lb lmo lt lv mg min
mk ml mn mr ms my mzn nap nds ne new nl nn
no oc pl pms pnb pt qu ro ru scn sco sh sk
sl sq sr su sv sw ta te tg th tl tr tt
uk ur uz vec vi vo wa war yi yo zh
2
0
−2
−2 0 2
log number of neighbors

For each language, we computed Spearman rank correlations between:

- log frequency and number of neighbors
- log frequency and orthographic probability
- orthographic probability and number of neighbors

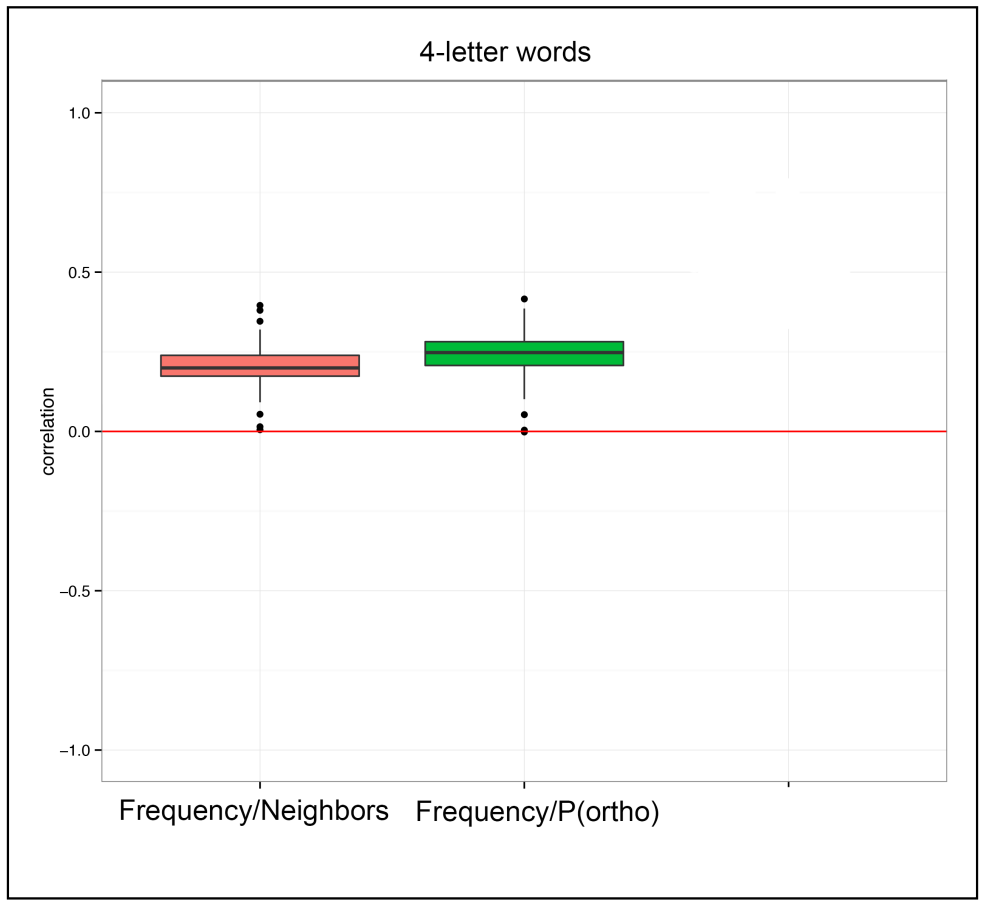
For each language, we computed Spearman rank correlations between:

- log frequency and number of neighbors
- log frequency and orthographic probability
- orthographic probability and number of neighbors



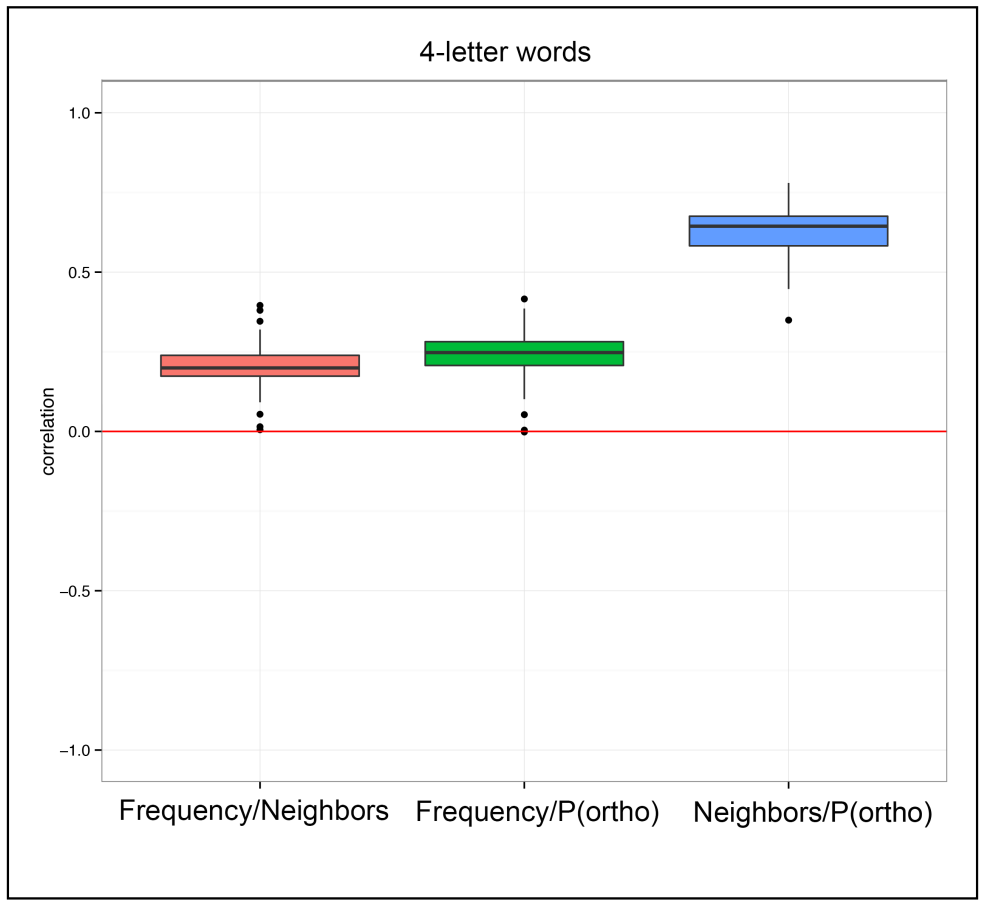
For each language, we computed Spearman rank correlations between:

- log frequency and number of neighbors
- log frequency and orthographic probability
- orthographic probability and number of neighbors



For each language, we computed Spearman rank correlations between:

- log frequency and number of neighbors
- log frequency and orthographic probability
- orthographic probability and number of neighbors



Results summary

Across all 115 languages we found:

- an (obvious) correlation b/w orthographic probability and number of neighbors
  ⇒ words that are more common are more likely to have neighbors
- correlations b/w number of neighbors and frequency and b/w orthographic probability and frequency
  ⇒ if a word has many neighbors and/or has high orthographic probability, it is more likely to be frequent.

This suggests a pressure for clumpiness that becomes stronger with use.

Yet, this effect is predicted by a model of word generation: more probable strings will be generated more often and higher string probability gives rise to more neighbors...

⇒ **we need a proper baseline**

Results summary

Across all 115 languages we found:

- an (obvious) correlation b/w orthographic probability and number of neighbors
  ⇒ words that are more common are more likely to have neighbors
- correlations b/w number of neighbors and frequency and b/w orthographic probability and frequency
  ⇒ if a word has many neighbors and/or has high orthographic probability, it is more likely to be frequent.

This suggests a pressure for clumpiness that becomes stronger with use.

Yet, this effect is predicted by a model of word generation: more probable strings will be generated more often and higher string probability gives rise to more neighbors...

⇒ **we need a proper baseline**

Results summary

Across all 115 languages we found:

- an (obvious) correlation b/w orthographic probability and number of neighbors
  ⇒ words that are more common are more likely to have neighbors
- correlations b/w number of neighbors and frequency and b/w orthographic probability and frequency
  ⇒ if a word has many neighbors and/or has high orthographic probability, it is more likely to be frequent.

This suggests a pressure for clumpiness that becomes stronger with use.

Yet, this effect is predicted by a model of word generation: more probable strings will be generated more often and higher string probability gives rise to more neighbors...

⇒ we need a proper baseline

Results summary

Across all 115 languages we found:

- an (obvious) correlation b/w orthographic probability and number of neighbors
  ⇒ words that are more common are more likely to have neighbors
- correlations b/w number of neighbors and frequency and b/w orthographic probability and frequency
  ⇒ if a word has many neighbors and/or has high orthographic probability, it is more likely to be frequent.

This suggests a pressure for clumpiness that becomes stronger with use.

Yet, this effect is predicted by a model of word generation: more probable strings will be generated more often and higher string probability gives rise to more neighbors...

⇒ **we need a proper baseline**

OUTLINE

Two competing pressures:

- A pressure for **dispersion** (communicative efficiency..)
- A pressure for **clumpiness** (systematicity, learnability, memory..)

The question

Is there evidence for dispersion or clumpiness of word forms in the lexicon?

Two approaches:

- Large-scale analyses of lexicons
- Comparison of natural languages to phonotactically-controlled "null" lexicons

OUTLINE

Two competing pressures:

- A pressure for **dispersion** (communicative efficiency..)
- A pressure for **clumpiness** (systematicity, learnability, memory..)

The question

Is there evidence for dispersion or clumpiness of word forms in the lexicon?

Two approaches:

- Large-scale analyses of lexicons
- Comparison of natural languages to phonotactically-controlled "null" lexicons

Simulating "null" lexicons

Is the lexicon clumpier than what would be expected by chance alone?

Good question but....how to determine the chance level?

New methodology

- Generate "null" lexicons for which there is no pressure for dispersion or clumpiness, beyond probabilistic phonotactics
- Compare real lexicons to these null lexicons on a variety of lexical measures to see how it is different.

Simulating "null" lexicons

Is the lexicon clumpier than what would be expected by chance alone?

Good question but....how to determine the chance level?

New methodology

- Generate "null" lexicons for which there is no pressure for dispersion or clumpiness, beyond probabilistic phonotactics
- Compare real lexicons to these null lexicons on a variety of lexical measures to see how it is different.

Simulating "null" lexicons

Is the lexicon clumpier than what would be expected by chance alone?

Good question but....how to determine the chance level?

New methodology

- Generate "null" lexicons for which there is no pressure for dispersion or clumpiness, beyond probabilistic phonotactics
- Compare real lexicons to these null lexicons on a variety of lexical measures to see how it is different.

Method

Real lexicons (phonemic transcription)
monomorphemic lemmas from Dutch, English, German (CELEX), French (Lexique)
(4000 to 6000 words, no homophone, frequency > 0)

Models of phonotactically controlled lexicons

- n-phone models ($n = 1$ to 6)
- n-syllable models ($n = 1$ to 2)
- PCFG over syllables

(with back-off and smoothing)
All matched for length to the real lexicons
Sample 30 "null" lexicons for each model

Selecting the best model
Each model trained on 75% of the lexicon and evaluated on the probability of generating the remaining 25% of the lexicon.

Best model: 5-phone model (50% of generated words are real)

Measures of word form similarity: number of minimal pair, average Levenshtein distance (edit), network measures...

Method

Real lexicons (phonemic transcription)
monomorphemic lemmas from Dutch, English, German (CELEX), French (Lexique)
(4000 to 6000 words, no homophone, frequency > 0)

Models of phonotactically controlled lexicons

- n-phone models ($n = 1$ to 6)
- n-syllable models ($n = 1$ to 2)
- PCFG over syllables

(with back-off and smoothing)
All matched for length to the real lexicons
Sample 30 "null" lexicons for each model

Selecting the best model
Each model trained on 75% of the lexicon and evaluated on the probability of generating the remaining 25% of the lexicon.

Best model: 5-phone model (50% of generated words are real)

Measures of word form similarity: number of minimal pair, average Levenshtein distance (edit), network measures...

Method

Real lexicons (phonemic transcription)
monomorphemic lemmas from Dutch, English, German (CELEX), French (Lexique)
(4000 to 6000 words, no homophone, frequency > 0)

Models of phonotactically controlled lexicons

- n-phone models ($n = 1$ to 6)
- n-syllable models ($n = 1$ to 2)
- PCFG over syllables

(with back-off and smoothing)
All matched for length to the real lexicons
Sample 30 "null" lexicons for each model

Selecting the best model
Each model trained on 75% of the lexicon and evaluated on the probability of generating the remaining 25% of the lexicon.

Best model: 5-phone model (50% of generated words are real)

Measures of word form similarity: number of minimal pair, average Levenshtein distance (edit), network measures...

Method

Real lexicons (phonemic transcription)
monomorphemic lemmas from Dutch, English, German (CELEX), French (Lexique)
(4000 to 6000 words, no homophone, frequency > 0)

Models of phonotactically controlled lexicons

- n-phone models ($n = 1$ to 6)
- n-syllable models ($n = 1$ to 2)
- PCFG over syllables

(with back-off and smoothing)
All matched for length to the real lexicons
Sample 30 "null" lexicons for each model

Selecting the best model
Each model trained on 75% of the lexicon and evaluated on the probability of generating the remaining 25% of the lexicon.

Best model: 5-phone model (50% of generated words are real)

Measures of word form similarity: number of minimal pair, average Levenshtein distance (edit), network measures...

Method

Real lexicons (phonemic transcription)
monomorphemic lemmas from Dutch, English, German (CELEX), French (Lexique)
(4000 to 6000 words, no homophone, frequency > 0)

Models of phonotactically controlled lexicons

- n-phone models ($n = 1$ to 6)
- n-syllable models ($n = 1$ to 2)
- PCFG over syllables

(with back-off and smoothing)
All matched for length to the real lexicons
Sample 30 "null" lexicons for each model

Selecting the best model
Each model trained on 75% of the lexicon and evaluated on the probability of generating the remaining 25% of the lexicon.

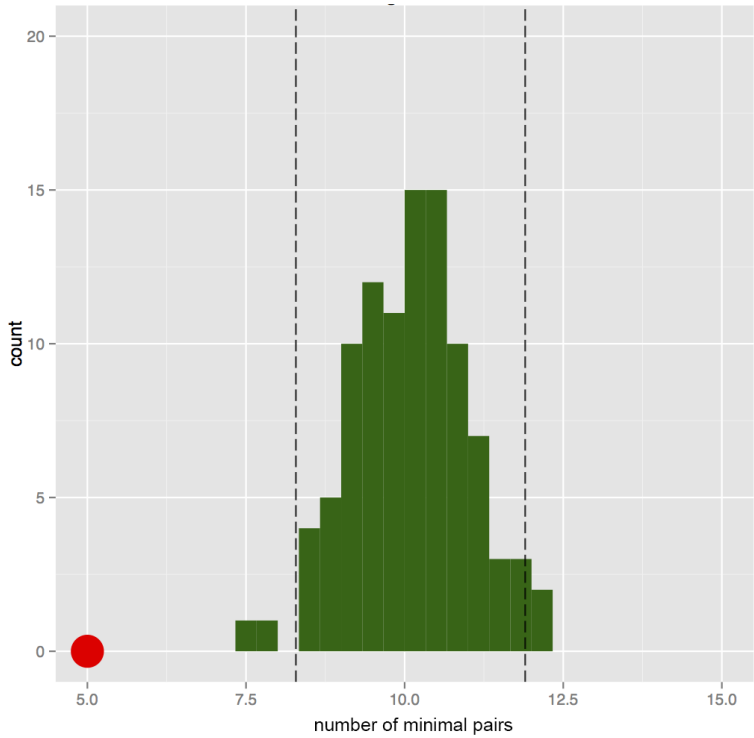
Best model: 5-phone model (50% of generated words are real)

Measures of word form similarity: number of minimal pair, average Levenshtein distance (edit), network measures...

Possible results: number of minimal pairs

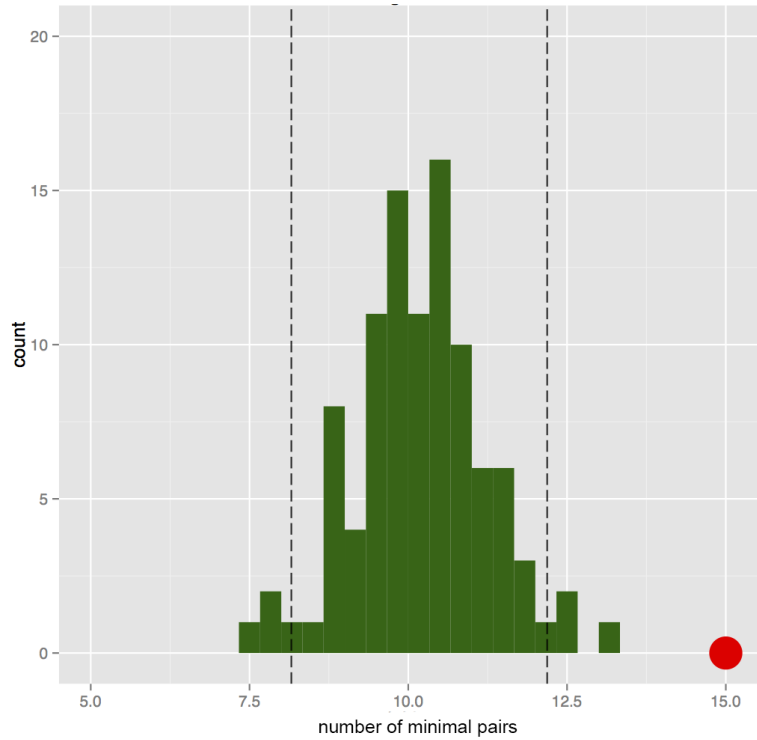
Dispersion

Minimal pairs (cat/bat) are avoided.



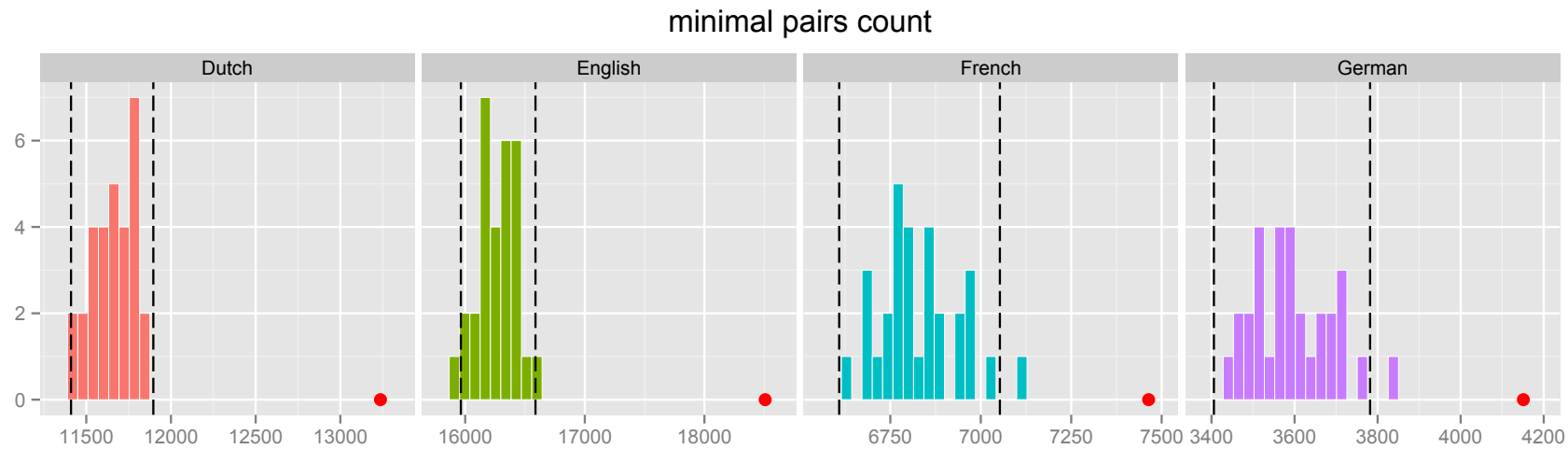
Clumpiness

Minimal pairs (cat/bat) are more likely.



red dot: the number of minimal pairs in the real lexicon.
green histogram: the distribution of the number of minimal pairs across simulated lexicons.
dotted lines: 95% interval

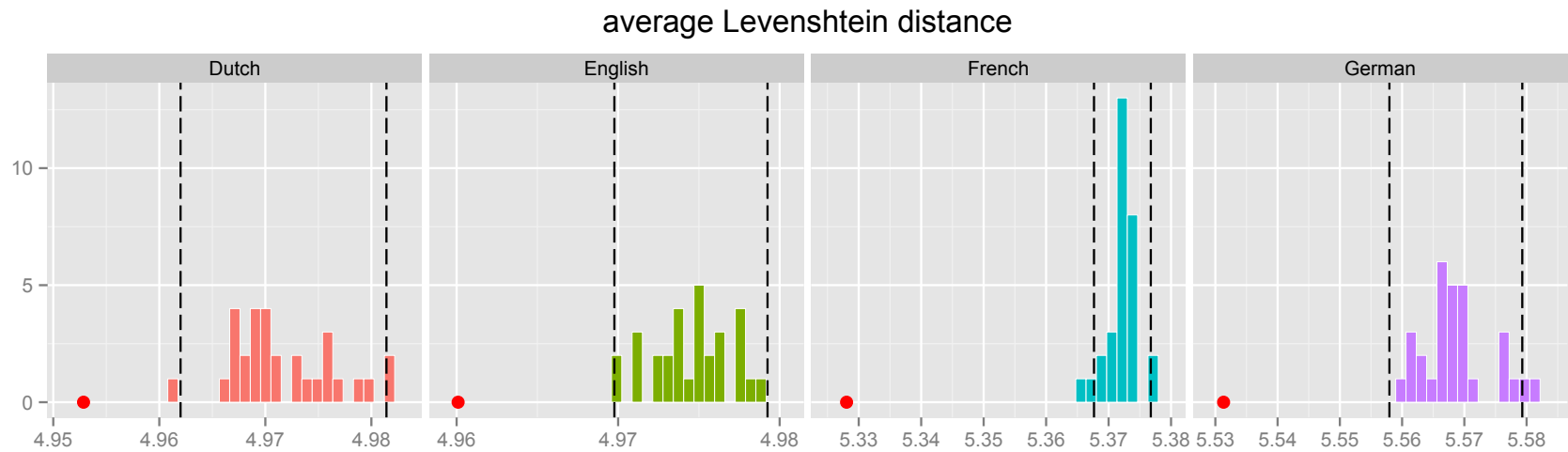
Actual result: number of minimal pairs



Result

More minimal pairs in all 4 real lexicons than expected by chance. (especially true for small lengths)

Another measure: Average Levenshtein distance

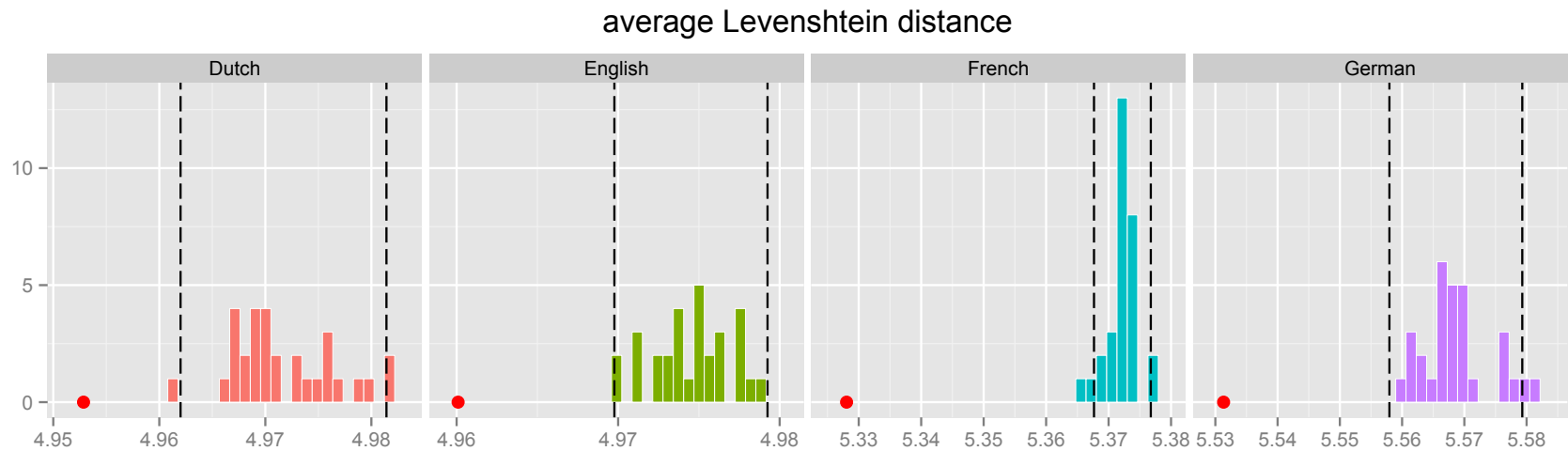


Result

Word forms are more similar in the real lexicon than expected by chance (lower Levenshtein distance)

Additional measures (network transitivity, clustering coefficients) tend towards the same conclusion: the lexicon seems clumpy!

Another measure: Average Levenshtein distance



Result

Word forms are more similar in the real lexicon than expected by chance (lower Levenshtein distance)

Additional measures (network transitivity, clustering coefficients) tend towards the same conclusion: the lexicon seems clumpy!

Null lexicon summary

- Most of the measures we reviewed suggest that **the lexicon is clumpier than what would be expected by chance.**
- This is not the result of morphological regularity since we focused on **monomorphemes**
- Nor the result of sound-to-sound probability (controlled by our generative model)

Global summary

Why does communicative efficiency not conflict with clumpiness in the lexicon?

- Clumpiness in the lexicon may not appear randomly but is organized along dimensions that maximize word form recoverability. (e.g., word onsets)
- Context is usually enough to disambiguate words regardless of how close they are in phonological space (Piantadosi et al., 2012)

- **Large-scale analysis results:** Most frequent words share more sound sequences pointing towards **clumpiness**
- **Null lexicon results:** The lexicon is **clumpier** than expected by chance across most measures of word form similarity

Global summary

Why does communicative efficiency not conflict with clumpiness in the lexicon?

- Clumpiness in the lexicon may not appear randomly but is organized along dimensions that maximize word form recoverability. (e.g., word onsets)
- Context is usually enough to disambiguate words regardless of how close they are in phonological space (Piantadosi et al., 2012)

- **Large-scale analysis results:** Most frequent words share more sound sequences pointing towards **clumpiness**
- **Null lexicon results:** The lexicon is **clumpier** than expected by chance across most measures of word form similarity

Global summary

Why does communicative efficiency not conflict with clumpiness in the lexicon?

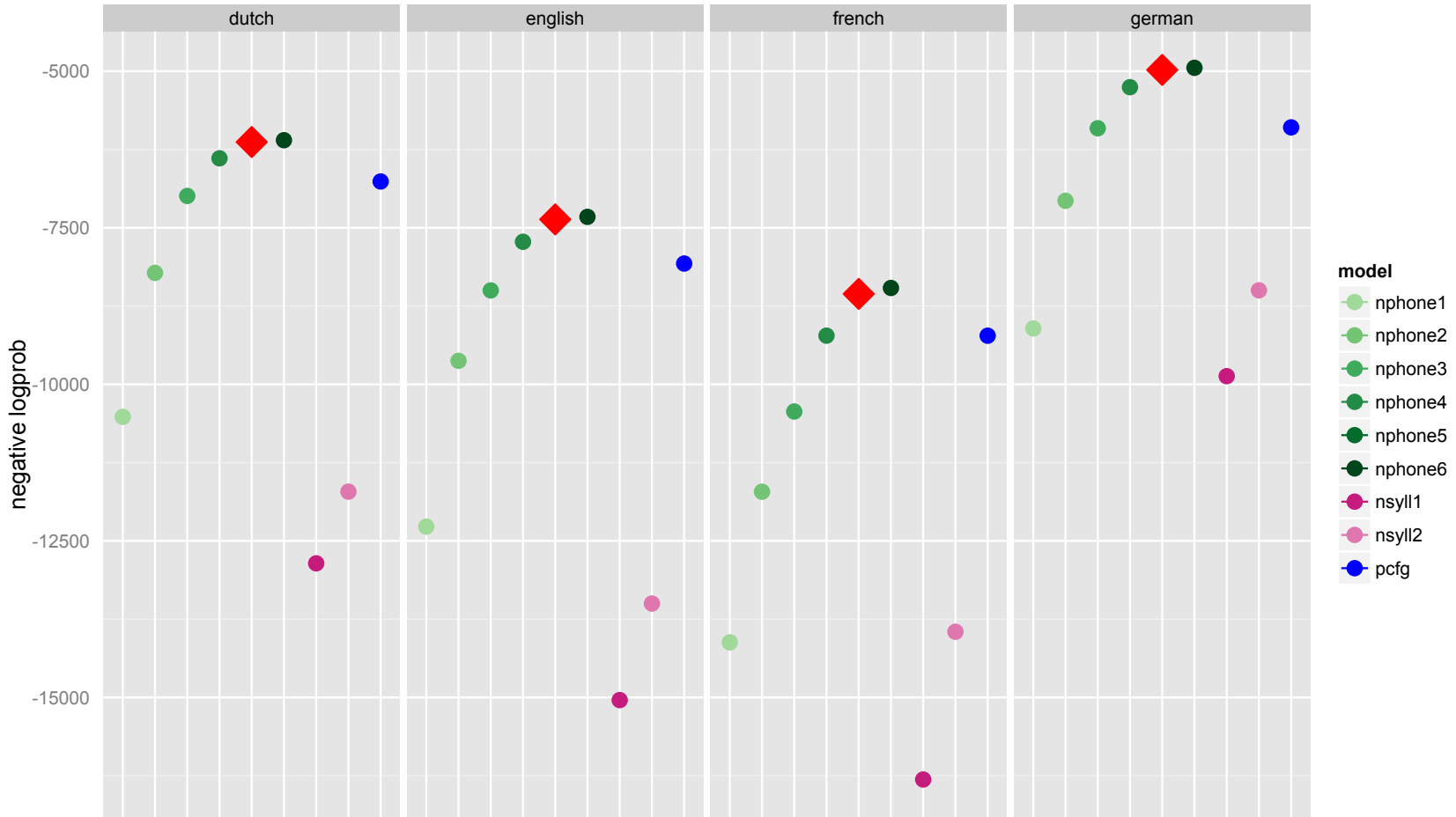
- Clumpiness in the lexicon may not appear randomly but is organized along dimensions that maximize word form recoverability. (e.g., word onsets)
- Context is usually enough to disambiguate words regardless of how close they are in phonological space (Piantadosi et al., 2012)

- **Large-scale analysis results:** Most frequent words share more sound sequences pointing towards **clumpiness**
- **Null lexicon results:** The lexicon is **clumpier** than expected by chance across most measures of word form similarity

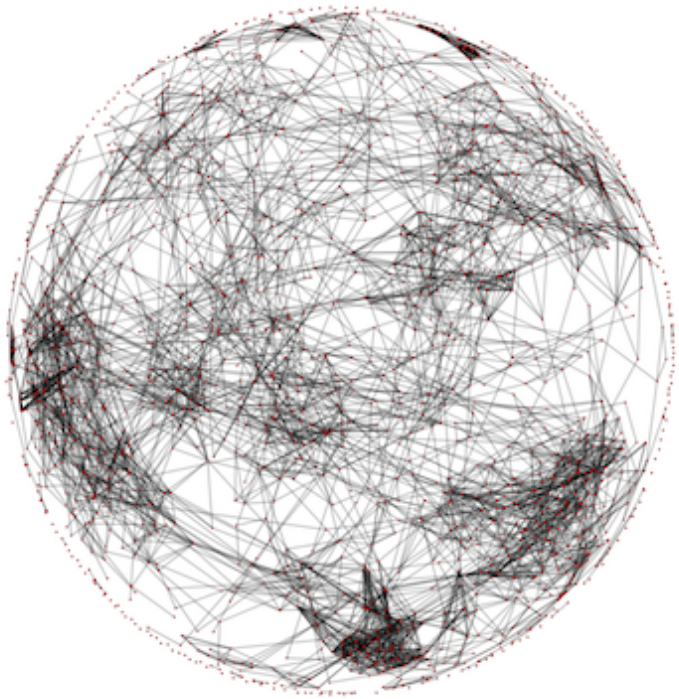
Thank you!



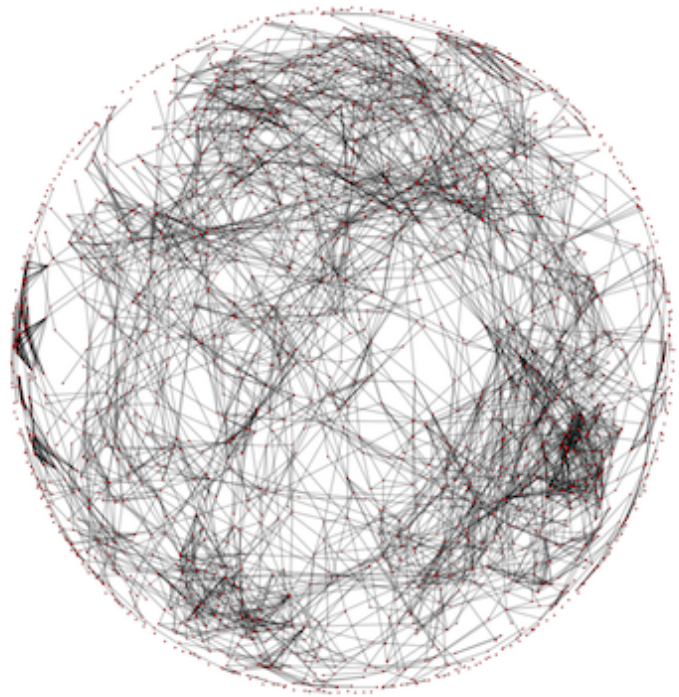

Evaluation


dutch
english
french
german
-5000
-7500
-10000
-12500
-15000
negative logprob
model
nphone1
nphone2
nphone3
nphone4
nphone5
nphone6
nsyll1
nsyll2
pcfg

Best model: 5-phone

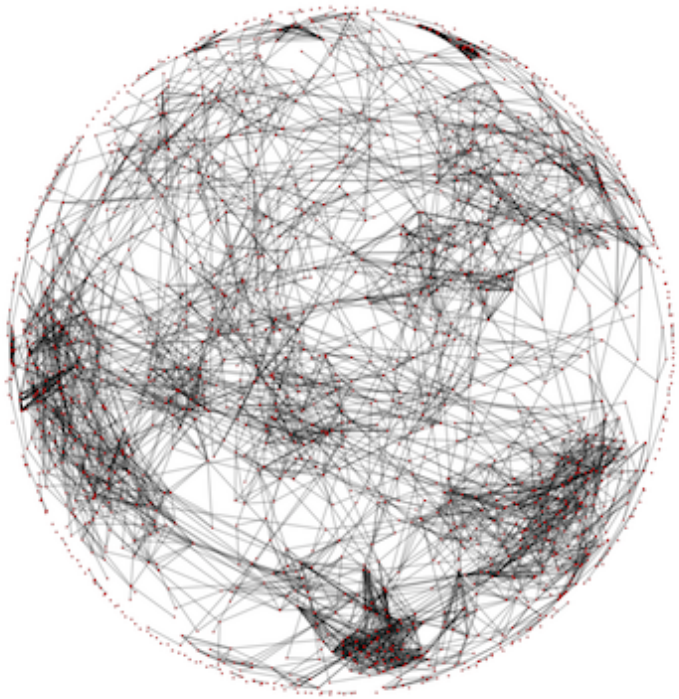


Real Lexicon

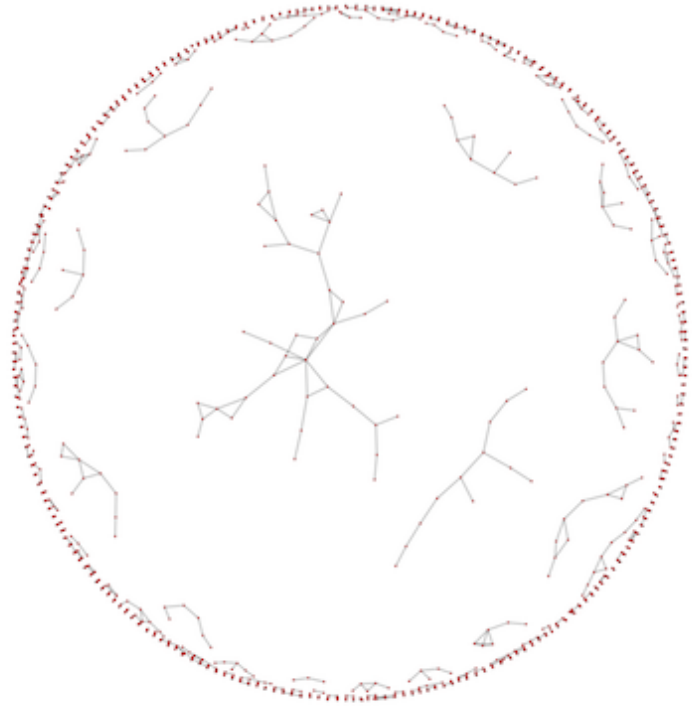


5-phone model

1-phone model



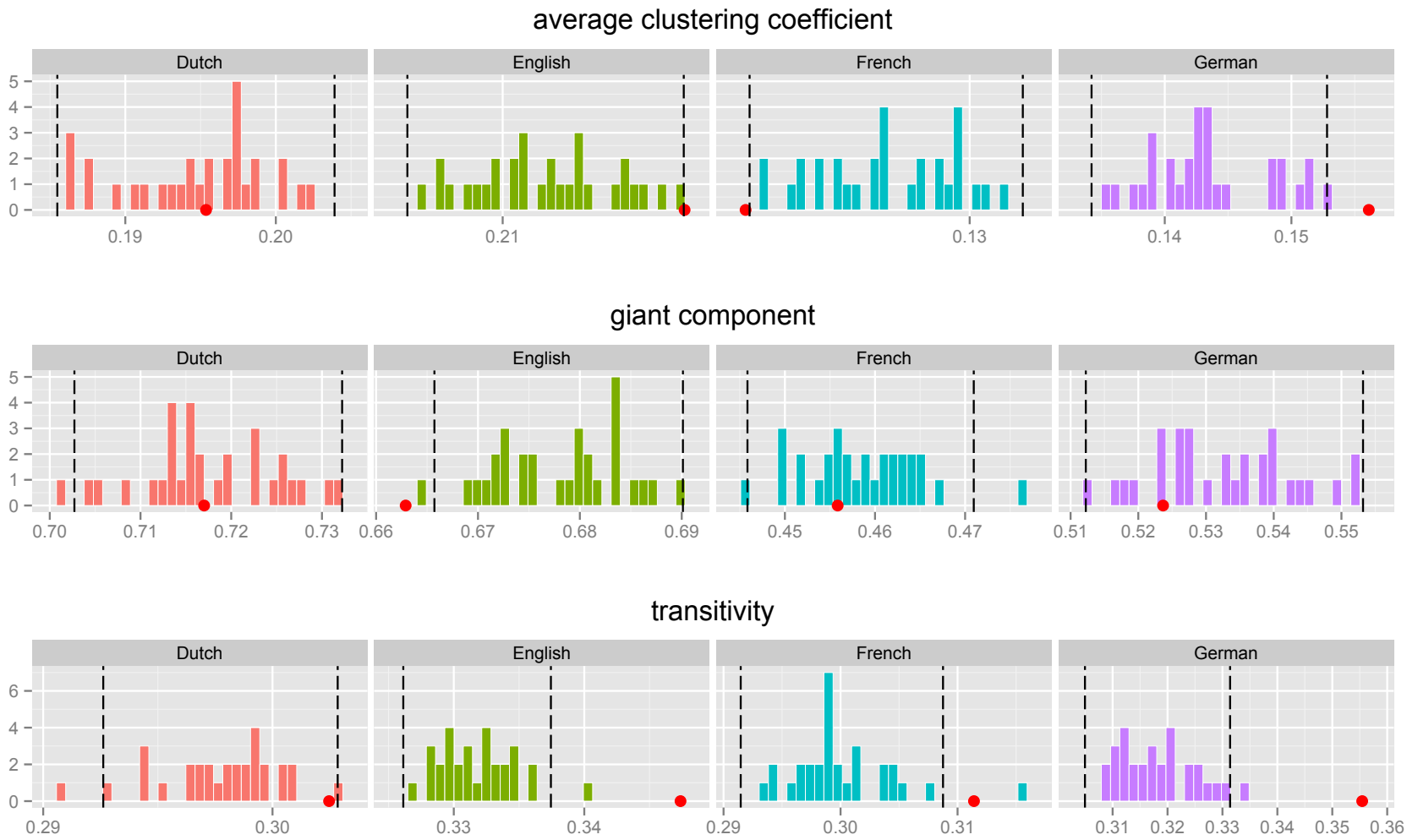
Real Lexicon



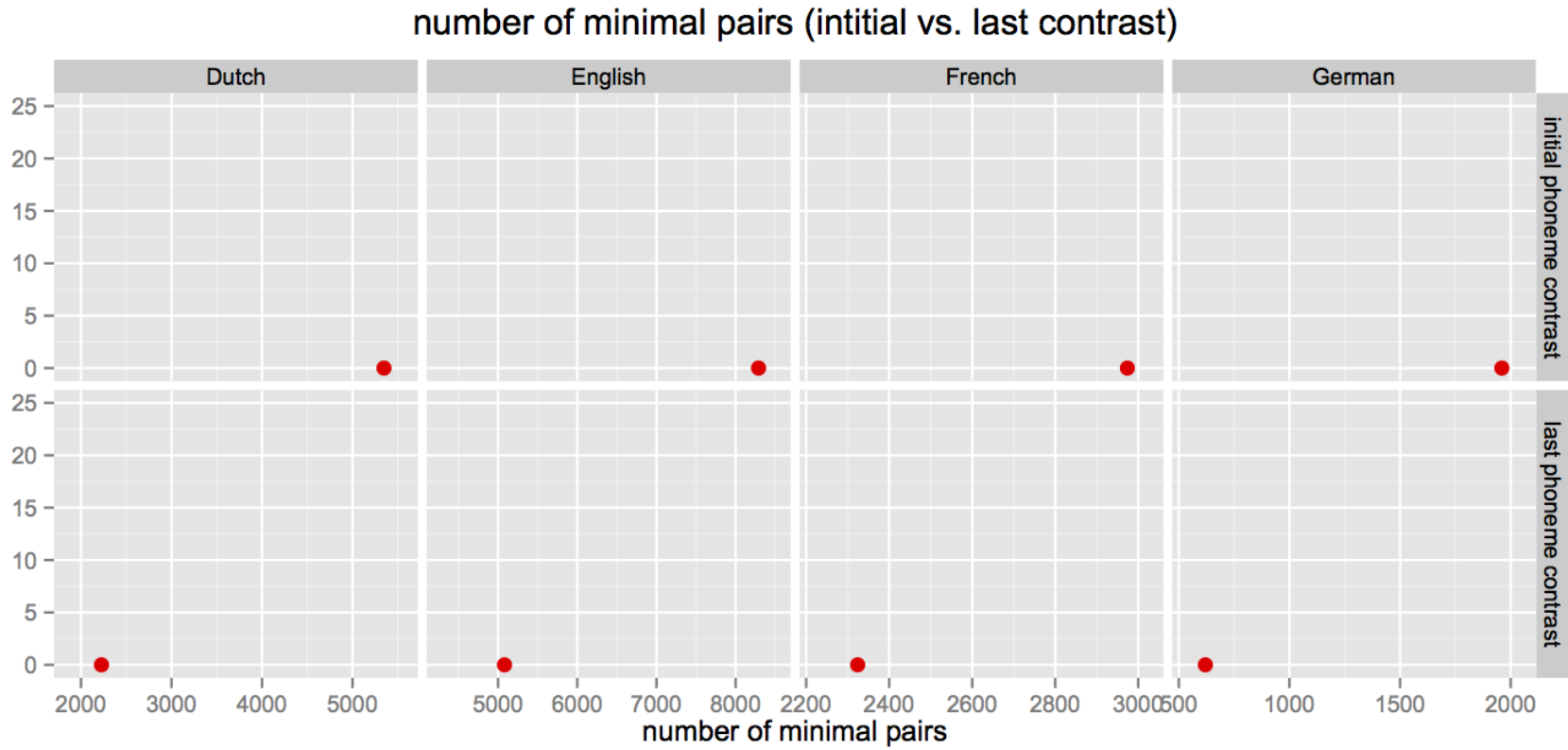
1-phone model

Number of minimal pairs by length

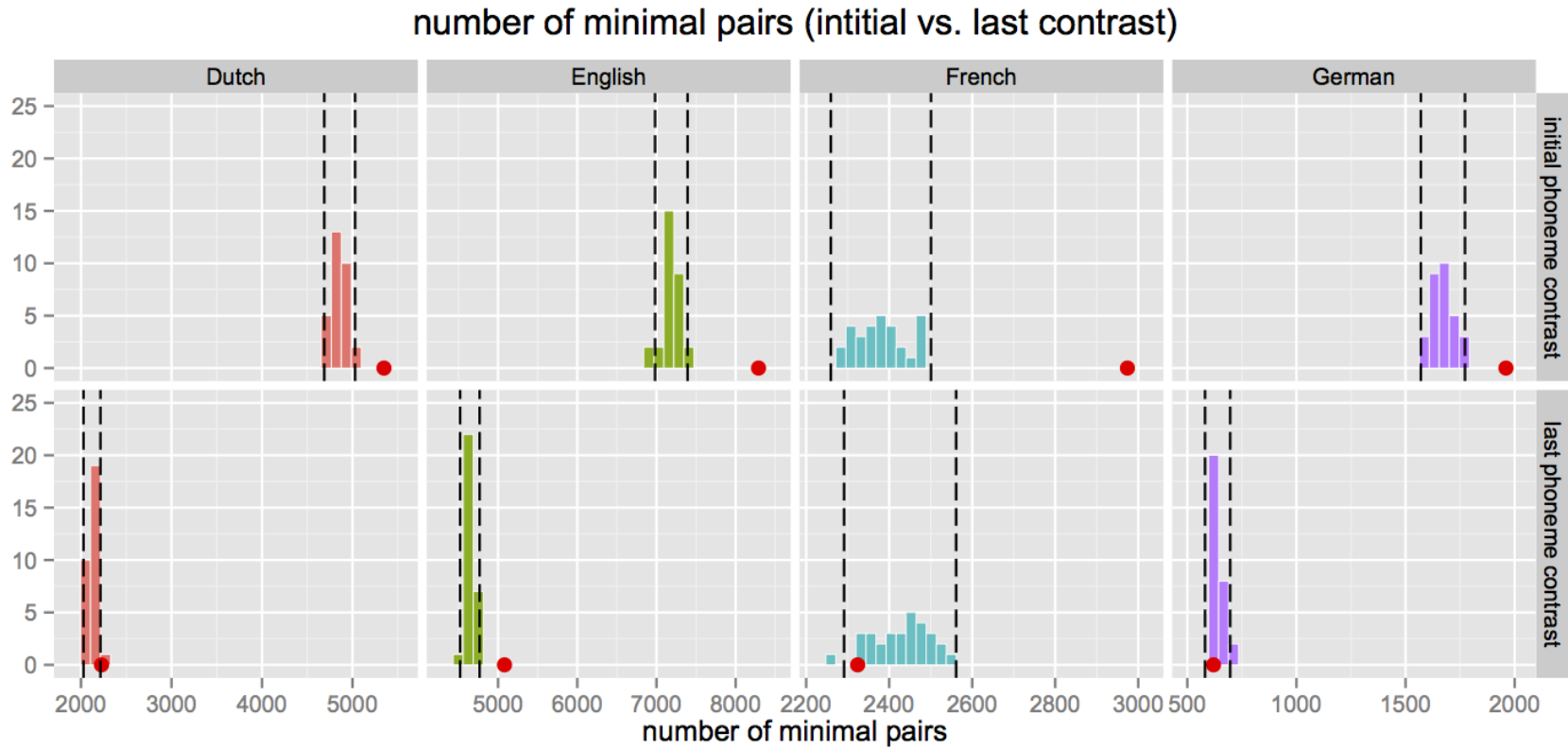
Network measures



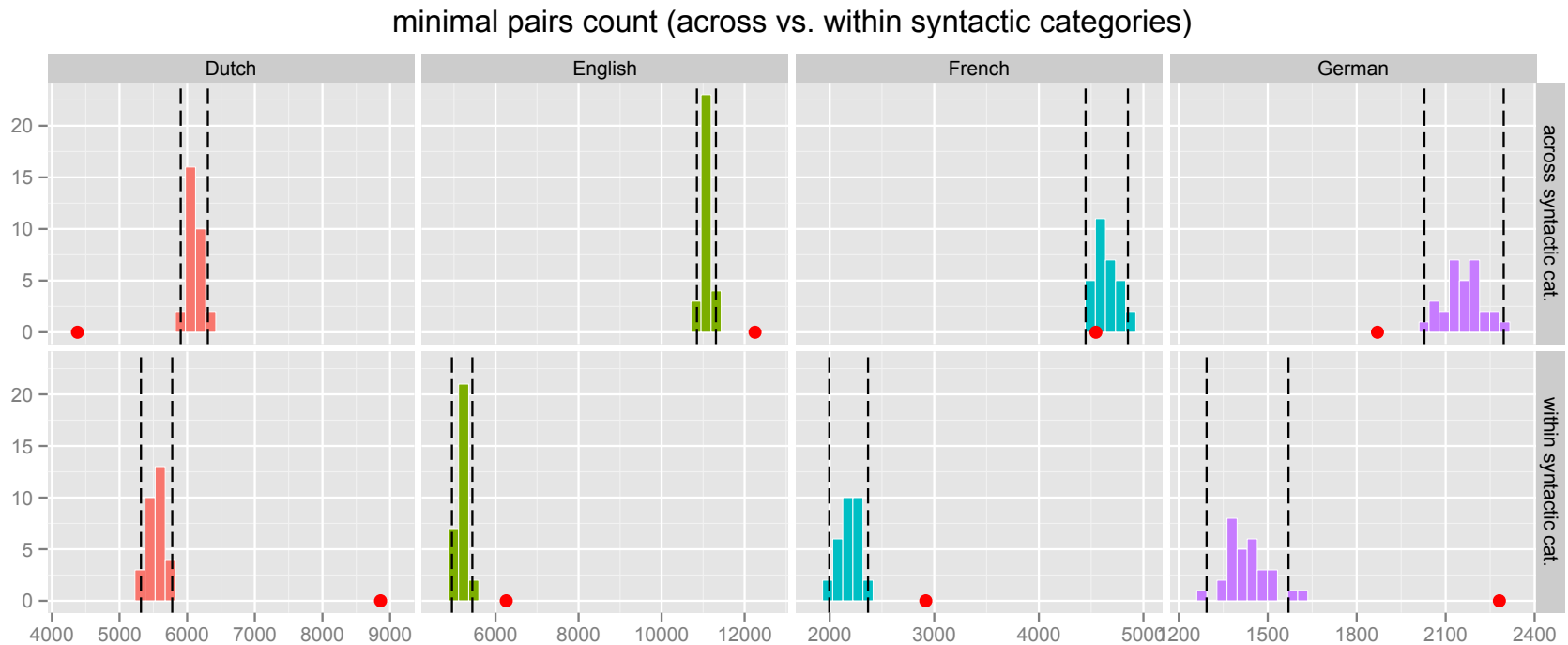
Word form similarity in word onsets



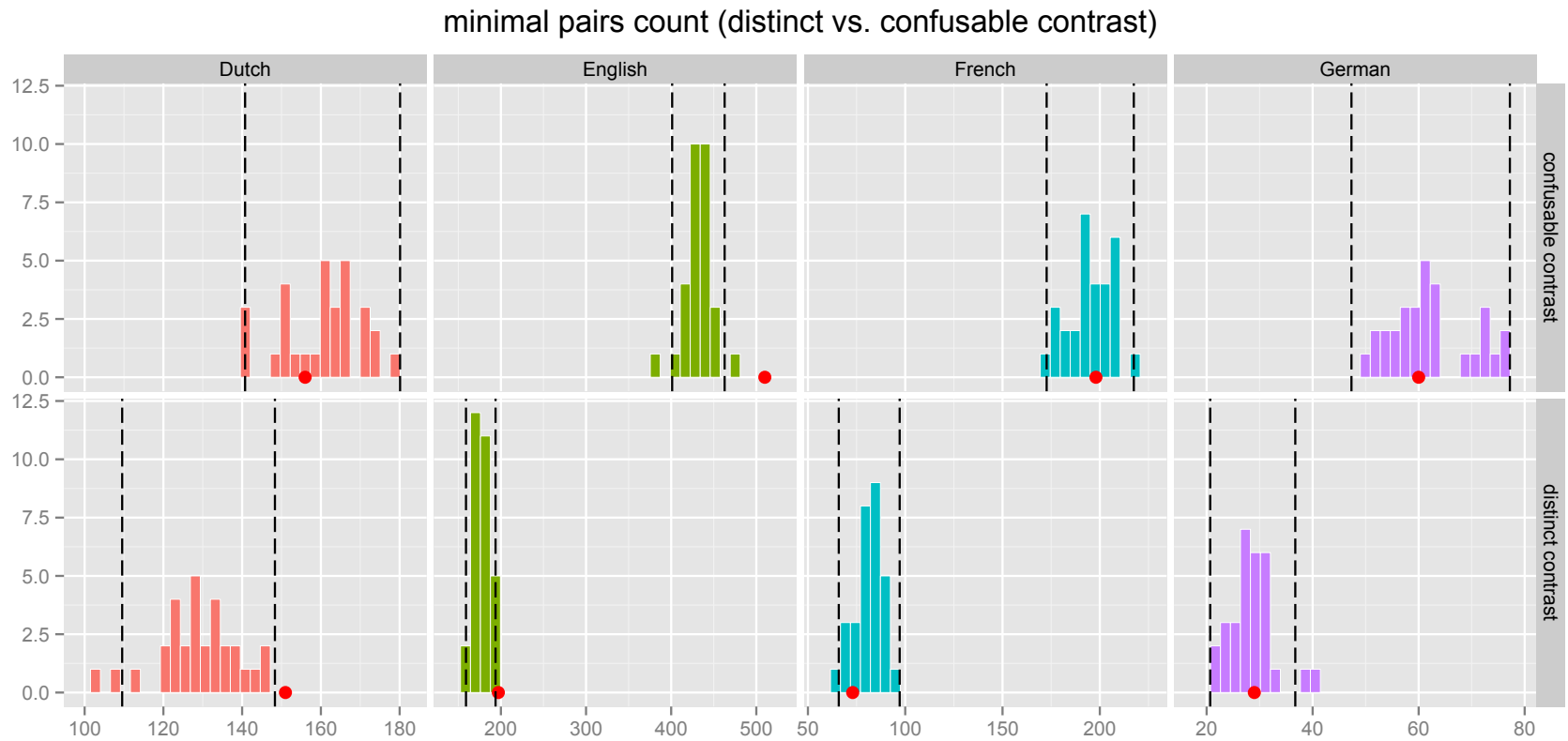
Word form similarity in word onsets



Word form similarity within and across grammatical categories



Confusable vs. non confusable minimal pairs



Semantic regularity in the lexicon

Is there a relationship between form and meaning for a large scale vocabulary?
(see also Monaghan et al., 2014)

Predictions:

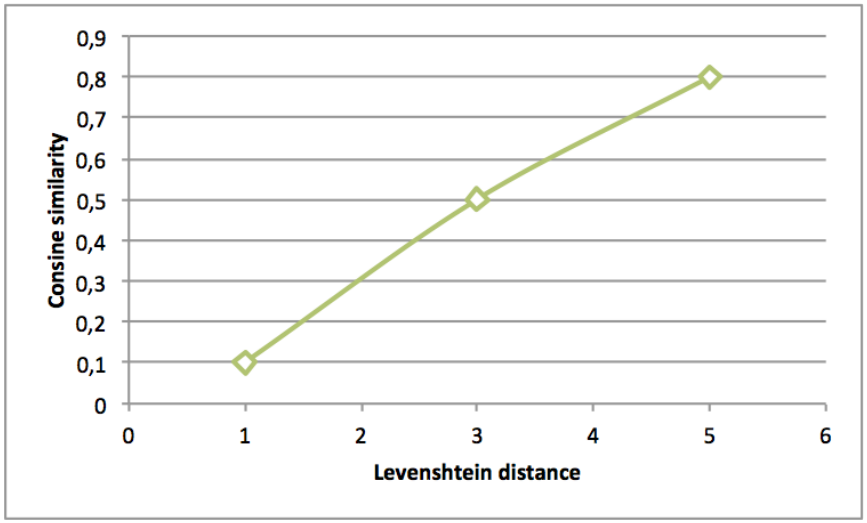
- **Dispersion**: phonologically similar words should be more semantically distant
- **Clumpiness**: phonologically similar words should be more semantically similar

Method:

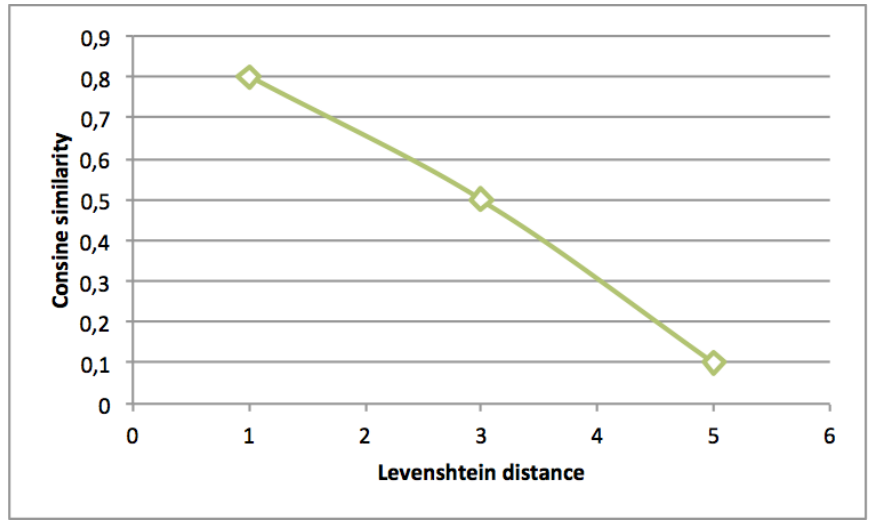
- 4 lexicons: Dutch, English German (CELEX), French (Lexique)
  all **monomorphemic** words (4000 to 6000 words)
- for each pairs of word of the same length, calculate:
  - its phonological similarity (edit string distance)
  - its semantic similarity (co-occurence vectors – LSA, but similar results with Wordnet)
    one word = one vector ; distance($word_1$, $word_2$) = cosine(v_1, v_2)

Possible results

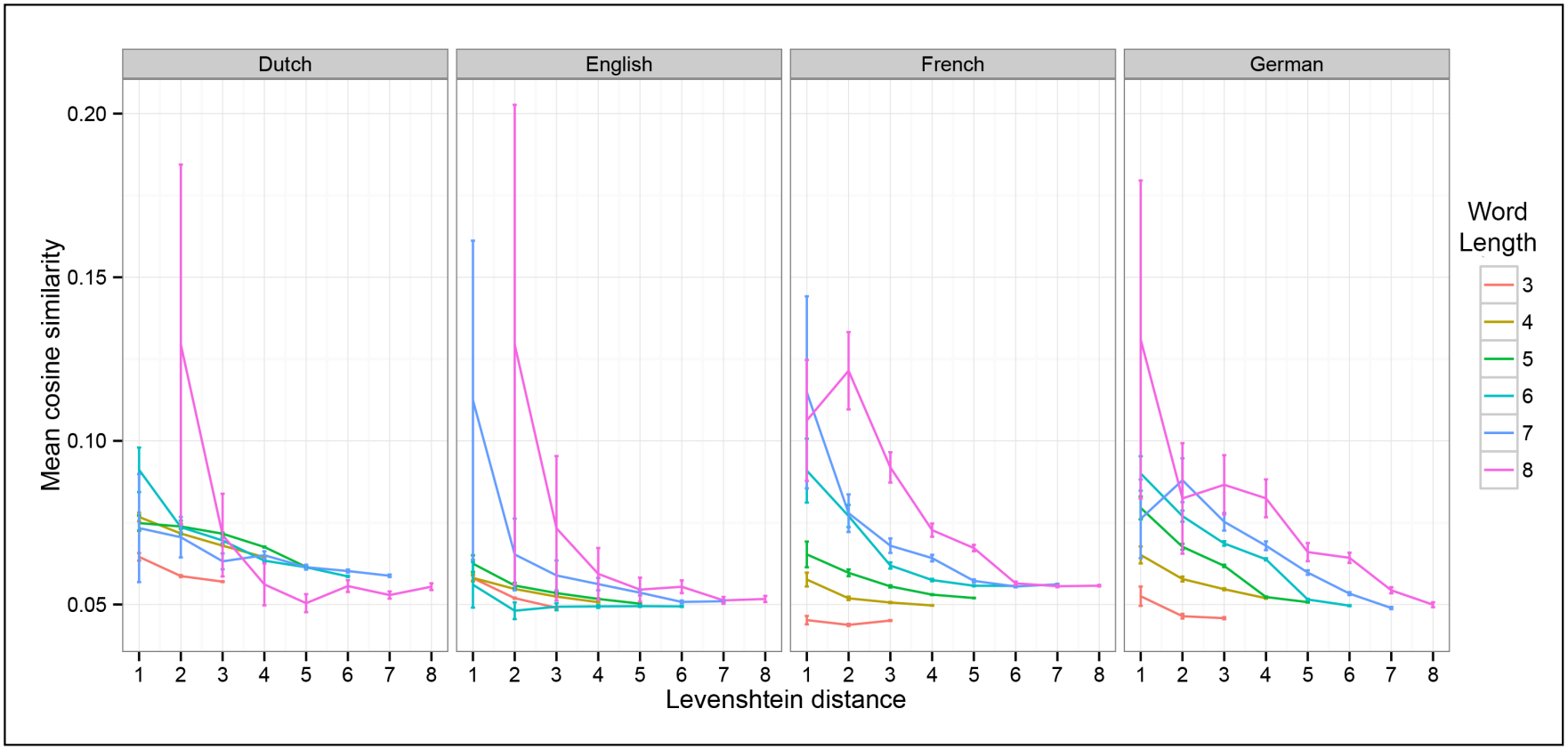
Dispersion
Phonologically similar words should be more semantically distant:



Clumpiness
Phonologically similar words should be more semantically similar:



Cosine similarity: the higher, the more semantically similar
Levenshtein distance: the lower, the more phonologically similar


Dutch
English
French
German
Mean cosine similarity
0.20
0.15
0.10
0.05
1 2 3 4 5 6 7 8
Levenshtein distance
Word Length
3
4
5
6
7
8

Results summary

High (but small) effect of systematicity in sound/meaning mappings among monomorphemic words across all 4 languages.

- Higher than expected by chance alone?
- A general property or an effect driven by small clusters of words? (e.g., *gl-* words) see Monaghan et al (2014)
- A consequence of etymology between words?

The trend for semantically similar words to be phonologically similar is evidence for lexical clumpiness.