Efficient phonological clustering in the mental lexicon

Kyle Mahowald, Isabelle Dautriche, Edward Gibson,
Anne Christophe, Steven T. Piantadosi

What we know about the lexicon

cat
kakis
gato
pusa
kottur
chat
Katze

What we know about the lexicon

- The relationship between word forms and their meanings is **arbitrary**.

cat
kakis
gato
pusa
kottur
chat
Katze

What we know about the lexicon

- The relationship between word forms and their meanings is **arbitrary**.
- As a result, there are many free parameters left for how we choose our word forms.

cat
kakis
gato
pusa
kottur
chat
Katze

What we know about the lexicon

- The relationship between word forms and their meanings is **arbitrary**.
- 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:

cat
kakis
gato
pusa
kottur
chat
Katze

What we know about the lexicon

- The relationship between word forms and their meanings is **arbitrary**.
- 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**

cat
kakis
gato
pusa
kottur
chat
Katze

What we know about the lexicon

- The relationship between word forms and their meanings is **arbitrary**.
- 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**
    - Efficient codes in the presence of noise have maximum distance between codewords

cat
kakis
gato
pusa
kottur
chat
Katze

What we know about the lexicon

- The relationship between word forms and their meanings is **arbitrary**.
- 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**
    - Efficient codes in the presence of noise have maximum distance between codewords
  - **ease of processing and learnability**

cat
kakis
gato
pusa
kottur
chat
Katze

What we know about the lexicon

- The relationship between word forms and their meanings is **arbitrary**.
- 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**
    - Efficient codes in the presence of noise have maximum distance between codewords
  - **ease of processing and learnability**
    - Easier to remember and produce things that are similar to things you already know

cat
kakis
gato
pusa
kottur
chat
Katze

Lexical design: Clumpish

Lexical design: Clumpish

Lexical design: Clumpish

fep

‘COLD’

Lexical design: Clumpish

fep
‘COLD’

feb
‘HOT’

Lexical design: Clumpish

fep ‘COLD’  feb ‘HOT’

- Allows you to re-use sounds you already know which makes it easy to learn and process, but you might get confused.

Lexical design: Sparsese

Lexical design: Sparsese

Lexical design: Sparsese

fep

‘COLD’

Lexical design: Sparsese

fep
‘COLD’

dax
‘HOT’

Lexical design: Sparsese

fep ‘COLD’  dax ‘HOT’

- More optimal from efficient coding perspective
- But you have to learn and store a new set of sounds.

Overview

Overview

- Two competing pressures:

Overview

- Two competing pressures:
  - A pressure for **dispersion** (communicative efficiency)

Overview

- Two competing pressures:
  - A pressure for **dispersion** (communicative efficiency)
  - A pressure for **clumpiness** (systematicity, learnability, memory)

Overview

- Two competing pressures:
  - A pressure for **dispersion** (communicative efficiency)
  - A pressure for **clumpiness** (systematicity, learnability, memory)

- **Question: Is there evidence for dispersion or clumpiness of word forms in the lexicon?**

Overview

- Two competing pressures:
  - A pressure for **dispersion** (communicative efficiency)
  - A pressure for **clumpiness** (systematicity, learnability, memory)
- **Question: Is there evidence for dispersion or clumpiness of word forms in the lexicon?**
- Two approaches:

Overview

- Two competing pressures:
  - A pressure for **dispersion** (communicative efficiency)
  - A pressure for **clumpiness** (systematicity, learnability, memory)
- **Question: Is there evidence for dispersion or clumpiness of word forms in the lexicon?**
- Two approaches:
  - Large-scale analysis of lexicons

Overview

- Two competing pressures:
  - A pressure for **dispersion** (communicative efficiency)
  - A pressure for **clumpiness** (systematicity, learnability, memory)

- **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

Large-scale analysis

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

Large-scale analysis

- If the lexicon is optimized in any direction, **frequent word forms** should be more optimized because they are used more.
- Can a pressure for dispersion /clumpiness be found in the frequency patterns of word use? (see Frauenfelder et al., 1993)

Large-scale analysis

- If the lexicon is optimized in any direction, **frequent word forms** should be more optimized because they are used more.
- Can a pressure for dispersion /clumpiness be found in the frequency patterns of word use? (see Frauenfelder et al., 1993)
  - **Dispersion**: the most frequent words should be more unique phonologically

Large-scale analysis

- If the lexicon is optimized in any direction, **frequent word forms** should be more optimized because they are used more.
- Can a pressure for dispersion /clumpiness be found in the frequency patterns of word use? (see Frauenfelder et al., 1993)
  - **Dispersion**: the most frequent words should be more unique phonologically
  - **Clumpiness**: the most frequent words should share more sequences

Large-scale analysis

Large-scale analysis

- Method:

Large-scale analysis

- Method:
  - orthographic lexicons from 115 languages extracted from Wikipedia (top 20k most frequent tokens)

Large-scale analysis

- Method:
  - orthographic lexicons from 115 languages extracted from Wikipedia (top 20k most frequent tokens)
  - For each word type, calculate:

Large-scale analysis

- Method:
  - orthographic lexicons from 115 languages extracted from Wikipedia (top 20k most frequent tokens)
- For each word type, calculate:
  - number of **neighbors** (car/cat and cat/cats)

Large-scale analysis

- Method:
  - orthographic lexicons from 115 languages extracted from Wikipedia (top 20k most frequent tokens)
- For each word type, calculate:
  - number of **neighbors** (car/cat and cat/cats)
  - token **frequency**

Large-scale analysis

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

Possible results

Possible results

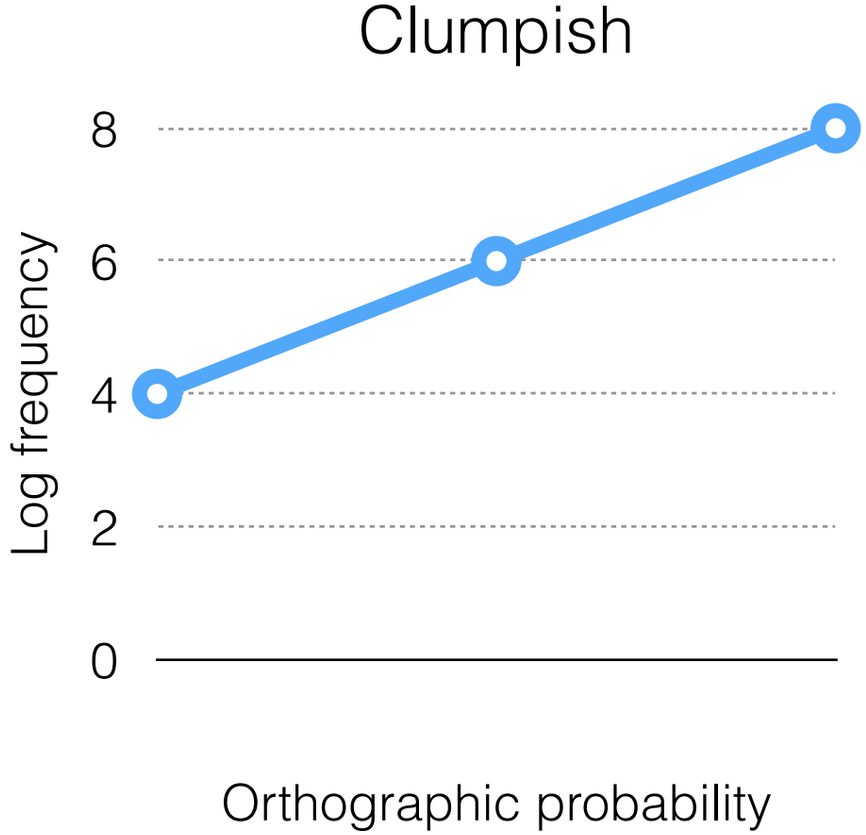
- Frequent words have more neighbors and higher probability

Possible results

- Frequent words have more neighbors and higher probability
- Frequent words have fewer neighbors and lower probability

Possible results

- Frequent words have more neighbors and higher probability

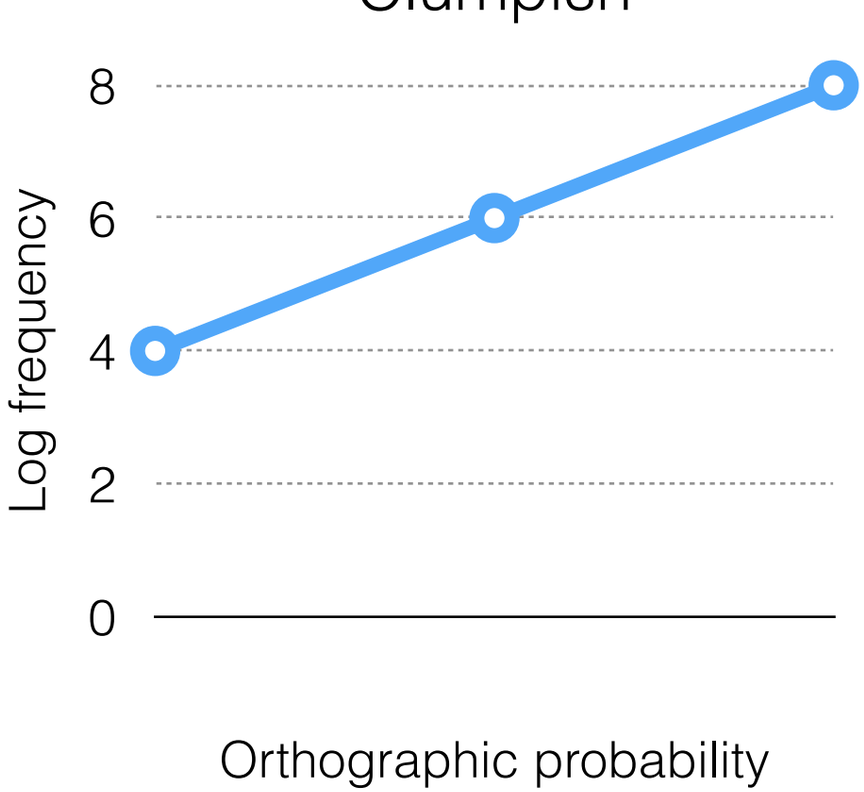


- Frequent words have fewer neighbors and lower probability

Possible results

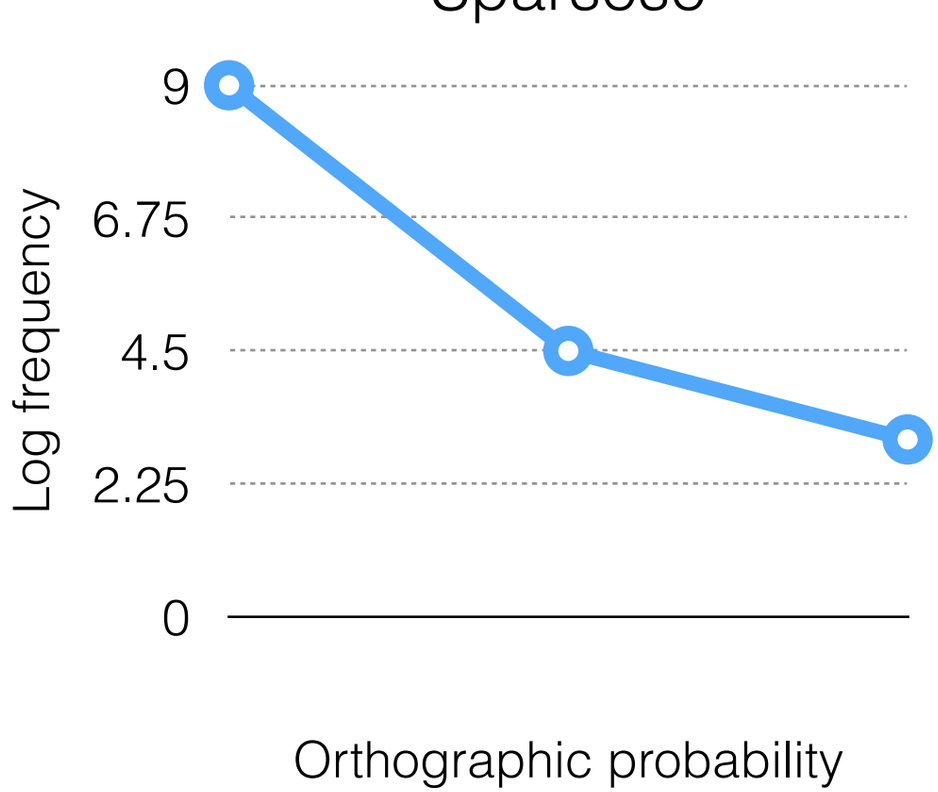
- Frequent words have more neighbors and higher probability

Clumpish

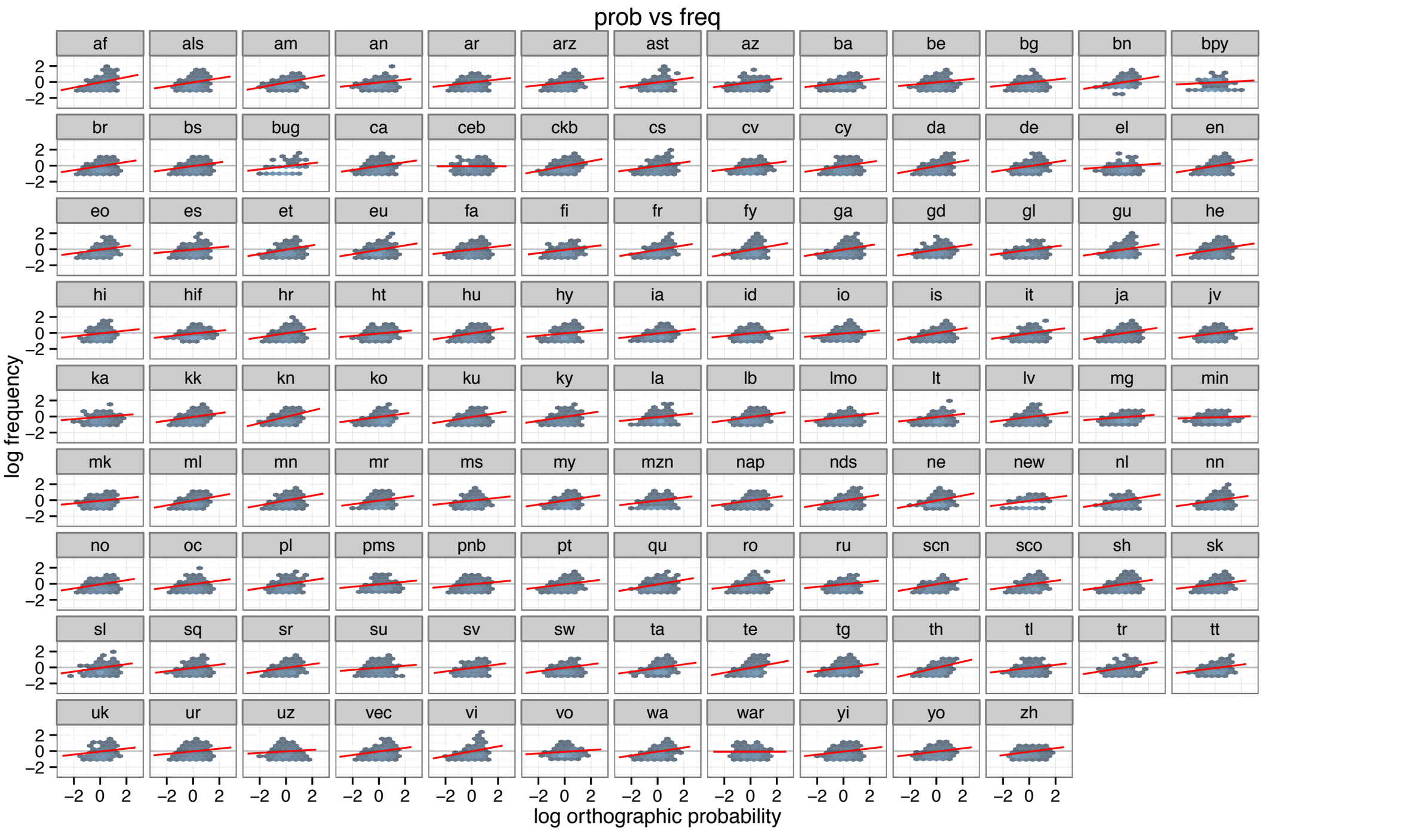


- Frequent words have fewer neighbors and lower probability

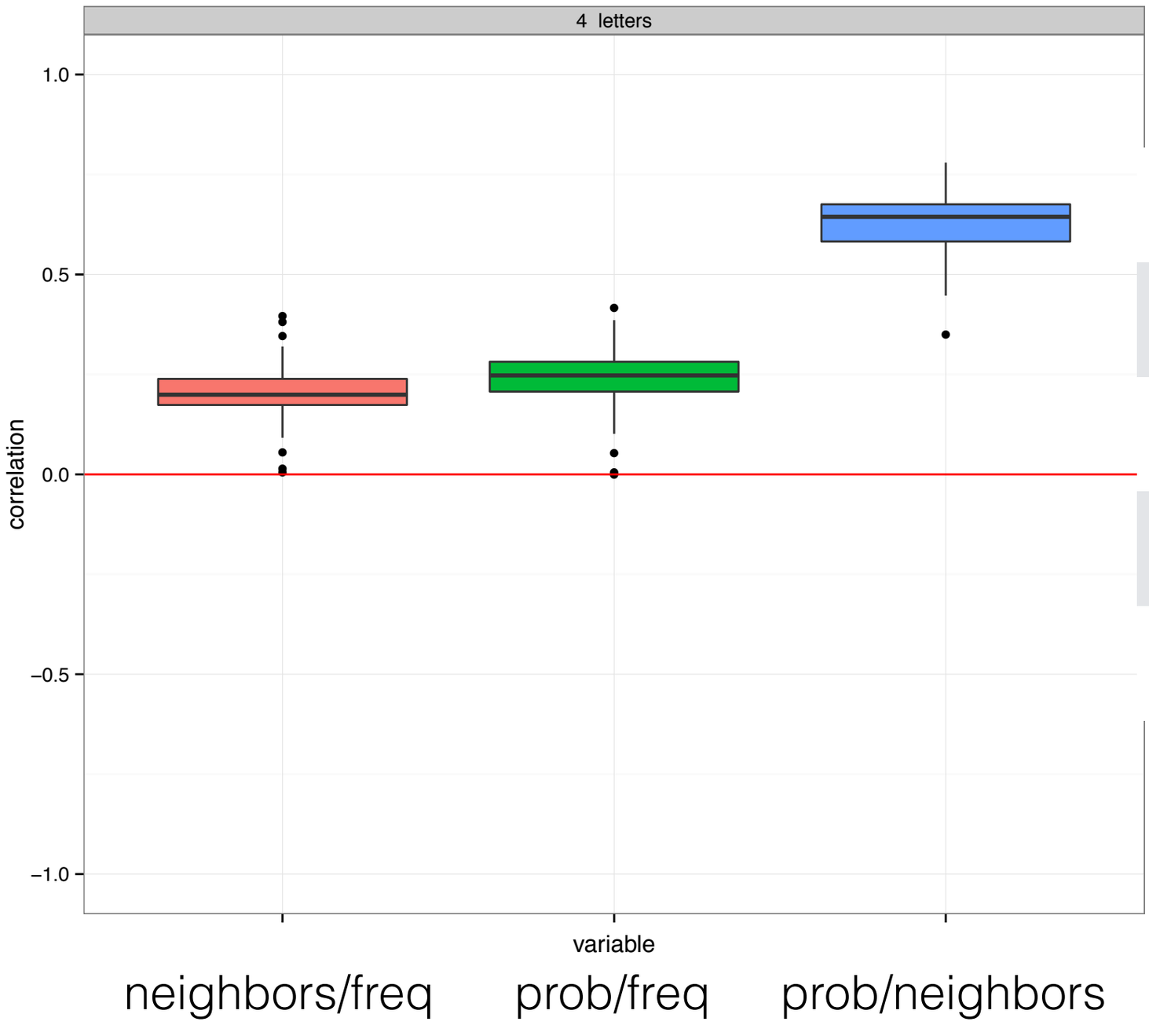
Sparsese



Wikipedia (4-letter words)



Wikipedia results



	mean	pos.	sig
neighbors/ freq	0.21	0.97	0.96
prob/freq	0.24	0.97	0.99
prob/ neighbors	0.63	1.00	1.00

(other lengths similar)

Results summary

- Across 115 languages, we found:
  - an obvious correlation between orthographic probability and number of neighbors (more common words have more neighbors)
  - correlations between number of neighbors and frequency
- **This suggests a pressure for clumpiness that becomes stronger with use.**
- Possible confound: morphemes
- Yet this effect is predicted by a model of word generation: more probable strings will be generated more often and higher string probabilities give rise to more neighbors.
  - We need a proper baseline.

Simulating “null” lexicons

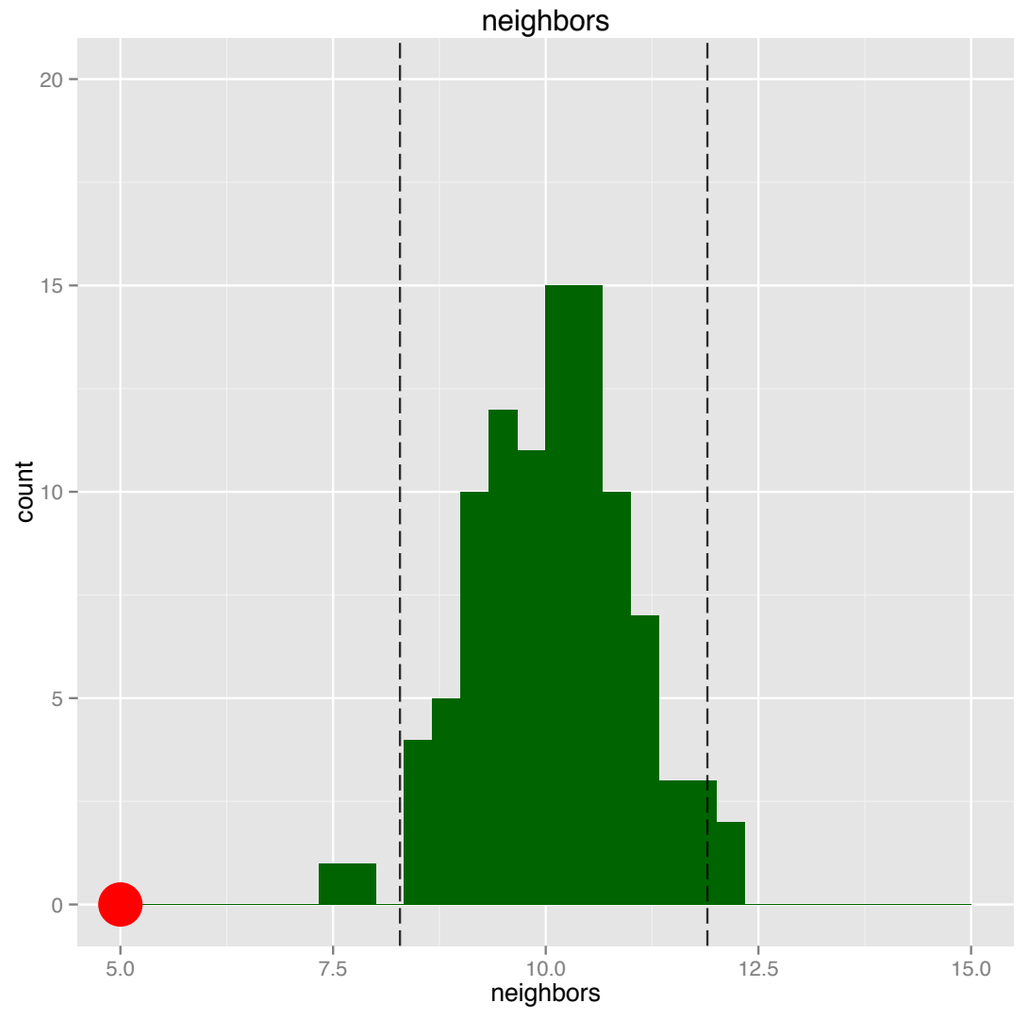
- Is the lexicon clumpier than what would be expected by chance alone?
  - How do we determine what a baseline is?
- New method
  - Generate “null” lexicons for which there is no pressure for dispersion or clumpiness, beyond phonotactics
  - Compare real lexicons to these null lexicons on a variety of lexical measures to see how they are different.

Method

- **Real lexicons (phonemic transcriptions)** are **monomorphemic** lemmas from Dutch, English, German (CELEX), French (Lexique)
  - 4000 to 6000 words, no homophones, freq > 0
- **Models of phonotactically controlled lexicons**
  - n-phone models (n=1 to n=6)
  - n-syllable models (n=1 to n=2)
  - PCFG over syllables
- All with back off and smoothing, all matched for length to the real lexicon
- Sample 30 “null” lexicons for each
- **Select the best model**
  - Each model trained on 75% of the real lexicon and evaluated on the remaining 25%
- **Best model: 5-phone model** (50% of generated words, on average, are real)
- **Measure of word form similarity:** number of minimal pairs, average Levenshtein distance (edit), network measures..

Hypothesis 1: Sparsese

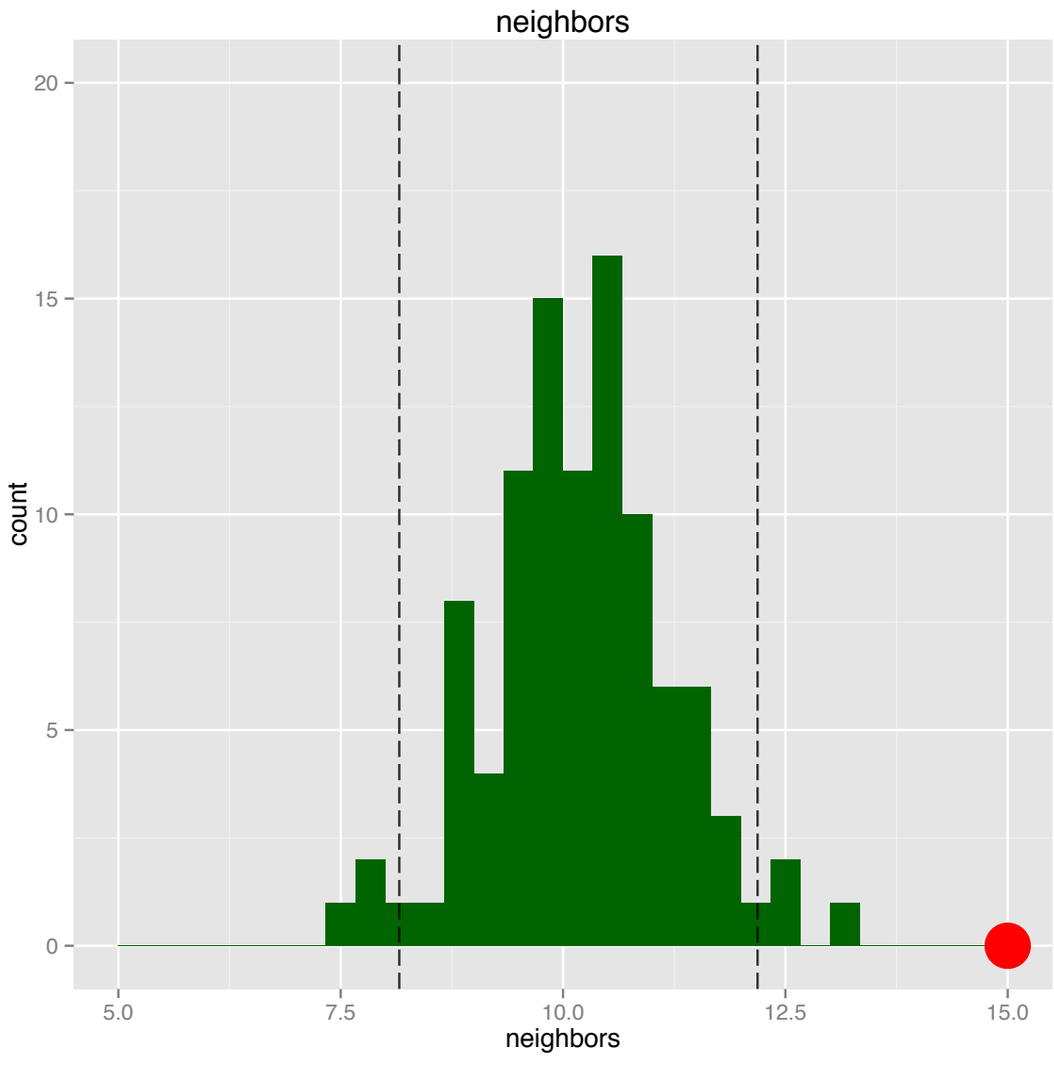
- Neighbors and minimal pairs (pig/big) are avoided in order to create a sparse code.



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

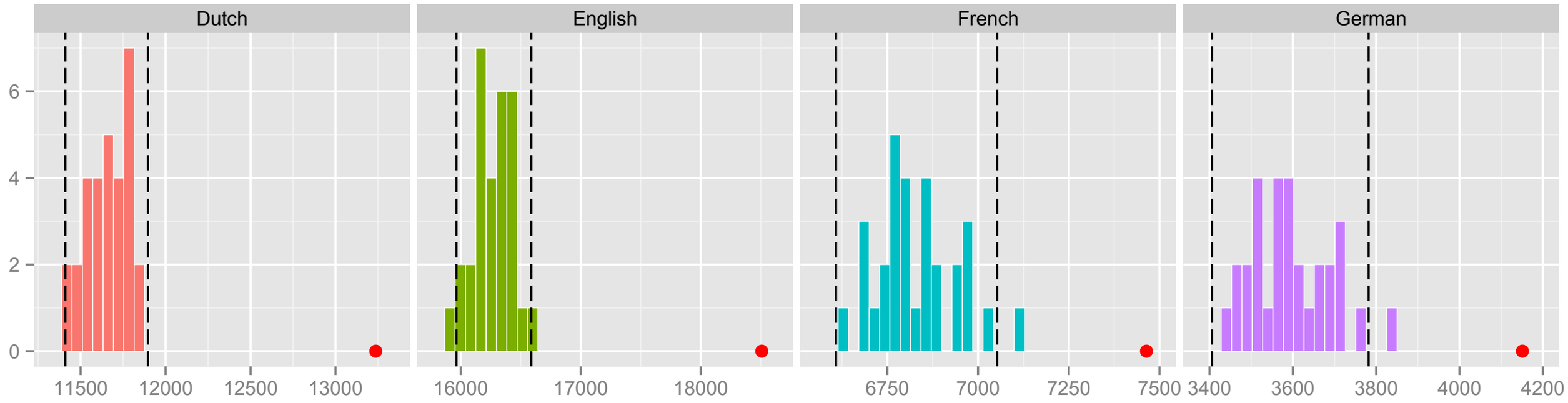
Hypothesis 2: Clumpish

- Minimal pairs and neighbors are more likely due to pressure for re-use and regularity (i.e., having *big* in the lexicon makes *pig* more likely)



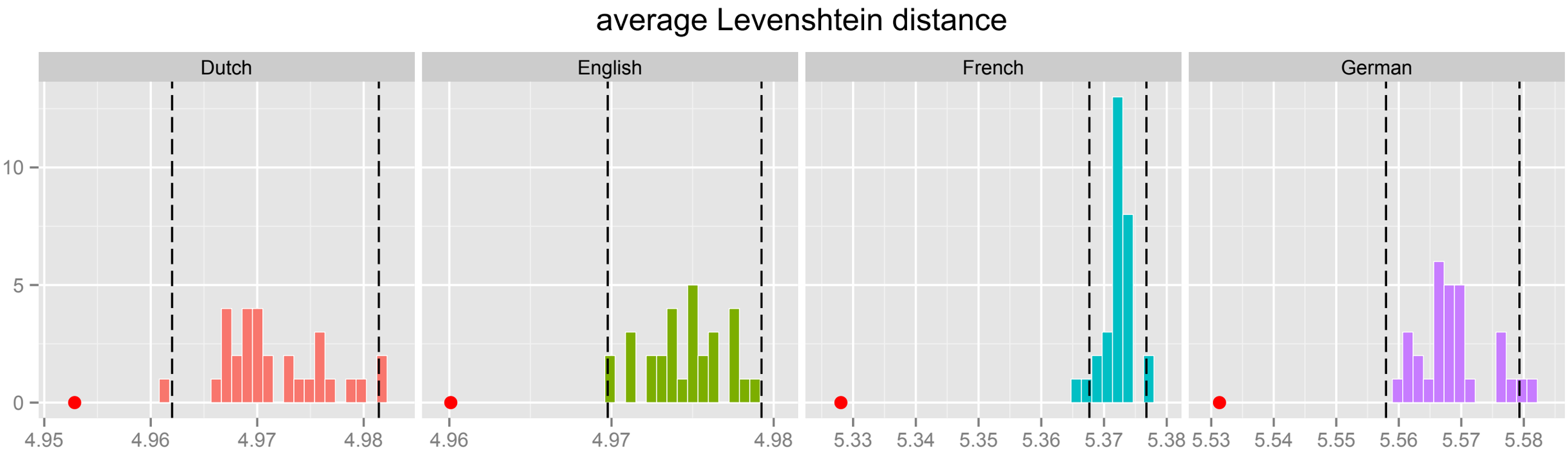
red dot: number of minimal pairs in the real lexicon
green histogram: the distribution of the number of minimal pairs across simulated lexicons
dotted line: 95% confidence interval

Result: number of minimal pairs



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

Another measure: Average Levenshtein distance



- Word forms are more similar in the real lexicon.
- Additional measures (network transitivity, clustering coefficient) tend towards the same conclusion: the lexicon is clumpy.

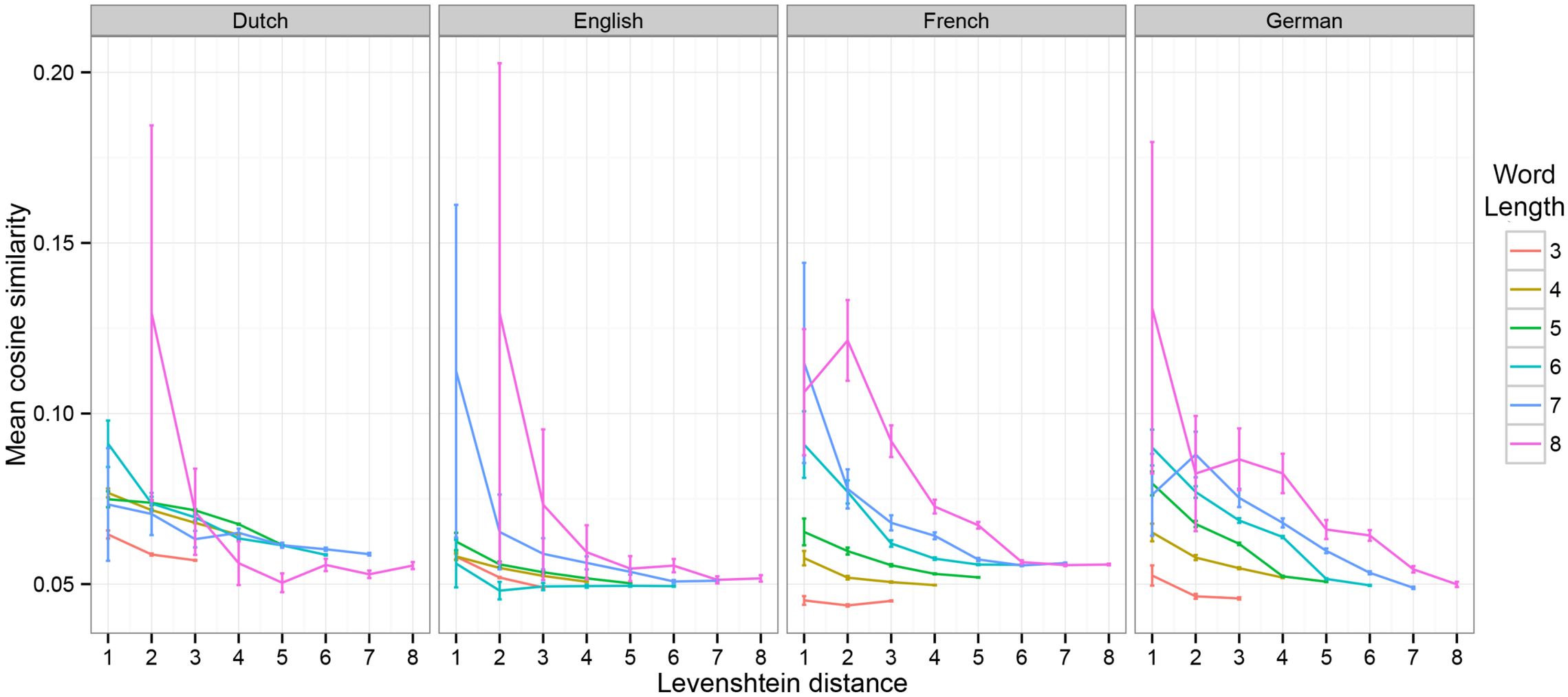
Null lexicon summary

- Most of the measures reviewed suggests that **the lexicon is clumpier than expected by chance.**
  - **Not** the result of morphological regularity since we focused on **monomorphemes**
  - Not the result of sound-to-sound transition probability controlled for by model

Possible role of semantics

- Is there a relationship between form and meaning at the large scale? (cf. Monaghan, et al. 2014)
- 4 lexicons, calculate for each word pair:
  - phonological edit distance
  - semantic similarity using co-occurrence vectors from LSA

Possible role of semantics



Even antonyms are clumpy

fep
'COLD'

feb
'HOT'

- 42 English antonym pairs (happy/sad) that are not morphologically related. Look at Levenshtein distance compared to all other words in set of 84.

- The average Levenshtein distance between a word and its antonym was 3.64 compared to 3.98 between a word and a non-antonym. Significant by mixed effect model (β=-.08, t=-4.11, $p < .001$).

Global summary

- Large-scale analysis of Wikipedia: 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

Thanks!

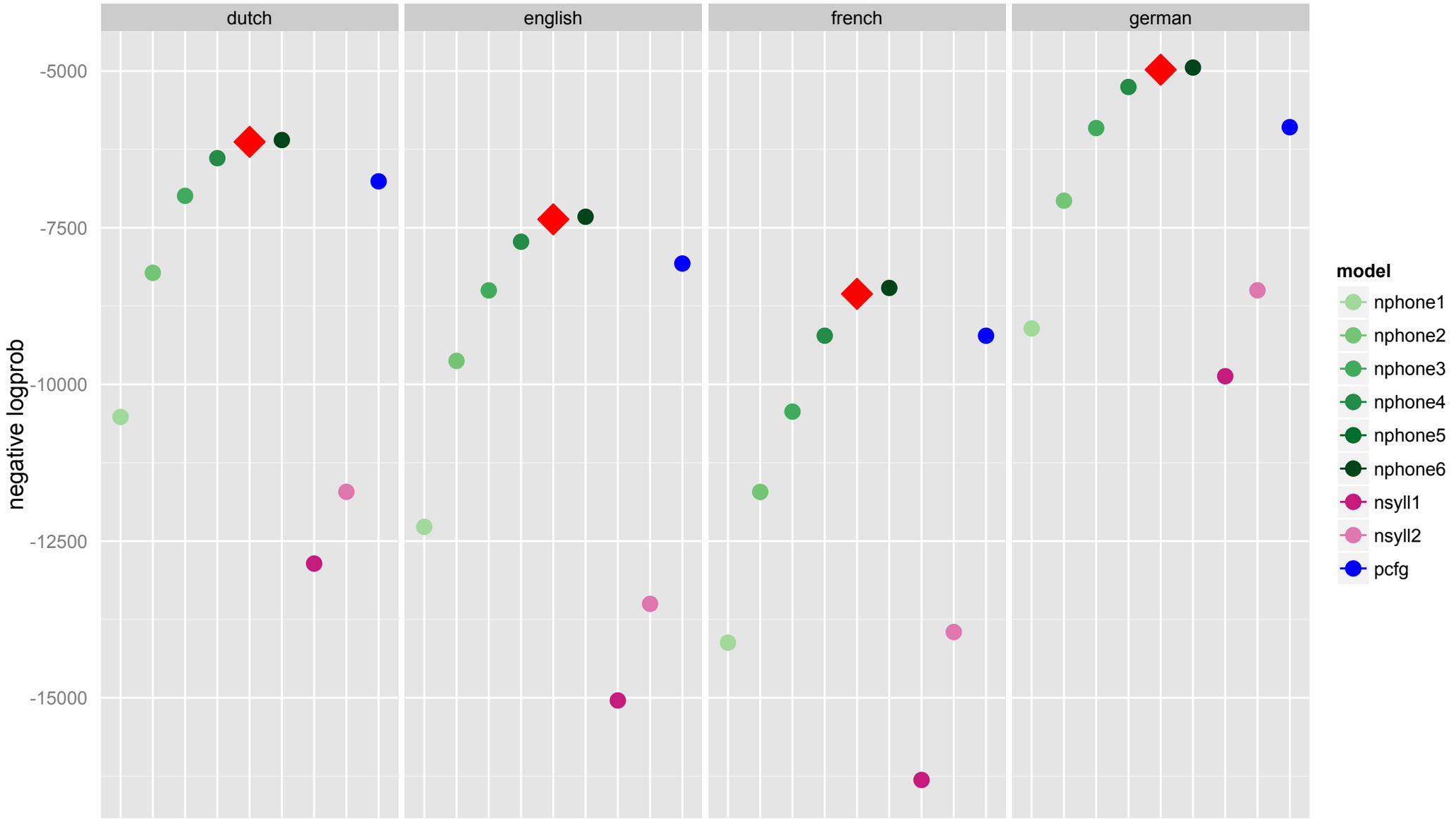



- Collaborators: Isabelle Dautriche, Ted Gibson, Anne Christophe, Steve Piantadosi

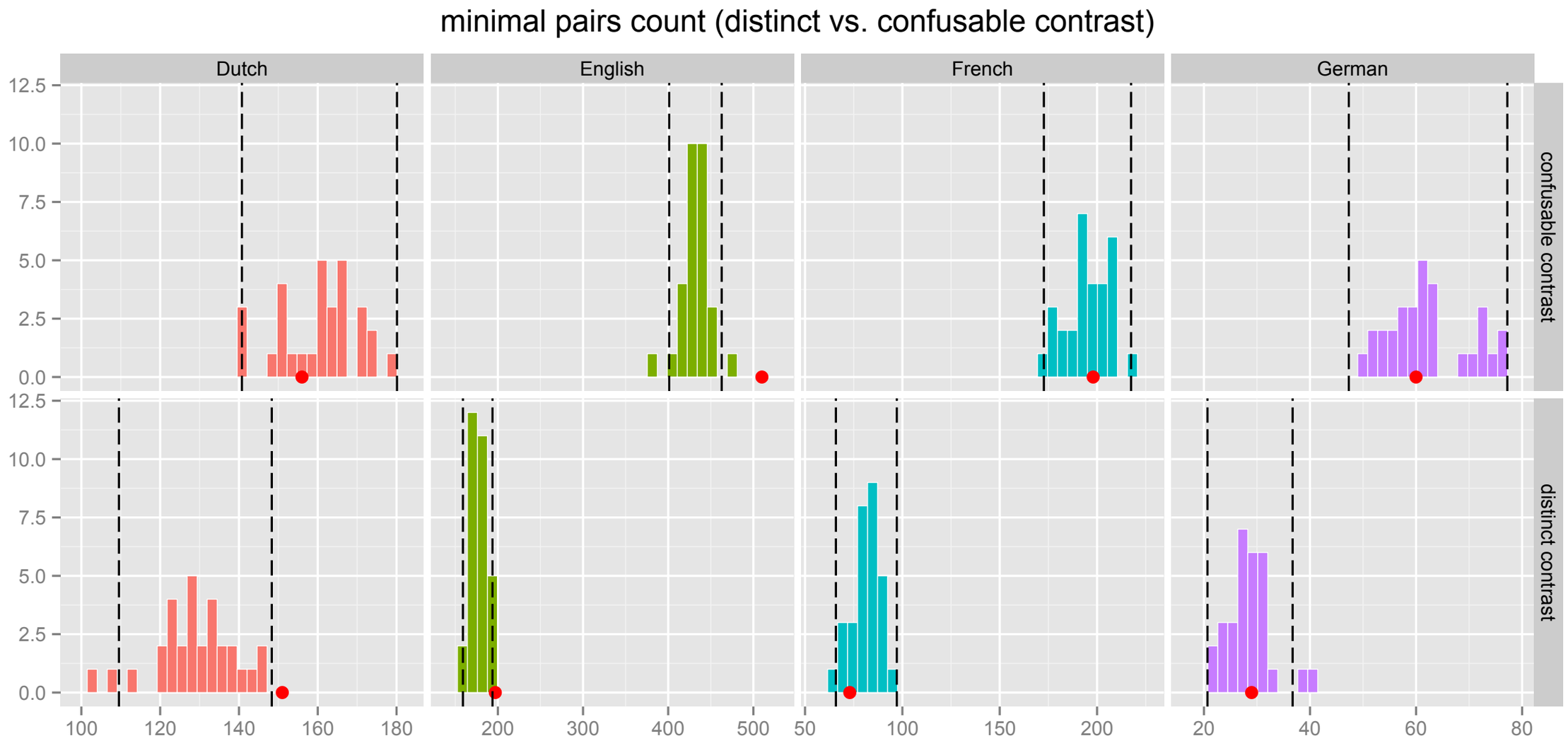
- Members of Tedlab

- Audience at AMLaP 2014

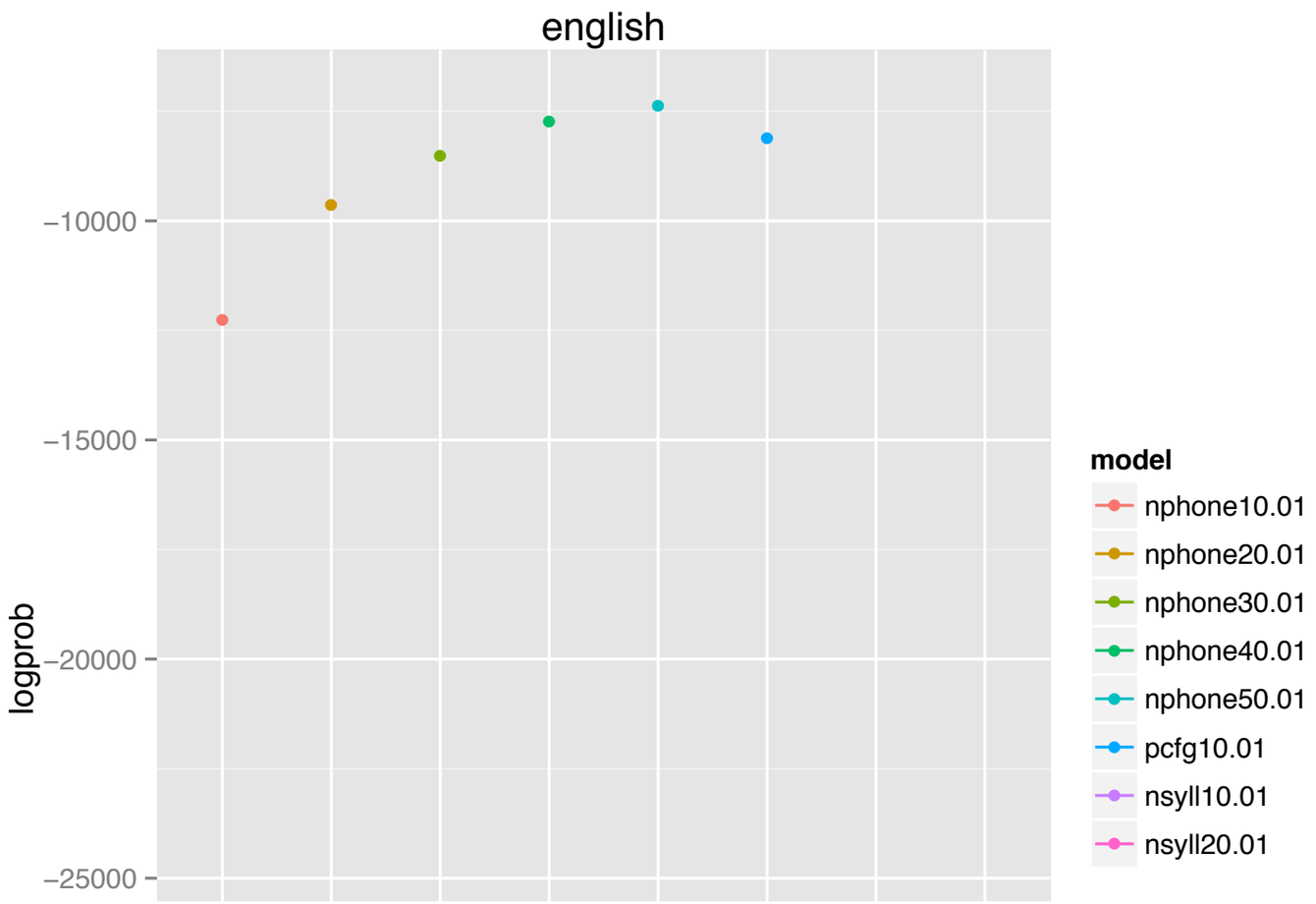
Evaluation


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

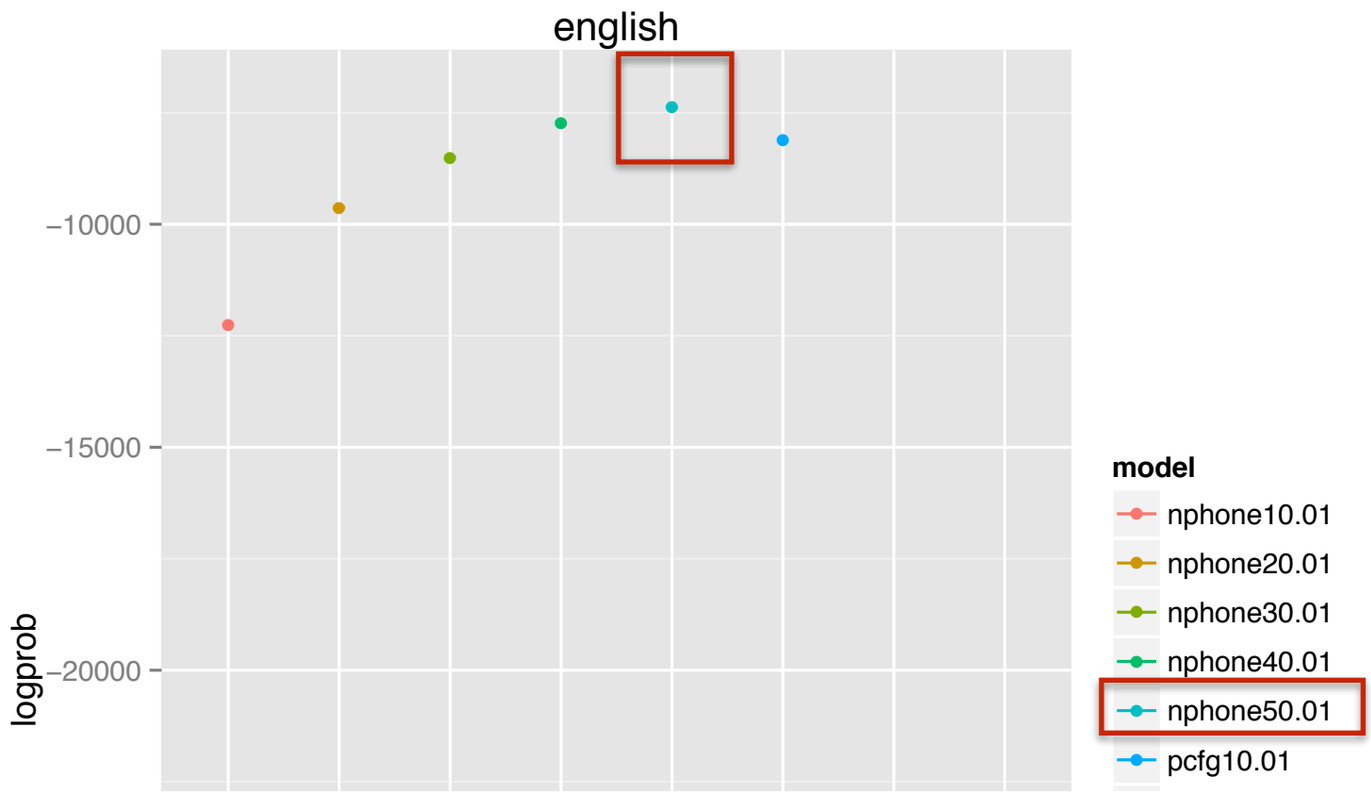
Confusable vs. non-confusable MPS



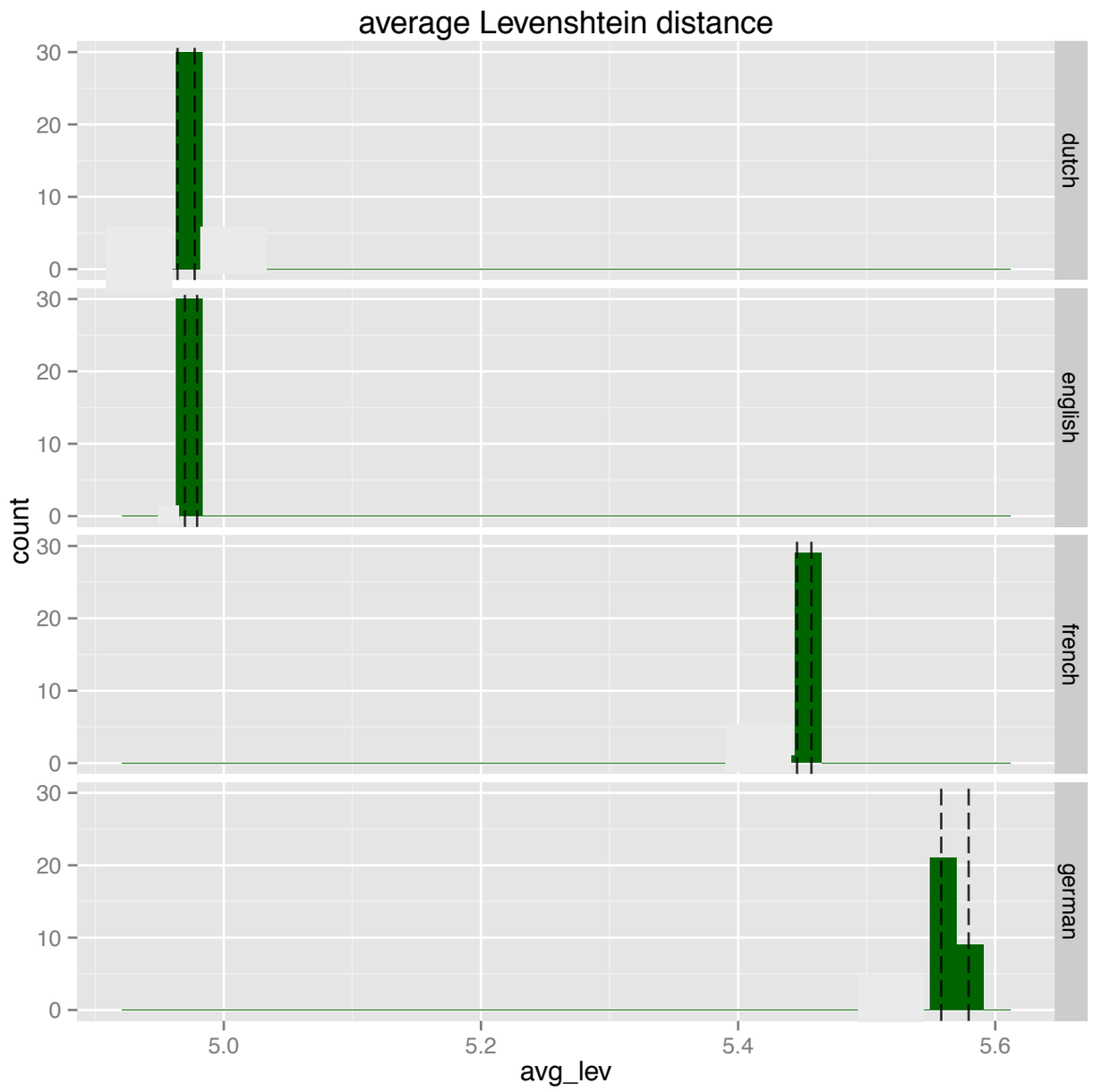
Clumpiness and semantic factors



Evaluating models (train on .75; test on .25)

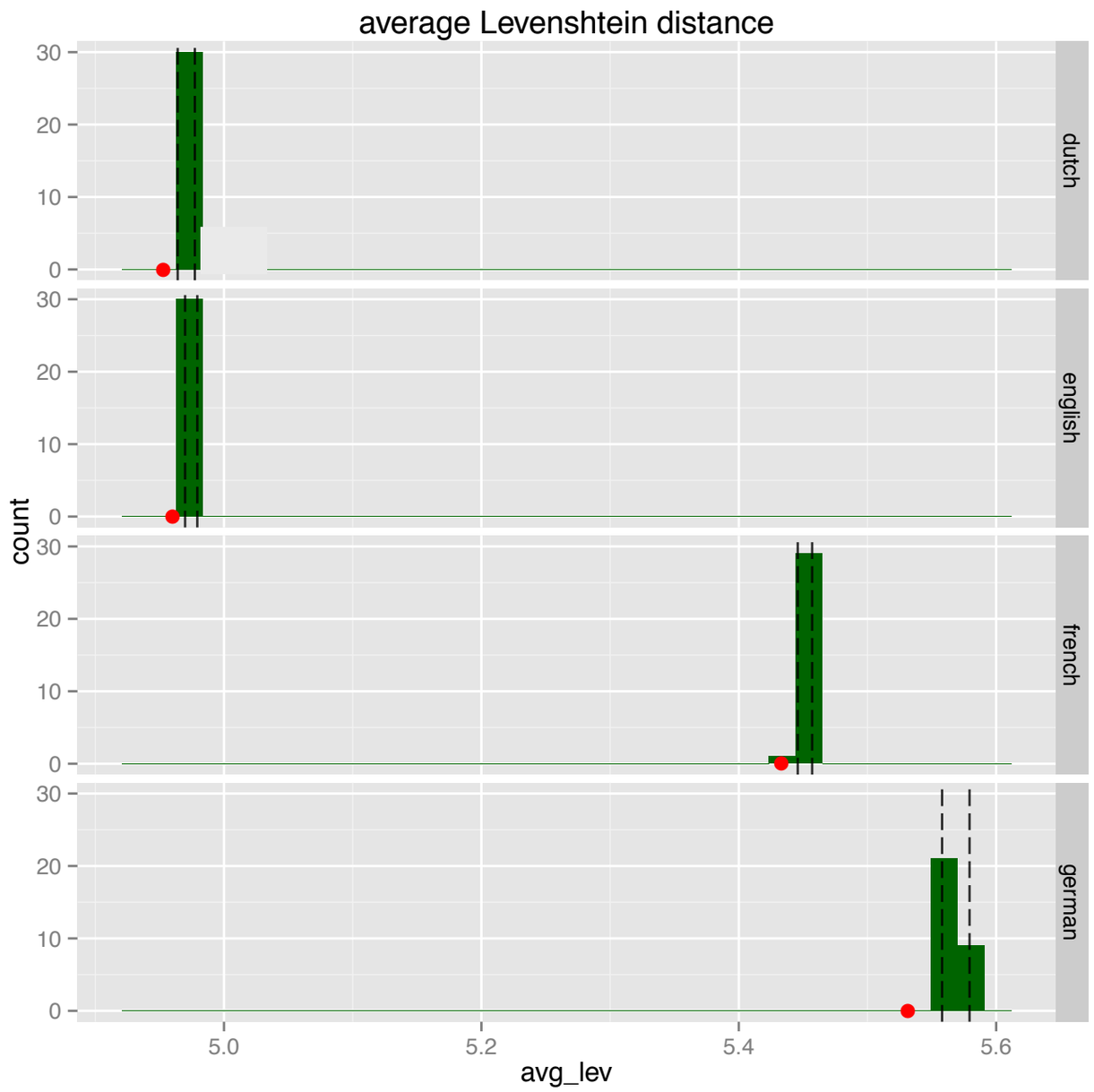


Results: Edit distance

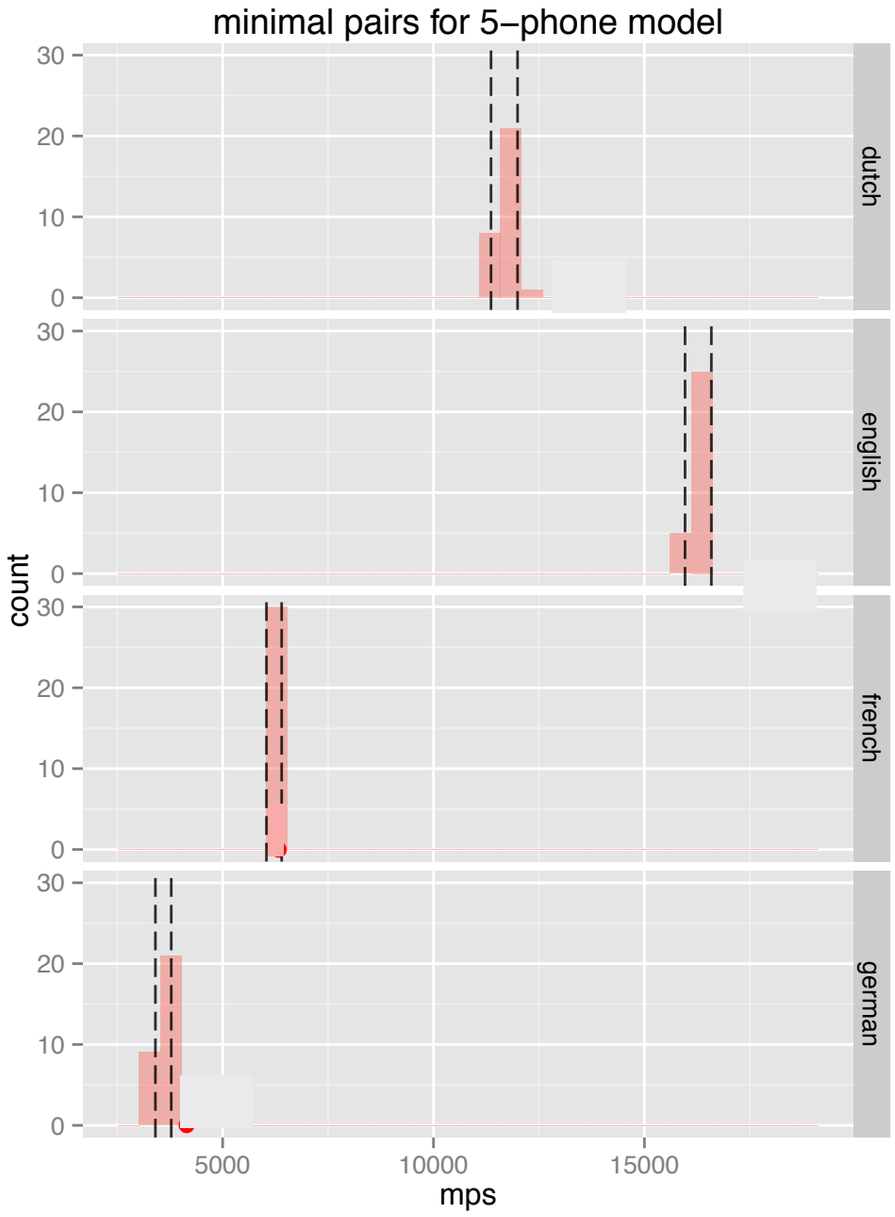


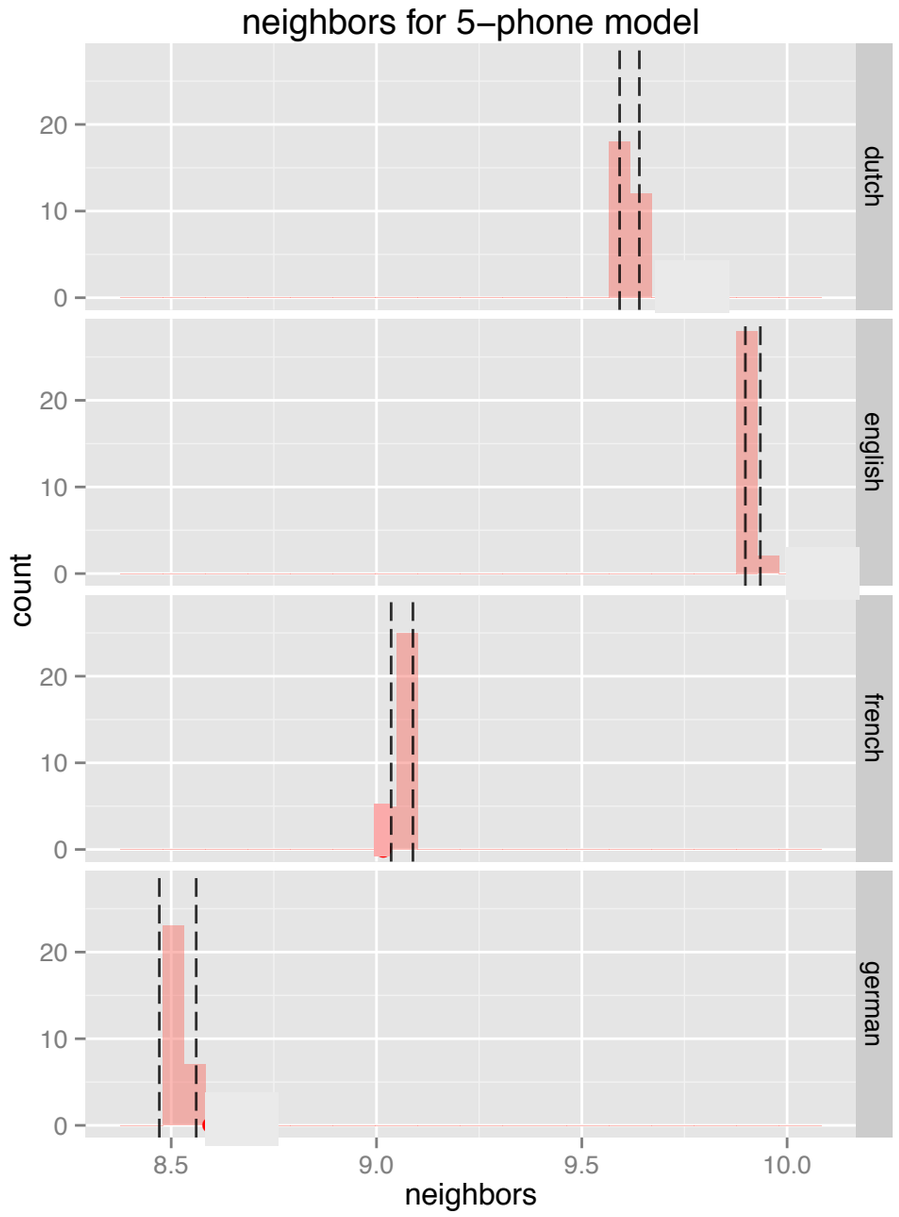
Results: Edit distance

clumpy



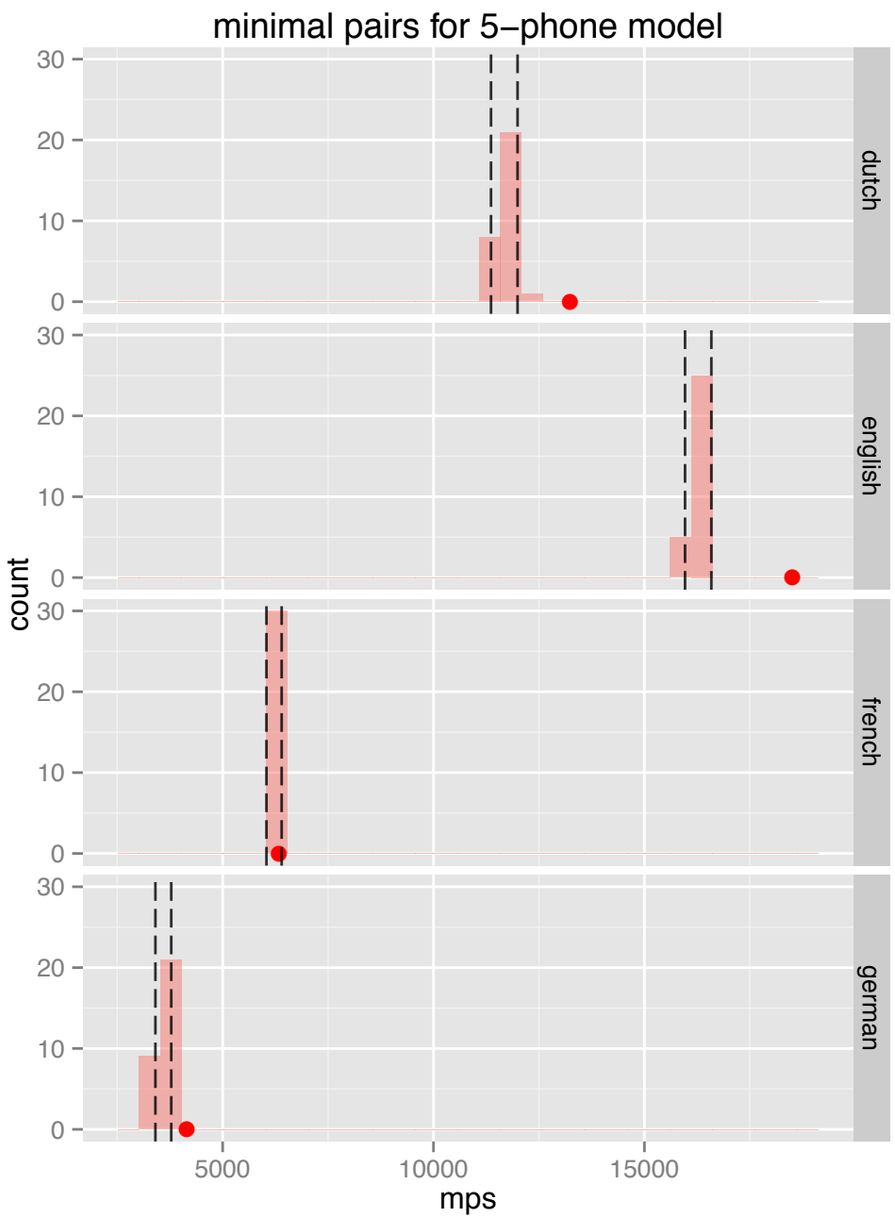
Results: Neighbors and minimal pairs

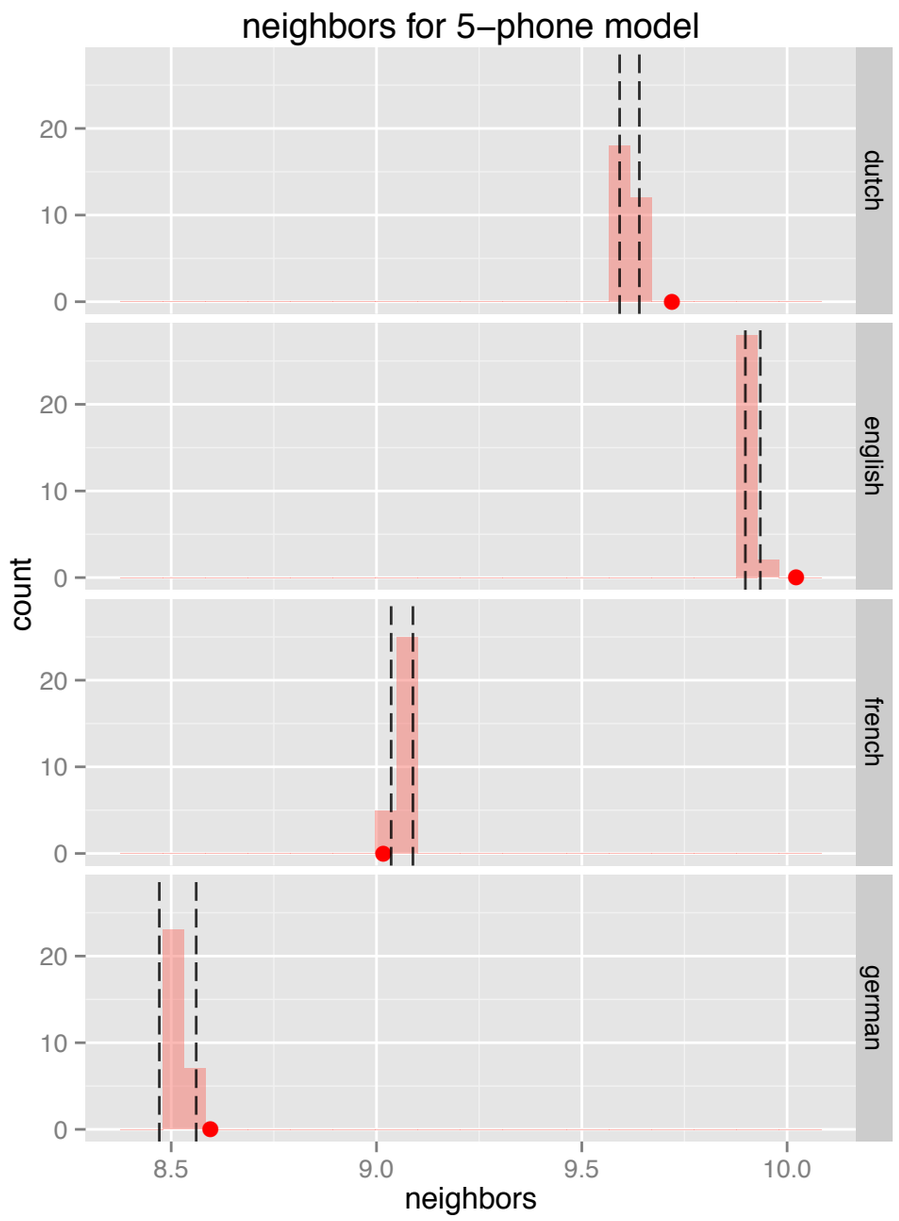




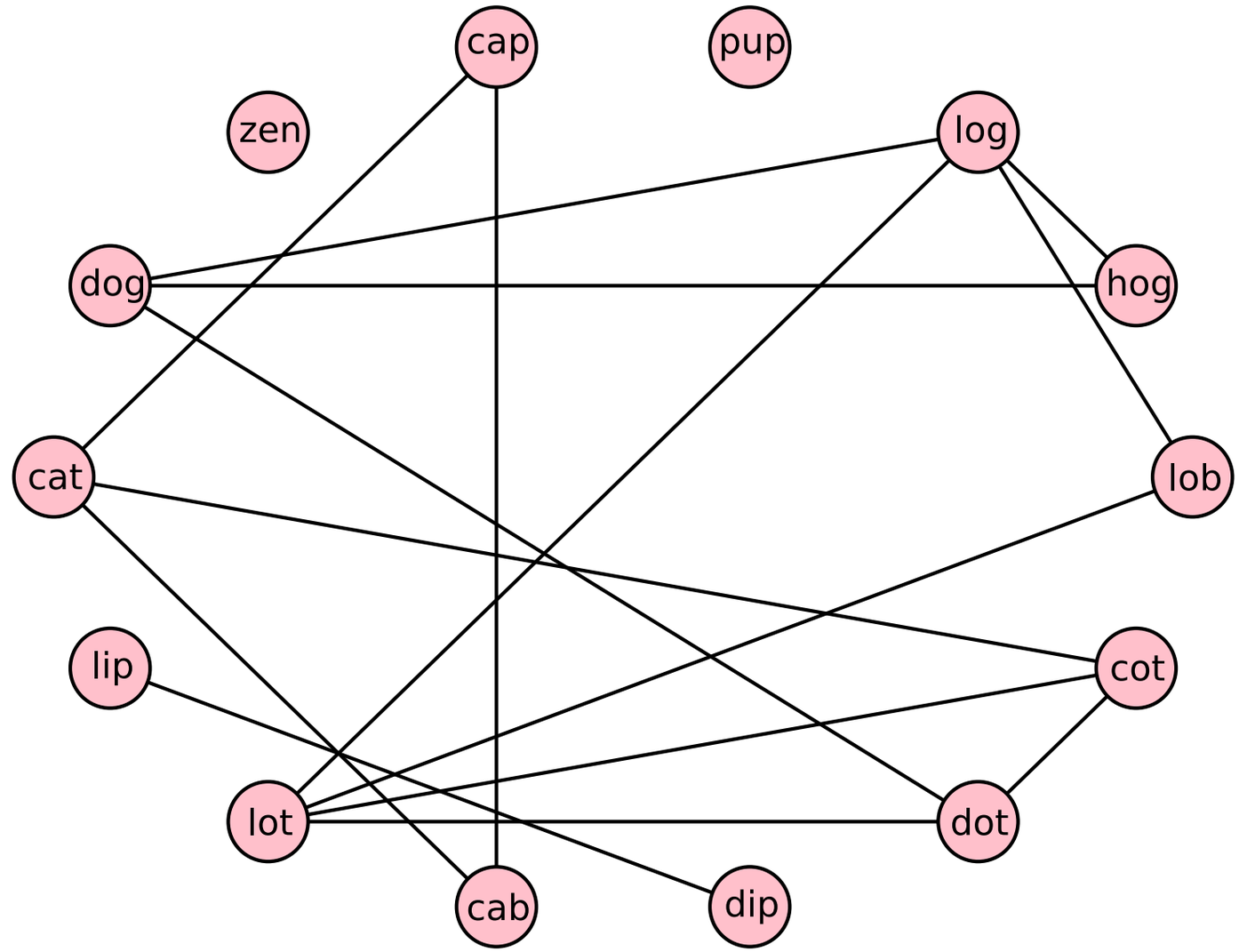
Results: Neighbors and minimal pairs

clumpy



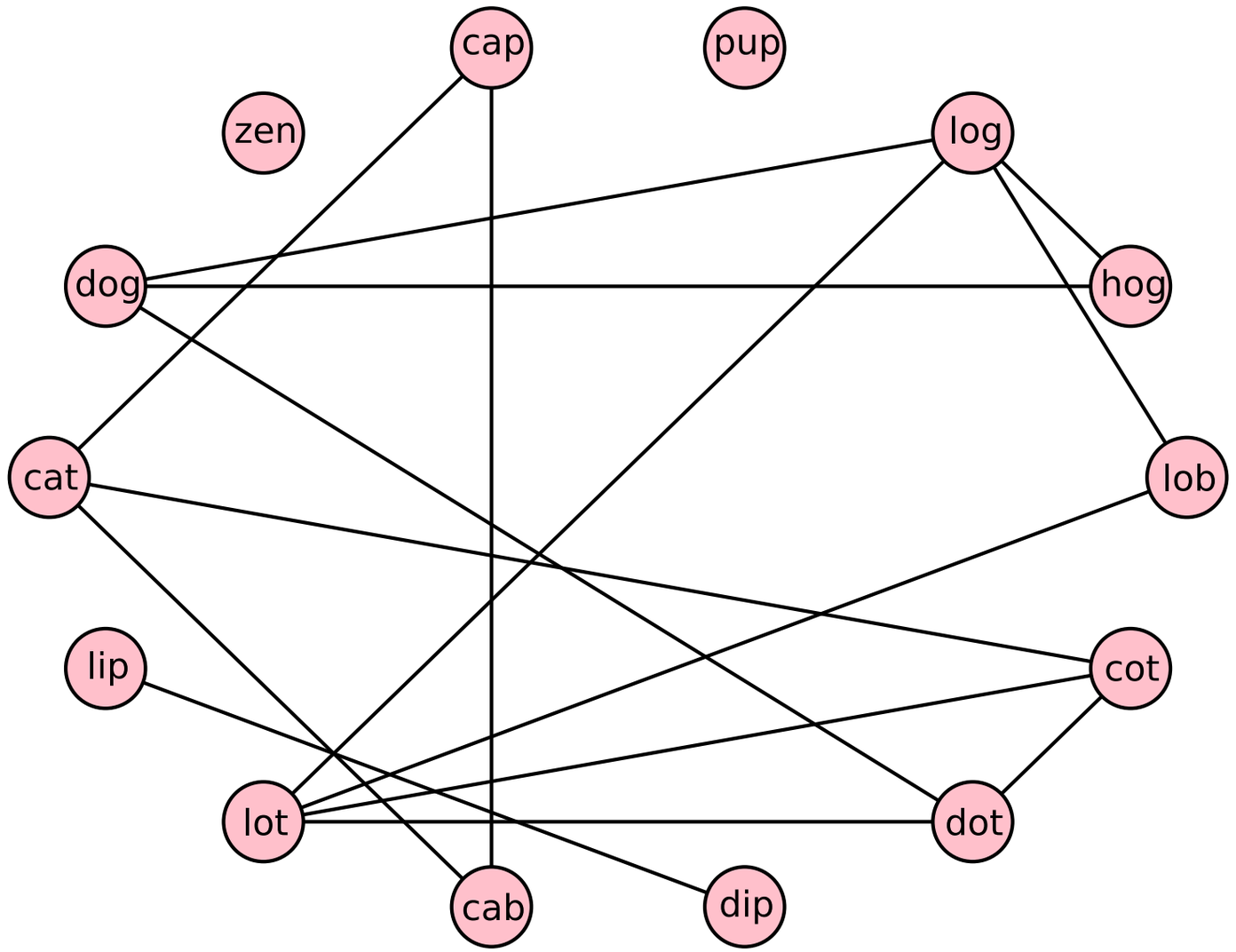


Results: Network measures

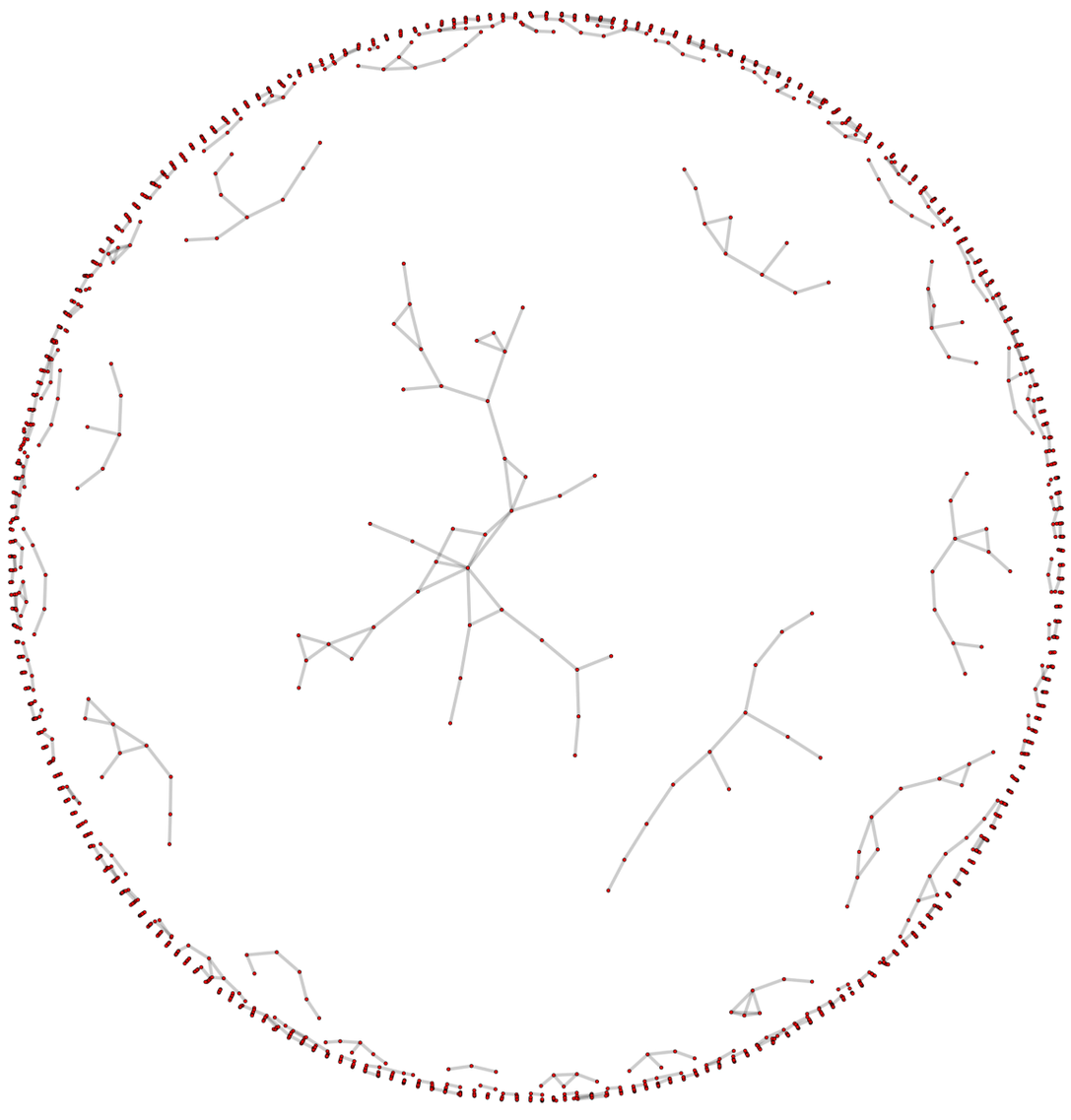


Results: Network measures

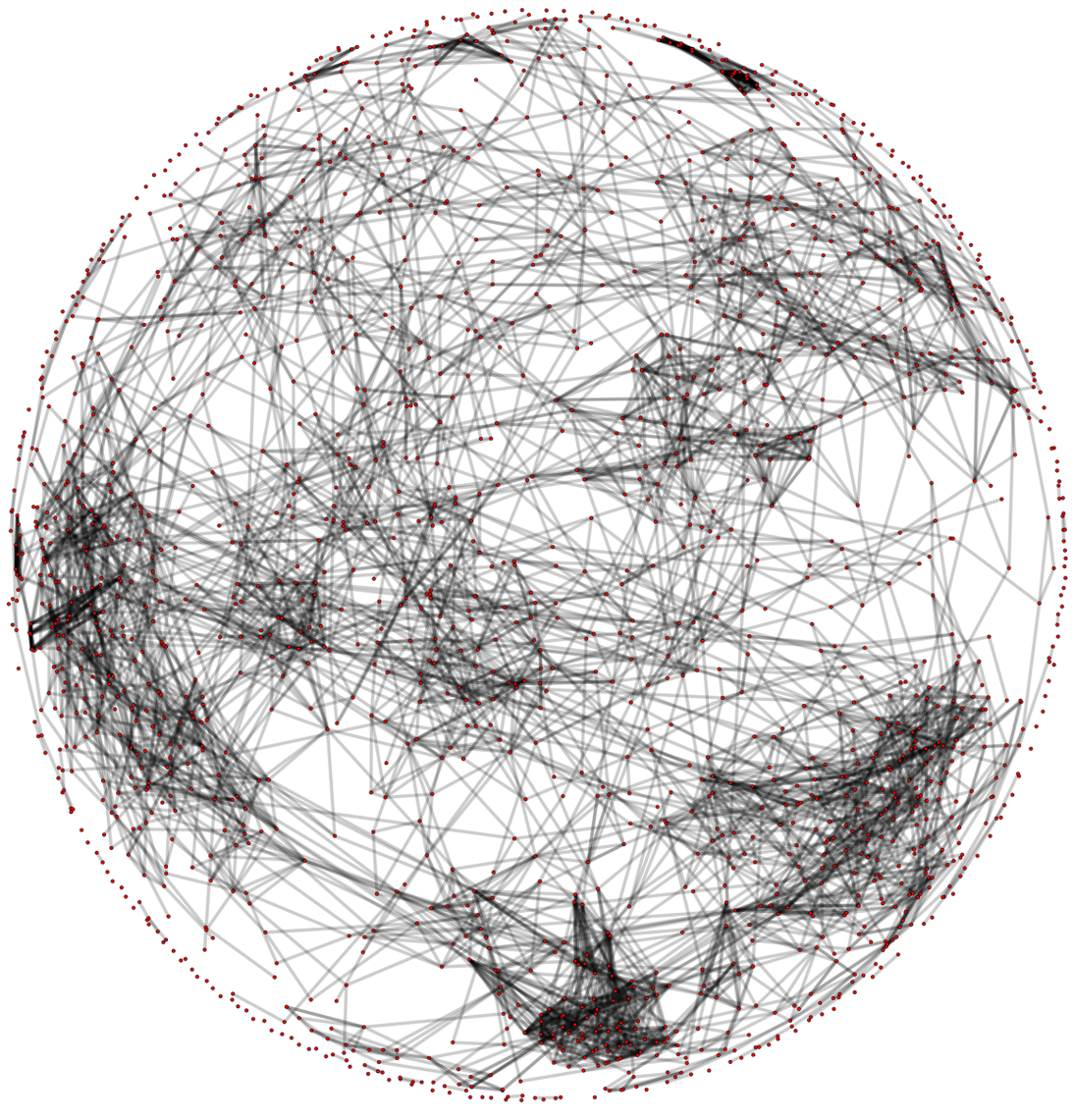
Connect any 2 nodes that are a minimal pair



1-phone model

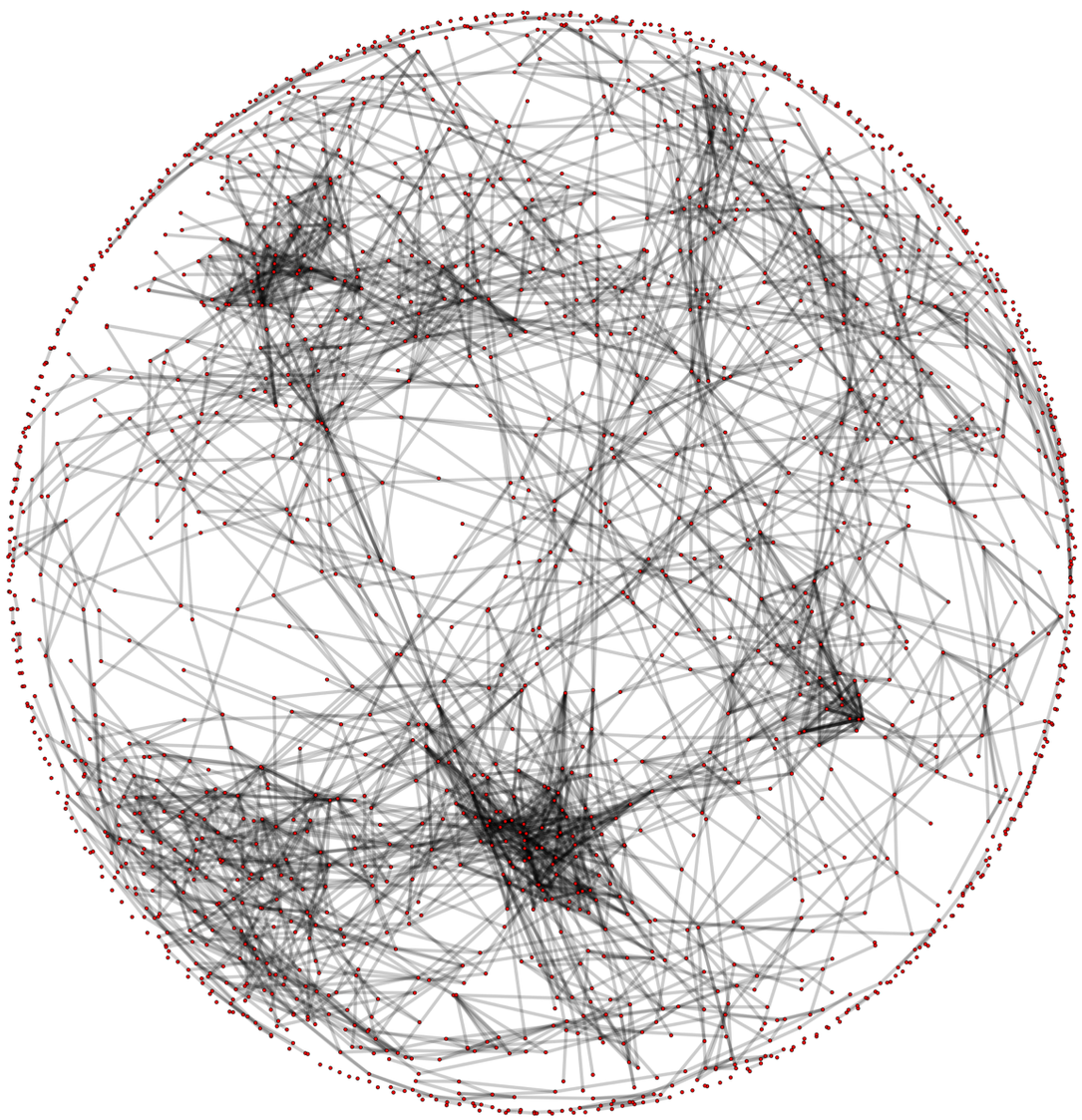


1-phone model

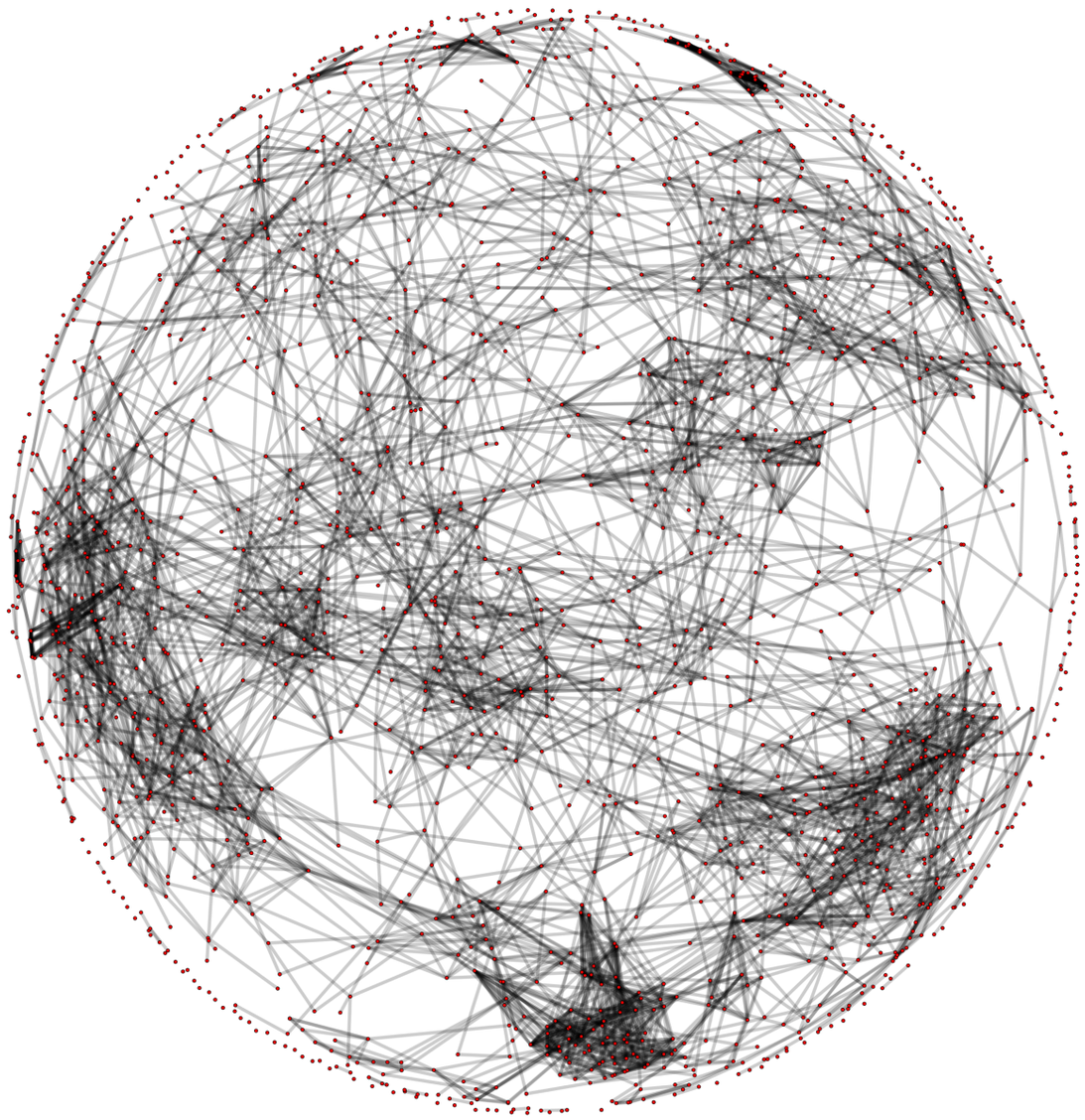


real lexicon

2-phone model

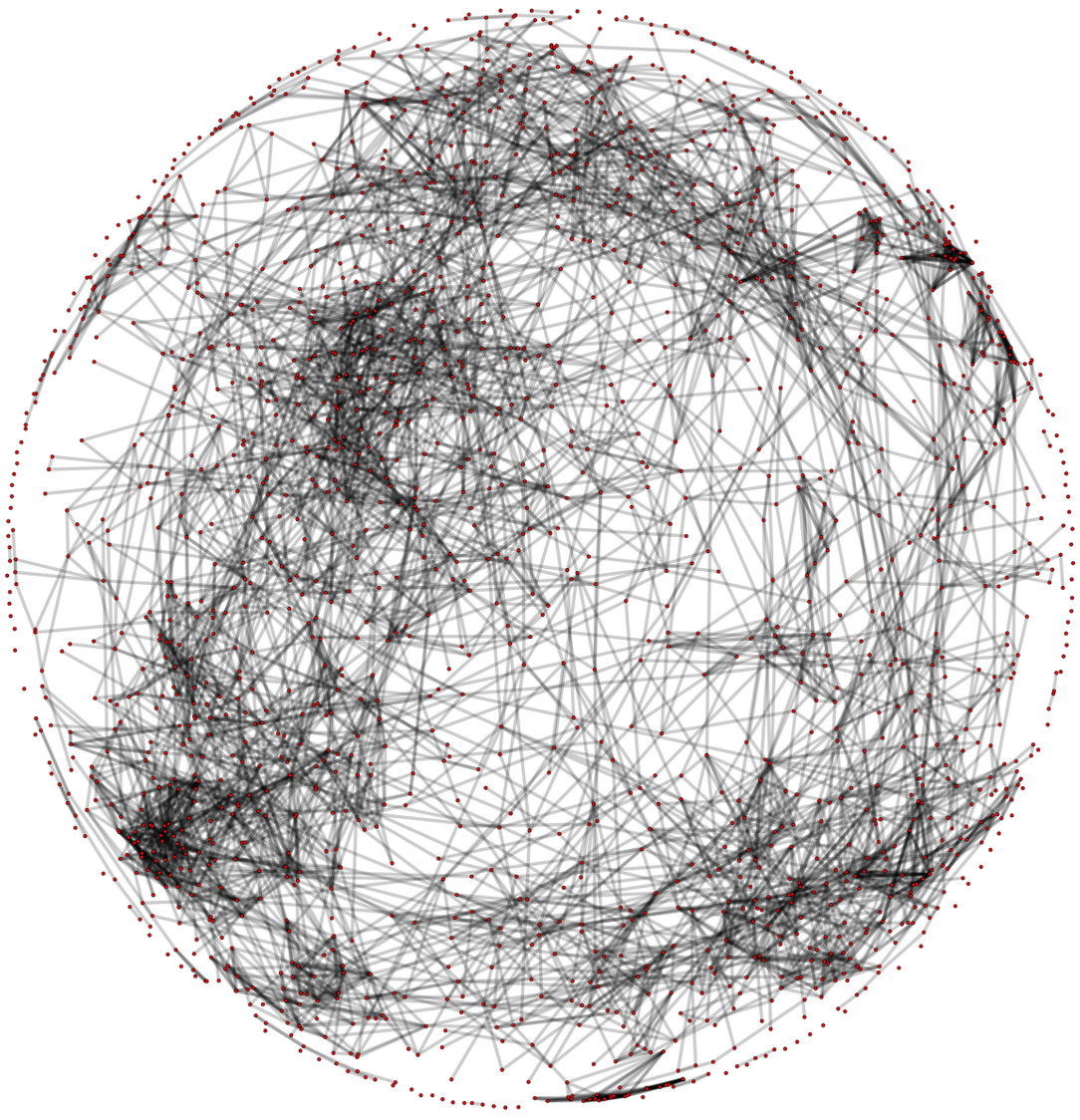


2-phone model

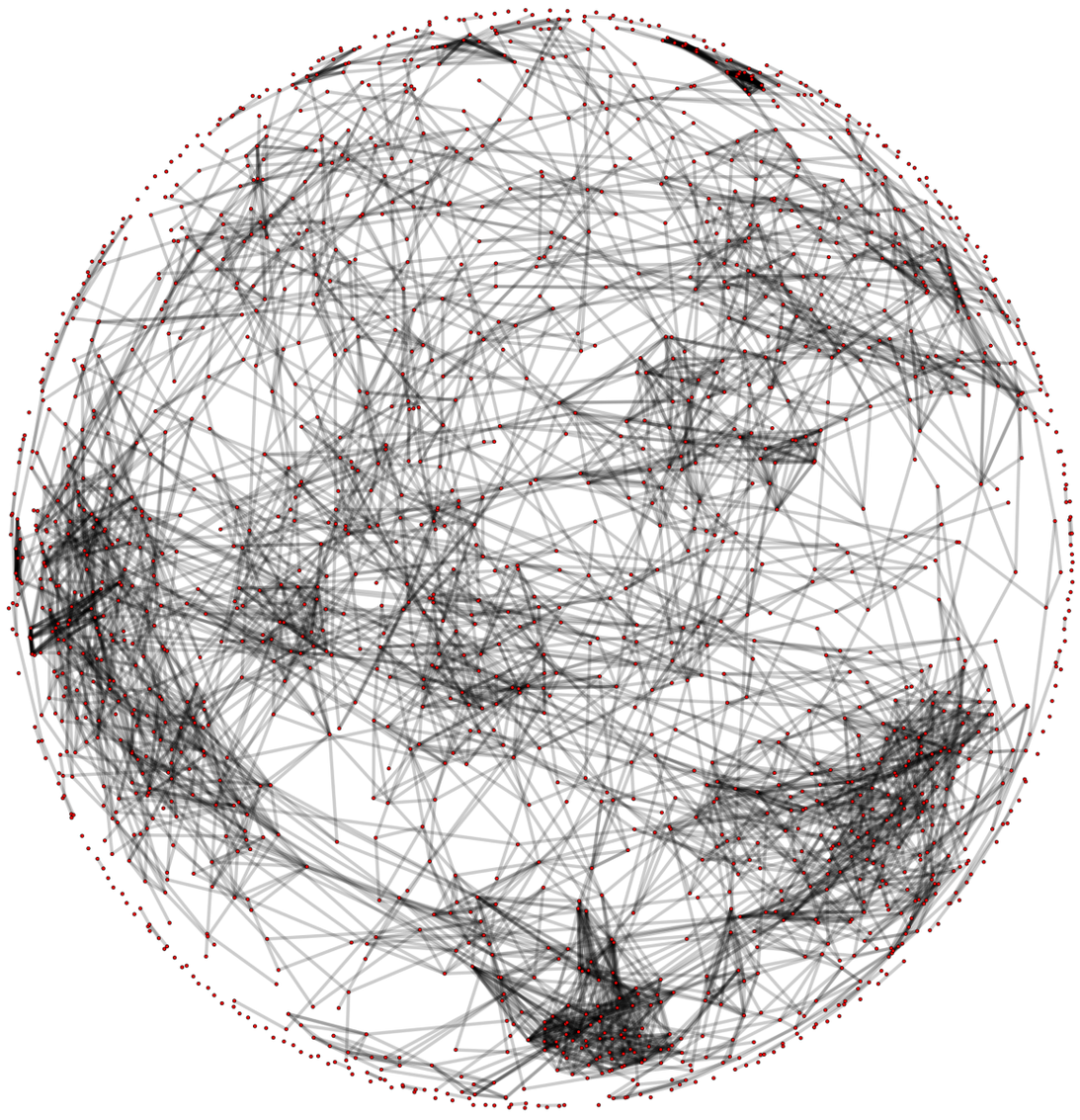


real lexicon

3-phone model

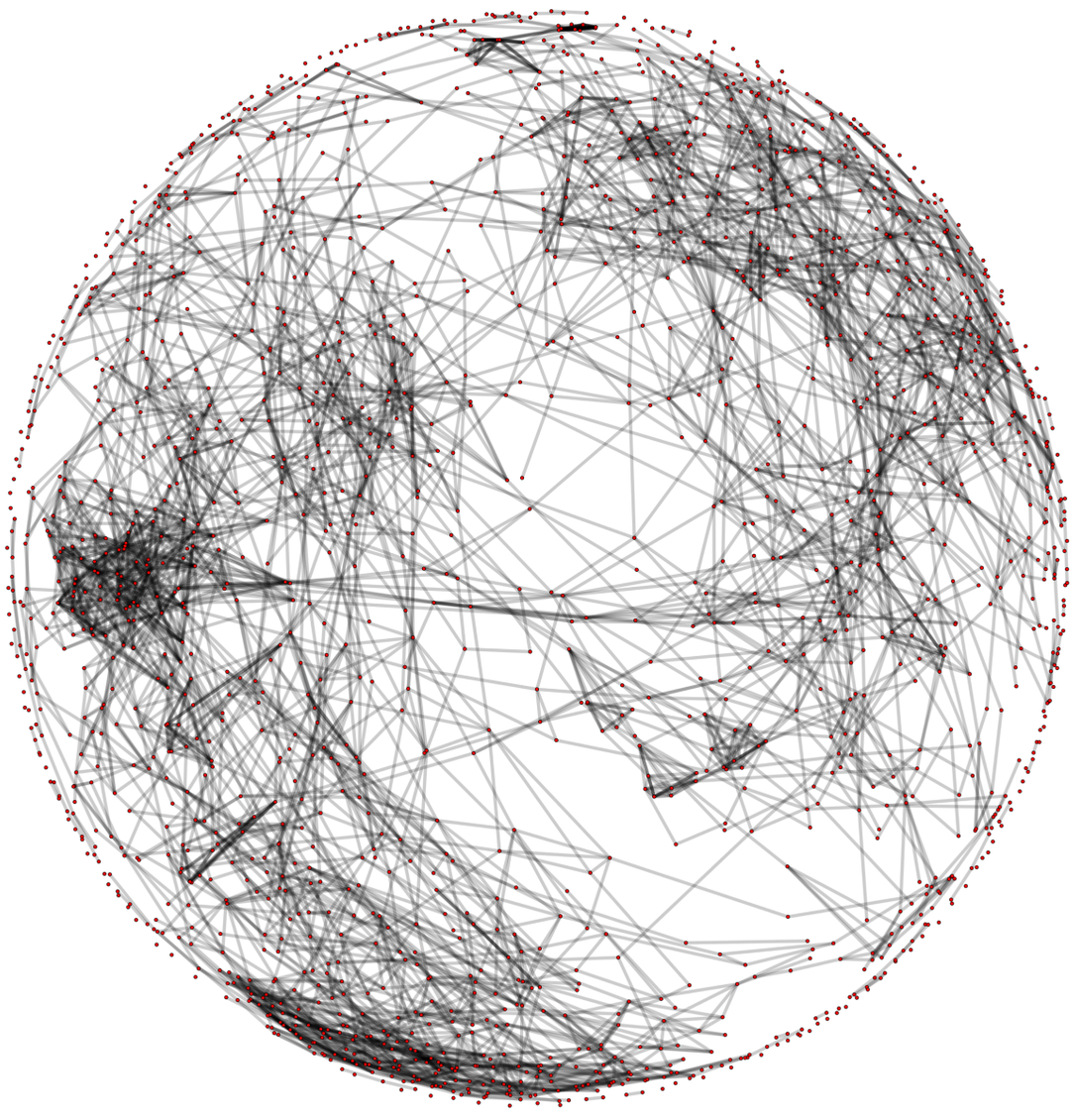


3-phone model

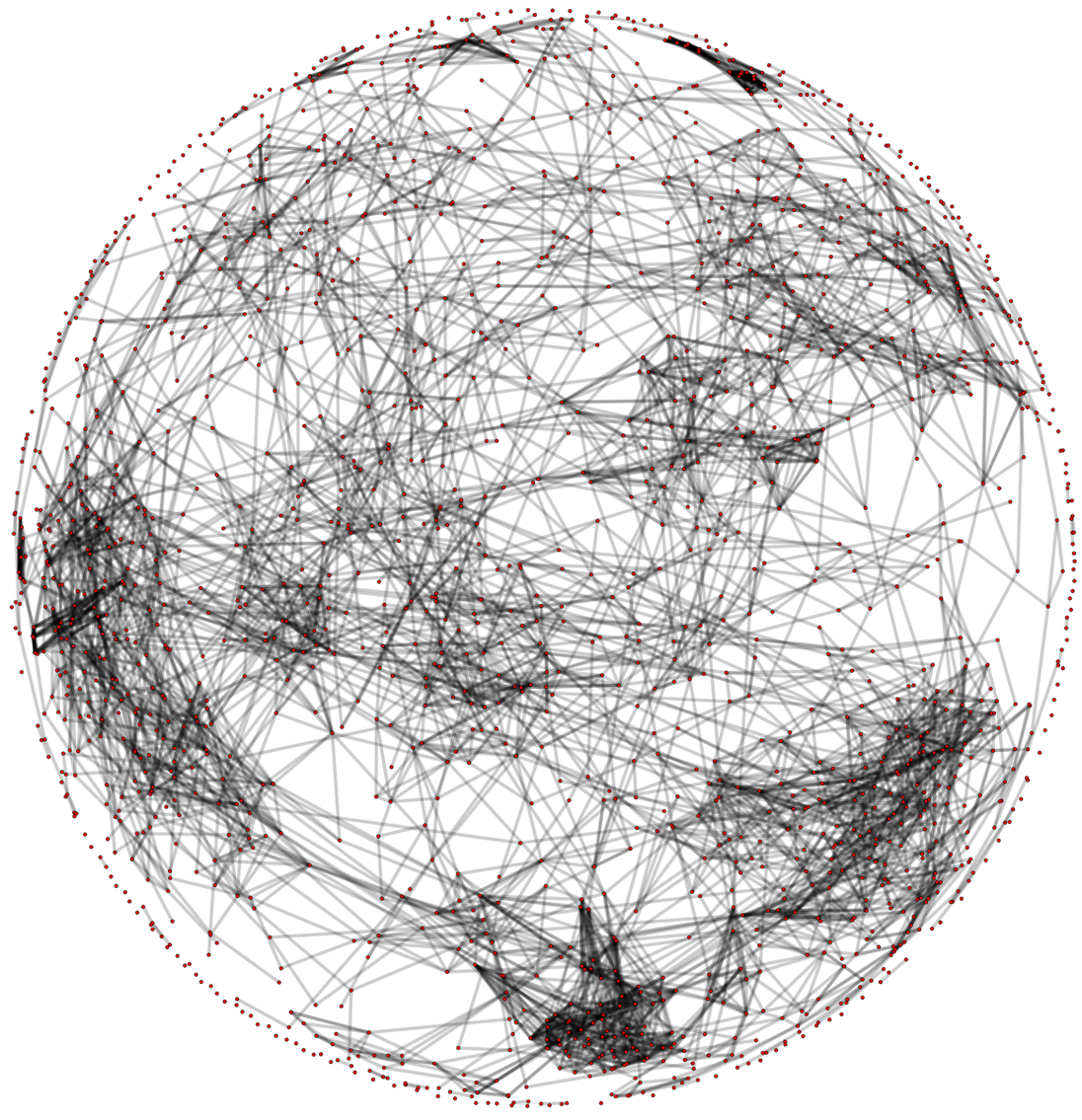


real lexicon

4-phone model

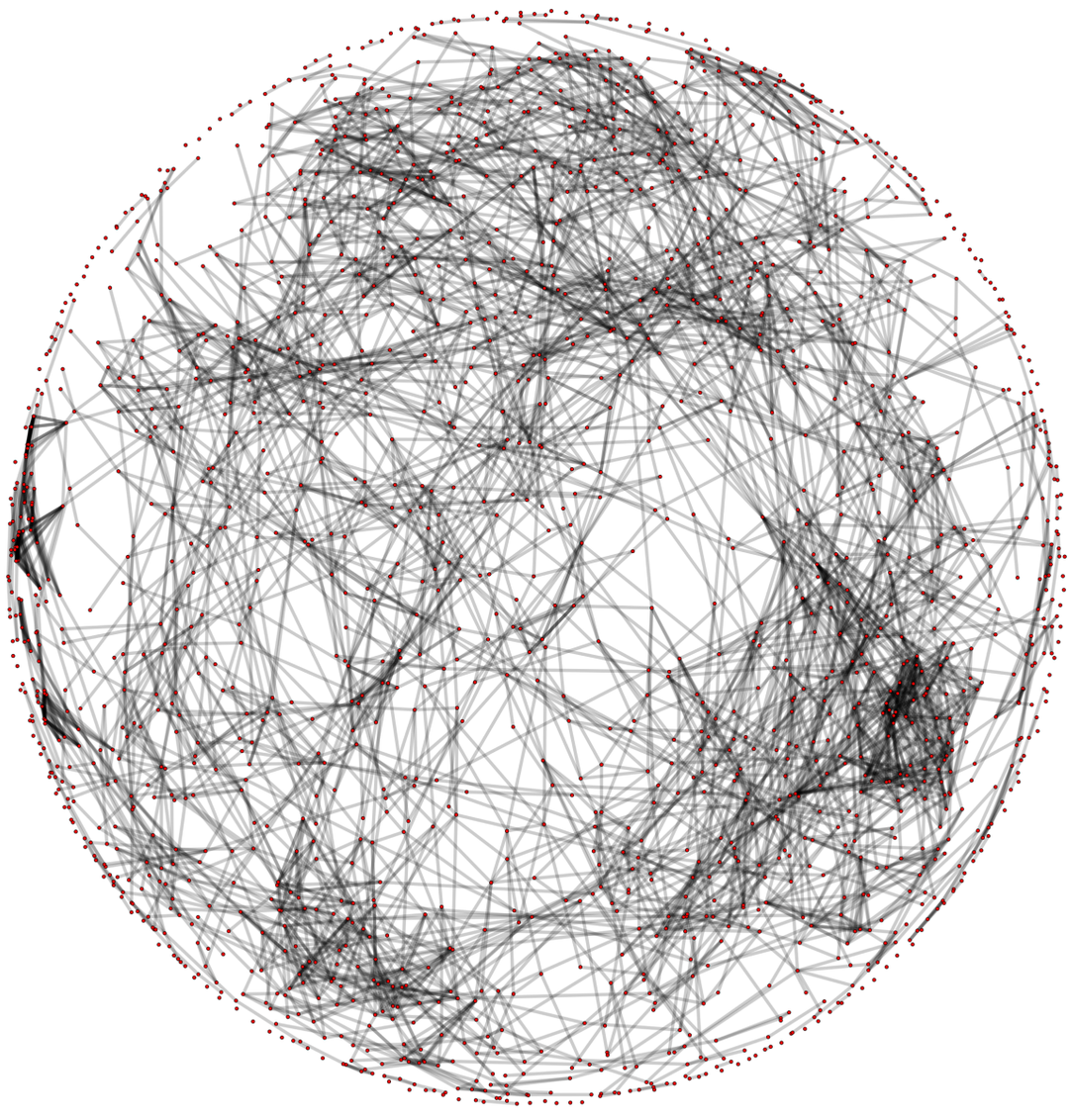


4-phone model

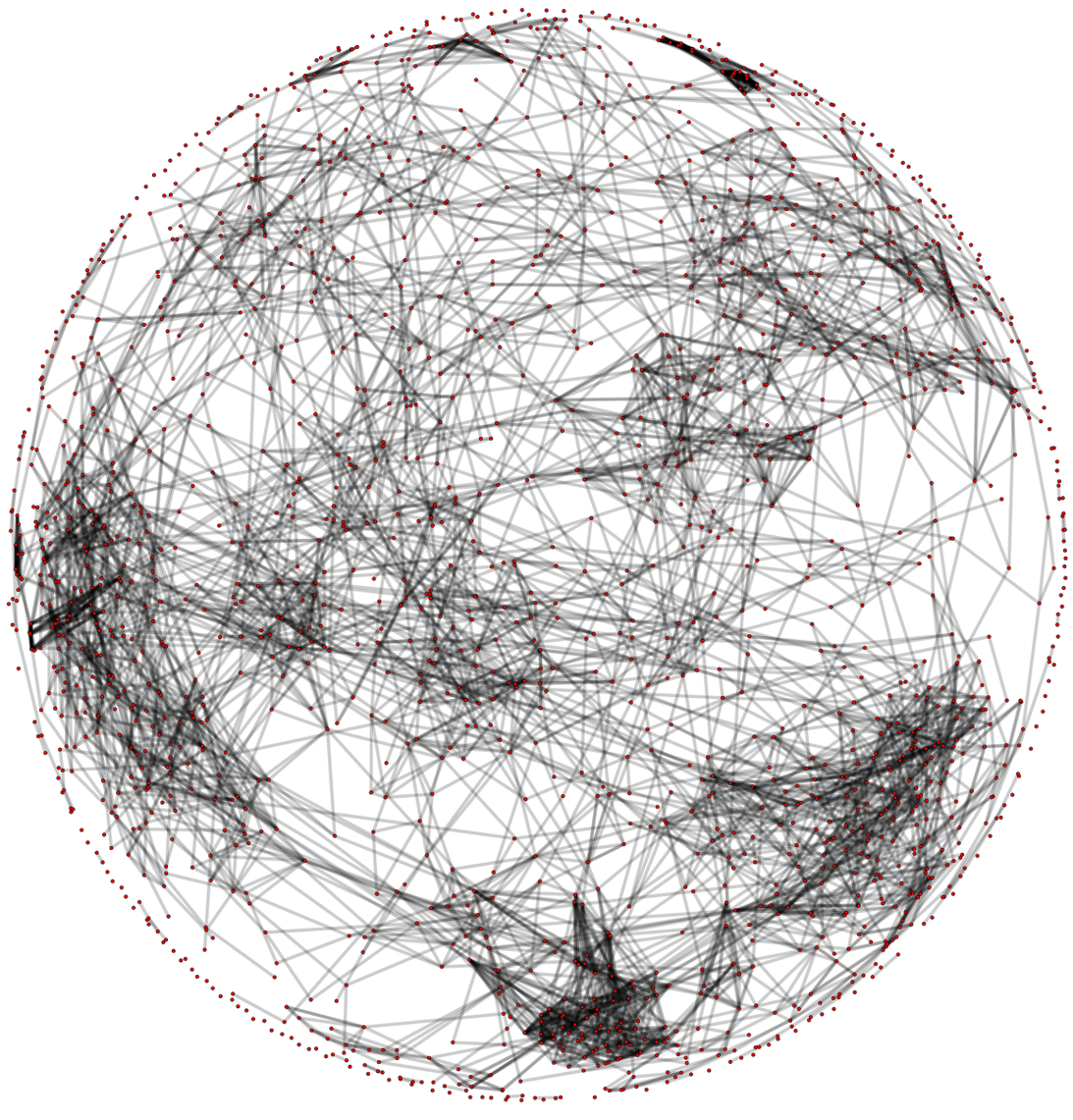


real lexicon

5-phone model

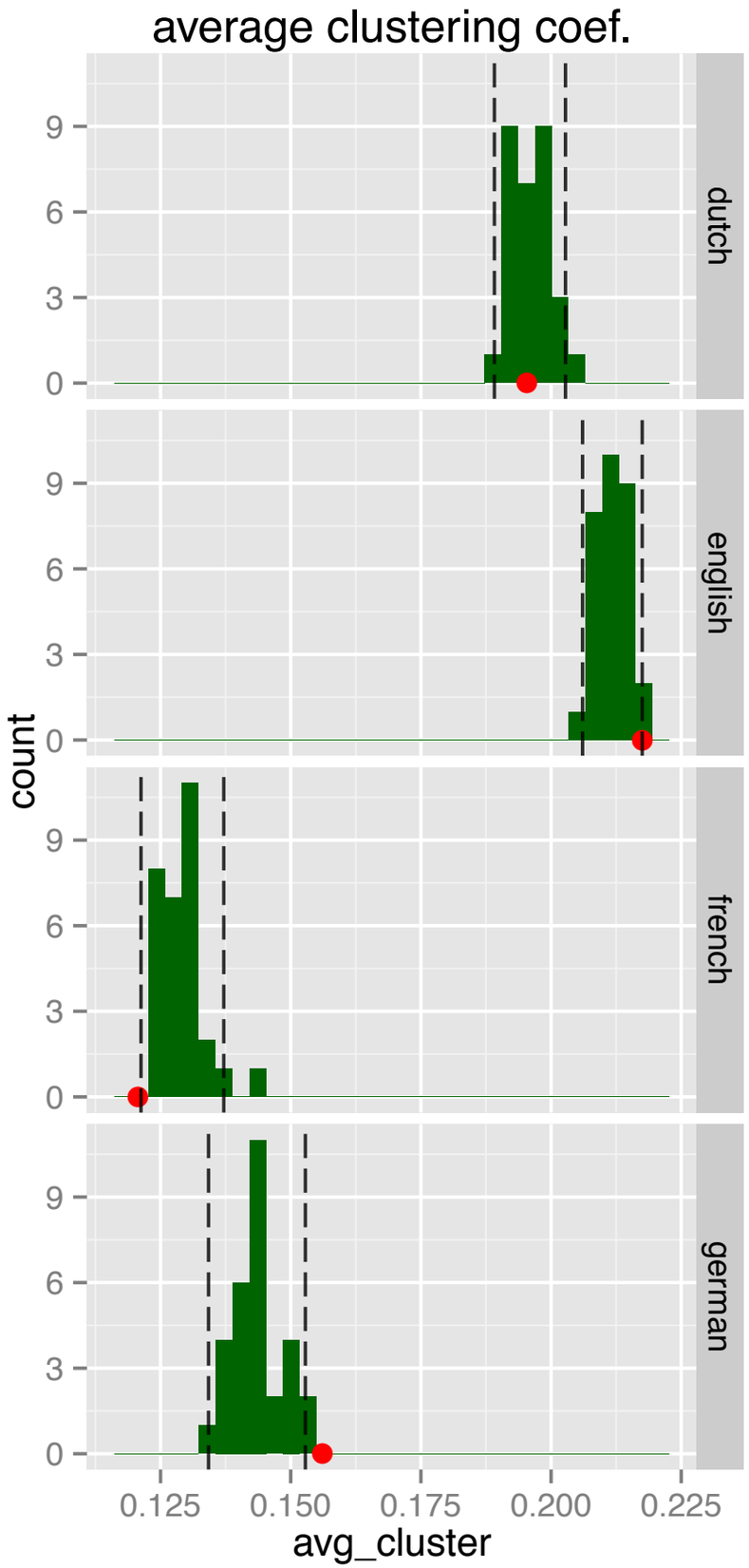


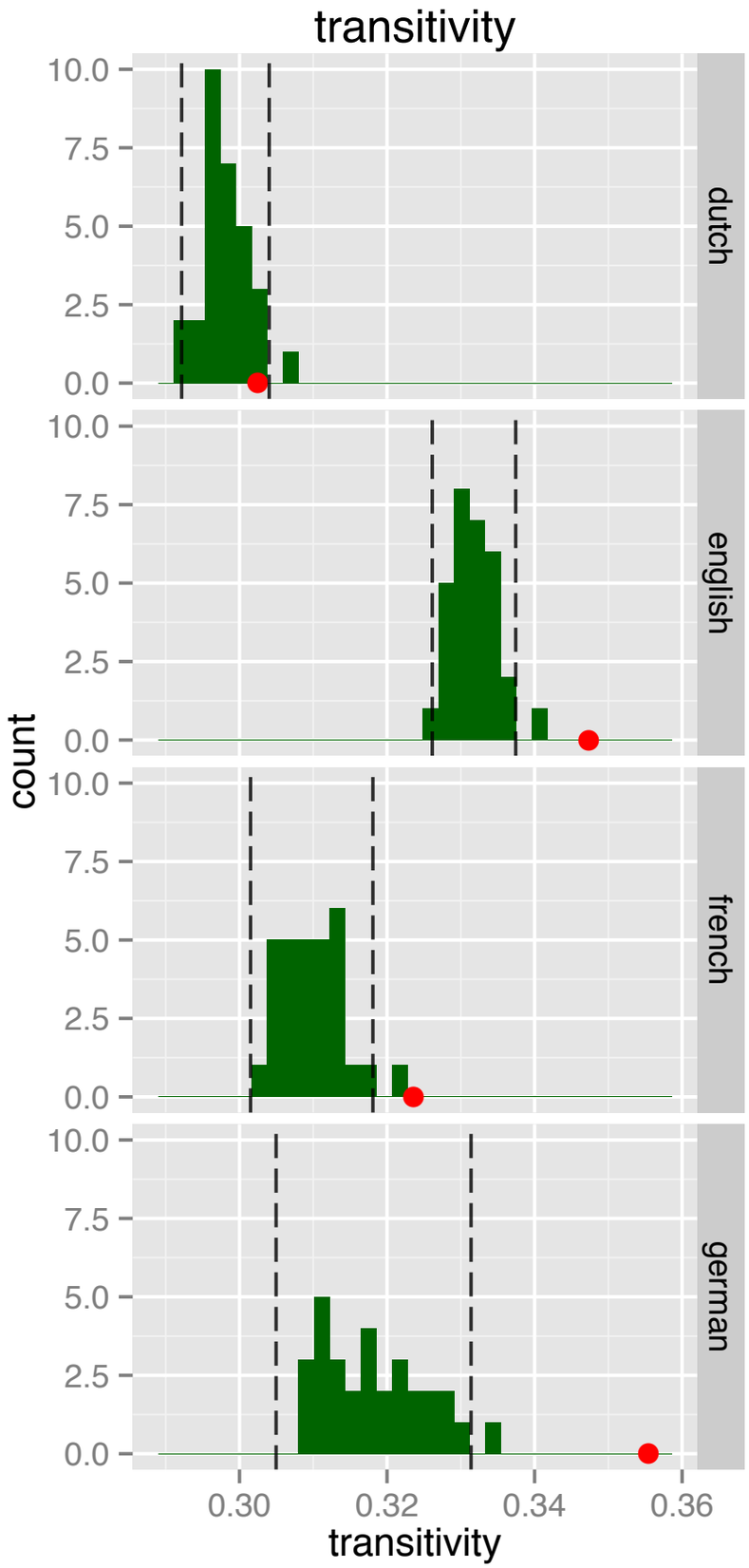
5-phone model



real lexicon

most network measures suggest clumpiness





Summary of null lexicons

- With the sometimes exception of French (and possibly word onset/coda measures), most real lexicon measures appear to be clumpy relative to the best baseline.

- Result not from morphology or meaningful subunits

- Not from sound to sound transition probability

- Ongoing work: effects of semantics? Effects of etymology.

Summary

- Info/information: Communicative pressures act on word length distributions.

- Wikipedia results: The lexicon is clumpy.

- Null lexicon results: Even compared to plausible baselines, the lexicon seems clumpy.

- Future work: Building models, exploring the underlying processes that give rise to the structure seen here.

Thanks!

- Collaborators on this work: Ted Gibson, Steve Piantadosi, Richard Futrell, Evelina Fedorenko, Isabelle Dautriche, Anne Christophe

- And members of Tedlab: Tim O’Donnell, Melissa Kline, Leon Bergen, Laura Stearns