Arbitrariness of Linguistic Sign Questioned: Correlation between Word Form and Meaning in Russian

Andrey Kutuzov
andreku@ifi.uio.no

University of Oslo

May 31, 2017

Introduction

Image from `https://seminalthought.blogspot.ru/`

Introduction

Image from `https://seminalthought.blogspot.ru/`

Prior knowledge

- Since Ferdinand de Saussure, we know that the linguistic sign is arbitrary:

Introduction

Image from `https://seminalthought.blogspot.ru/`

Prior knowledge

- Since Ferdinand de Saussure, we know that the linguistic sign is arbitrary:
- any meaning can be conveyed by any sequence of sounds or characters;

Introduction

Image from `https://seminalthought.blogspot.ru/`

Prior knowledge

- Since Ferdinand de Saussure, we know that the linguistic sign is arbitrary:
- any meaning can be conveyed by any sequence of sounds or characters;
- form and semantics are not related.

But...

But...

- There are exceptions from this law:

But...

- There are exceptions from this law:
- Onomatopoeia (imitating the sound with the word form);

Introduction

But...

- There are exceptions from this law:
- Onomatopoeia (imitating the sound with the word form);
  - ‘мяукать’

But...

- There are exceptions from this law:
- Onomatopoeia (imitating the sound with the word form);
  - ‘мяукать’
- Phonaesthemes (parts of words with consistently linked form and meaning):

But...

- There are exceptions from this law:
- Onomatopoeia (imitating the sound with the word form);
  - ‘мяукать’
- Phonaesthemes (parts of words with consistently linked form and meaning):
  - *‘gl-’* related to vision and light in English [Bergen, 2004];

But...

- There are exceptions from this law:
- Onomatopoeia (imitating the sound with the word form);
  - ‘мяукать’
- Phonaesthemes (parts of words with consistently linked form and meaning):
  - *‘gl-’* related to vision and light in English [Bergen, 2004];
  - ‘-стр-’ related to quickness or streaming in Russian [Mikhalev, 2008];
  - etc...

But...

- There are exceptions from this law:
- Onomatopoeia (imitating the sound with the word form);
  - 'мяукать'
- Phonaesthemes (parts of words with consistently linked form and meaning):
  - *'gl-'* related to vision and light in English [Bergen, 2004];
  - '-стр-' related to quickness or streaming in Russian [Mikhalev, 2008];
  - etc...

Can we quantify this systematic iconicity in the language as a whole?

Quantifying form and meaning

- ‘Surface’ and ‘semantic’ differences between word pairs;

Introduction

Quantifying form and meaning

- 'Surface' and 'semantic' differences between word pairs;
- if these differences are correlated, it would mean that the form to some extent does predict the meaning (or vice versa);

Quantifying form and meaning

- 'Surface' and 'semantic' differences between word pairs;
- if these differences are correlated, it would mean that the form to some extent does predict the meaning (or vice versa);
- the strength of this correlation shows how systematic is the vocabulary we deal with;

Quantifying form and meaning

- ‘Surface’ and ‘semantic’ differences between word pairs;
- if these differences are correlated, it would mean that the form to some extent does predict the meaning (or vice versa);
- the strength of this correlation shows how systematic is the vocabulary we deal with;
- surface differences: Levenshtein edit distances;

Quantifying form and meaning

- 'Surface' and 'semantic' differences between word pairs;
- if these differences are correlated, it would mean that the form to some extent does predict the meaning (or vice versa);
- the strength of this correlation shows how systematic is the vocabulary we deal with;
- surface differences: Levenshtein edit distances;
- semantic differences: cosine distances between word vectors in the word embedding models.

Quantifying form and meaning

- ‘Surface’ and ‘semantic’ differences between word pairs;
- if these differences are correlated, it would mean that the form to some extent does predict the meaning (or vice versa);
- the strength of this correlation shows how systematic is the vocabulary we deal with;
- surface differences: Levenshtein edit distances;
- semantic differences: cosine distances between word vectors in the word embedding models.

Findings for Russian

- We analyzed the link between the graphic forms and meanings of frequent monosyllabic Russian nouns;

Introduction

Quantifying form and meaning

- 'Surface' and 'semantic' differences between word pairs;
- if these differences are correlated, it would mean that the form to some extent does predict the meaning (or vice versa);
- the strength of this correlation shows how systematic is the vocabulary we deal with;
- surface differences: Levenshtein edit distances;
- semantic differences: cosine distances between word vectors in the word embedding models.

Findings for Russian

- We analyzed the link between the graphic forms and meanings of frequent monosyllabic Russian nouns;
- There is a strongly statistically significant systematicity in this data;

Introduction

Quantifying form and meaning

- ‘Surface’ and ‘semantic’ differences between word pairs;
- if these differences are correlated, it would mean that the form to some extent does predict the meaning (or vice versa);
- the strength of this correlation shows how systematic is the vocabulary we deal with;
- surface differences: Levenshtein edit distances;
- semantic differences: cosine distances between word vectors in the word embedding models.

Findings for Russian

- We analyzed the link between the graphic forms and meanings of frequent monosyllabic Russian nouns;
- There is a strongly statistically significant systematicity in this data;
- The correlation is even higher than the one reported in similar experiments for English.

Some previous work

- The form space and meaning in English were shown to be related in [Monaghan et al., 2014];

Some previous work

- The form space and meaning in English were shown to be related in [Monaghan et al., 2014];
- indeed, there are regions in the lexicon, where the arbitrariness principle is violated;

Some previous work

- The form space and meaning in English were shown to be related in [Monaghan et al., 2014];
- indeed, there are regions in the lexicon, where the arbitrariness principle is violated;
- [Gutierrez et al., 2016] further proved this with modern word embedding models and kernel regresssion (best paper award at ACL-2016);

Some previous work

- The form space and meaning in English were shown to be related in [Monaghan et al., 2014];
- indeed, there are regions in the lexicon, where the arbitrariness principle is violated;
- [Gutierrez et al., 2016] further proved this with modern word embedding models and kernel regresssion (best paper award at ACL-2016);
- [Blasi et al., 2016] showed that there are strong cross-linguistic sound-meaning associations.

Introduction

Some previous work

- The form space and meaning in English were shown to be related in [Monaghan et al., 2014];
- indeed, there are regions in the lexicon, where the arbitrariness principle is violated;
- [Gutierrez et al., 2016] further proved this with modern word embedding models and kernel regresssion (best paper award at ACL-2016);
- [Blasi et al., 2016] showed that there are strong cross-linguistic sound-meaning associations.

What about Russian?

- The problem was studied in [Zhuravlev, 1991] and other works of the same author;

Introduction

Some previous work

- The form space and meaning in English were shown to be related in [Monaghan et al., 2014];
- indeed, there are regions in the lexicon, where the arbitrariness principle is violated;
- [Gutierrez et al., 2016] further proved this with modern word embedding models and kernel regresssion (best paper award at ACL-2016);
- [Blasi et al., 2016] showed that there are strong cross-linguistic sound-meaning associations.

What about Russian?

- The problem was studied in [Zhuravlev, 1991] and other works of the same author;
- the results were unstable, hardly verifiable and generally disputable.

Some previous work

- The form space and meaning in English were shown to be related in [Monaghan et al., 2014];
- indeed, there are regions in the lexicon, where the arbitrariness principle is violated;
- [Gutierrez et al., 2016] further proved this with modern word embedding models and kernel regresssion (best paper award at ACL-2016);
- [Blasi et al., 2016] showed that there are strong cross-linguistic sound-meaning associations.

What about Russian?

- The problem was studied in [Zhuravlev, 1991] and other works of the same author;
- the results were unstable, hardly verifiable and generally disputable.

Now we can quantify it properly.

Experimental setting

Data sources

4 test sets were produced from the Russian National Corpus (**RNC**):

Experimental setting

Data sources

4 test sets were produced from the Russian National Corpus (**RNC**):

1. **Mono**: all monosyllabic nouns with frequency > 100 (1 729 words);

Experimental setting

Data sources

4 test sets were produced from the Russian National Corpus (**RNC**):

1. **Mono**: all monosyllabic nouns with frequency > 100 (1 729 words);
2. **Bi**: monosyllabic and bisyllabic words with frequency > 1000 (2 900 words);

Experimental setting

Data sources

4 test sets were produced from the Russian National Corpus (**RNC**):

1. **Mono**: all monosyllabic nouns with frequency > 100 (1 729 words);
2. **Bi**: monosyllabic and bisyllabic words with frequency > 1000 (2 900 words);
3. **Bi_NoDim**: the same as **Bi**, w/o the nouns ending with the diminutive suffixes '-ок', '-ек' and '-ка; (2 633 words);

Experimental setting

Data sources

4 test sets were produced from the Russian National Corpus (**RNC**):

1. **Mono**: all monosyllabic nouns with frequency > 100 (1 729 words);
2. **Bi**: monosyllabic and bisyllabic words with frequency > 1000 (2 900 words);
3. **Bi_NoDim**: the same as **Bi**, w/o the nouns ending with the diminutive suffixes '-ок', '-ек' and '-ка; (2 633 words);
4. **All**: all nouns with frequency > 1000 (6 715 words).

Experimental setting

Data sources

4 test sets were produced from the Russian National Corpus (**RNC**):

1. **Mono**: all monosyllabic nouns with frequency > 100 (1 729 words);
2. **Bi**: monosyllabic and bisyllabic words with frequency > 1000 (2 900 words);
3. **Bi_NoDim**: the same as **Bi**, w/o the nouns ending with the diminutive suffixes '-ок', '-ек' and '-ка; (2 633 words);
4. **All**: all nouns with frequency > 1000 (6 715 words).

Excluded:

Experimental setting

Data sources

4 test sets were produced from the Russian National Corpus (**RNC**):

1. **Mono**: all monosyllabic nouns with frequency > 100 (1 729 words);
2. **Bi**: monosyllabic and bisyllabic words with frequency > 1000 (2 900 words);
3. **Bi_NoDim**: the same as **Bi**, w/o the nouns ending with the diminutive suffixes '-ок', '-ек' and '-ка; (2 633 words);
4. **All**: all nouns with frequency > 1000 (6 715 words).

Excluded:

- nouns less than 3 characters;

Experimental setting

Data sources

4 test sets were produced from the Russian National Corpus (**RNC**):

1. **Mono**: all monosyllabic nouns with frequency > 100 (1 729 words);
2. **Bi**: monosyllabic and bisyllabic words with frequency > 1000 (2 900 words);
3. **Bi_NoDim**: the same as **Bi**, w/o the nouns ending with the diminutive suffixes '-ок', '-ек' and '-ка; (2 633 words);
4. **All**: all nouns with frequency > 1000 (6 715 words).

Excluded:

- nouns less than 3 characters;
- nouns with non-Cyrillic characters and digits;

Experimental setting

Data sources

4 test sets were produced from the Russian National Corpus (**RNC**):

1. **Mono**: all monosyllabic nouns with frequency > 100 (1 729 words);
2. **Bi**: monosyllabic and bisyllabic words with frequency > 1000 (2 900 words);
3. **Bi_NoDim**: the same as **Bi**, w/o the nouns ending with the diminutive suffixes '-ок', '-ек' and '-ка; (2 633 words);
4. **All**: all nouns with frequency > 1000 (6 715 words).

Excluded:

- nouns less than 3 characters;
- nouns with non-Cyrillic characters and digits;
- proper names and toponyms (as detected by *Mystem*).

Experimental setting

Data sources

4 test sets were produced from the Russian National Corpus (**RNC**):

1. **Mono**: all monosyllabic nouns with frequency > 100 (1 729 words);
2. **Bi**: monosyllabic and bisyllabic words with frequency > 1000 (2 900 words);
3. **Bi_NoDim**: the same as **Bi**, w/o the nouns ending with the diminutive suffixes '-ок', '-ек' and '-ка; (2 633 words);
4. **All**: all nouns with frequency > 1000 (6 715 words).

Excluded:

- nouns less than 3 characters;
- nouns with non-Cyrillic characters and digits;
- proper names and toponyms (as detected by *Mystem*).

Experimental setting

Distributional model

- For orthographic differences, the edit distance is enough;

Experimental setting

Distributional model

- For orthographic differences, the edit distance is enough;
- for semantic differences, we need a distributional semantic model.

Experimental setting

Distributional model

- For orthographic differences, the edit distance is enough;
- for semantic differences, we need a distributional semantic model.

Continuous Skipgram model [Mikolov et al., 2013] was trained on the lemmatized and PoS-tagged RNC:

Experimental setting

Distributional model

- For orthographic differences, the edit distance is enough;
- for semantic differences, we need a distributional semantic model.

Continuous Skipgram model [Mikolov et al., 2013] was trained on the lemmatized and PoS-tagged RNC:

- vector size 300;

Experimental setting

Distributional model

- For orthographic differences, the edit distance is enough;
- for semantic differences, we need a distributional semantic model.

Continuous Skipgram model [Mikolov et al., 2013] was trained on the lemmatized and PoS-tagged RNC:

- vector size 300;
- symmetric context window 10;

Experimental setting

Distributional model

- For orthographic differences, the edit distance is enough;
- for semantic differences, we need a distributional semantic model.

Continuous Skipgram model [Mikolov et al., 2013] was trained on the lemmatized and PoS-tagged RNC:

- vector size 300;
- symmetric context window 10;
- other hyperparameters set as default.

Experimental setting

Intrinsic evaluation of the model:

- Russian part of *Multilingual SimLex999* [Leviant and Reichart, 2015]: 0.36;

Experimental setting

Intrinsic evaluation of the model:

- Russian part of *Multilingual SimLex999* [Leviant and Reichart, 2015]: 0.36;
- Russian translation of *Google Analogies* dataset [Mikolov et al., 2013]: 0.65.

Experimental setting

Intrinsic evaluation of the model:

- Russian part of *Multilingual SimLex999* [Leviant and Reichart, 2015]: 0.36;
- Russian translation of *Google Analogies* dataset [Mikolov et al., 2013]: 0.65.

These results are comparable to state-of-the-art for English and Russian.

Experimental setting

Intrinsic evaluation of the model:

- Russian part of *Multilingual SimLex999* [Leviant and Reichart, 2015]: 0.36;
- Russian translation of *Google Analogies* dataset [Mikolov et al., 2013]: 0.65.

These results are comparable to state-of-the-art for English and Russian.

Thus, the model is good enough to build further experiments upon it.

Measuring correlation

Workflow

1. calculate pairwise orthographic and semantic distances between words;

Workflow

1. calculate pairwise orthographic and semantic distances between words;
   - semantic distance: $1 - CosSim$, where $CosSim$ is the cosine similarity between word embeddings;

Measuring correlation

Workflow

1. calculate pairwise orthographic and semantic distances between words;
    - semantic distance: $1 - CosSim$, where $CosSim$ is the cosine similarity between word embeddings;
    - $CosSim = 0$ if $CosSim < 0$ (the distance is always within [0...1])

Measuring correlation

Workflow

1. calculate pairwise orthographic and semantic distances between words;
   - semantic distance: $1 - CosSim$, where $CosSim$ is the cosine similarity between word embeddings;
   - $CosSim = 0$ if $CosSim < 0$ (the distance is always within [0...1])
   - for n words, the number of pairs is $n \times (n-1)/2$:

Measuring correlation

Workflow

1. calculate pairwise orthographic and semantic distances between words;
   - semantic distance: $1 - CosSim$, where $CosSim$ is the cosine similarity between word embeddings;
   - $CosSim = 0$ if $CosSim < 0$ (the distance is always within [0...1])
   - for n words, the number of pairs is $n \times (n - 1)/2$:
     - **Mono**: 1 493 856 distances

Measuring correlation

Workflow

1. calculate pairwise orthographic and semantic distances between words;
   - semantic distance: $1 - CosSim$, where $CosSim$ is the cosine similarity between word embeddings;
   - $CosSim = 0$ if $CosSim < 0$ (the distance is always within [0...1])
   - for n words, the number of pairs is $n \times (n-1)/2$:
     - **Mono**: 1 493 856 distances
     - **Bi_NoDim**: 3.5 million distances

Measuring correlation

Workflow

1. calculate pairwise orthographic and semantic distances between words;
   - semantic distance: $1 - CosSim$, where $CosSim$ is the cosine similarity between word embeddings;
   - $CosSim = 0$ if $CosSim < 0$ (the distance is always within [0...1])
   - for n words, the number of pairs is $n \times (n-1)/2$:
     - **Mono**: 1 493 856 distances
     - **Bi_NoDim**: 3.5 million distances
     - **Bi**: 4 million distances

Measuring correlation

Workflow

1. calculate pairwise orthographic and semantic distances between words;
   - semantic distance: $1 - \mathit{CosSim}$, where CosSim is the cosine similarity between word embeddings;
   - $\mathit{CosSim} = 0$ if $\mathit{CosSim} < 0$ (the distance is always within [0...1])
   - for n words, the number of pairs is $n \times (n-1)/2$:
     - **Mono**: 1 493 856 distances
     - **Bi_NoDim**: 3.5 million distances
     - **Bi**: 4 million distances
     - **All**: 22.5 million distances

Measuring correlation

Workflow

1. calculate pairwise orthographic and semantic distances between words;
   - semantic distance: $1 - CosSim$, where $CosSim$ is the cosine similarity between word embeddings;
   - $CosSim = 0$ if $CosSim < 0$ (the distance is always within [0...1])
   - for n words, the number of pairs is $n \times (n-1)/2$:
     - **Mono**: 1 493 856 distances
     - **Bi_NoDim**: 3.5 million distances
     - **Bi**: 4 million distances
     - **All**: 22.5 million distances
2. for each dataset, produce 2 sets of distances (edit and cosine);

Measuring correlation

Workflow

1. calculate pairwise orthographic and semantic distances between words;
    - semantic distance: $1 - CosSim$, where $CosSim$ is the cosine similarity between word embeddings;
    - $CosSim = 0$ if $CosSim < 0$ (the distance is always within [0...1])
    - for n words, the number of pairs is $n \times (n-1)/2$:
        - **Mono**: 1 493 856 distances
        - **Bi_NoDim**: 3.5 million distances
        - **Bi**: 4 million distances
        - **All**: 22.5 million distances
2. for each dataset, produce 2 sets of distances (edit and cosine);
    - $Edit(\text{квас},\ \text{пас}) = 2$
    - $Cosine(\text{квас},\ \text{пас}) = 0.89$

Measuring correlation

Workflow

1. calculate pairwise orthographic and semantic distances between words;
   - semantic distance: $1 - \mathit{CosSim}$, where CosSim is the cosine similarity between word embeddings;
   - $\mathit{CosSim} = 0$ if $\mathit{CosSim} < 0$ (the distance is always within [0...1])
   - for n words, the number of pairs is $n \times (n-1)/2$:
     - **Mono**: 1 493 856 distances
     - **Bi_NoDim**: 3.5 million distances
     - **Bi**: 4 million distances
     - **All**: 22.5 million distances
2. for each dataset, produce 2 sets of distances (edit and cosine);
   - $\mathit{Edit}(\text{квас},\ \text{пас}) = 2$
   - $\mathit{Cosine}(\text{квас},\ \text{пас}) = 0.89$
3. calculate Spearman rank correlation (ρ) between these 2 sets;

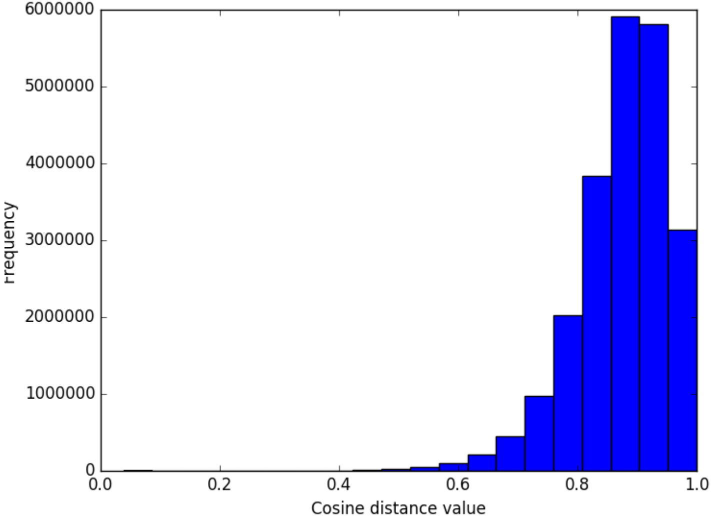
Measuring correlation

Workflow

1. calculate pairwise orthographic and semantic distances between words;
   - semantic distance: $1 - CosSim$, where $CosSim$ is the cosine similarity between word embeddings;
   - $CosSim = 0$ if $CosSim < 0$ (the distance is always within [0...1])
   - for n words, the number of pairs is $n \times (n-1)/2$:
     - **Mono**: 1 493 856 distances
     - **Bi_NoDim**: 3.5 million distances
     - **Bi**: 4 million distances
     - **All**: 22.5 million distances
2. for each dataset, produce 2 sets of distances (edit and cosine);
   - $Edit(\text{квас}, \text{пас}) = 2$
   - $Cosine(\text{квас}, \text{пас}) = 0.89$
3. calculate Spearman rank correlation (ρ) between these 2 sets;
4. pairs similar in form tend to be more similar in meaning?

Measuring correlation

NB: the distances are skewed to the right and not normally distributed:



*Distribution of pairwise cosine distances in the **All** dataset*

Testing significance

Testing significance

- pairwise distances are not independent: changing one character in a word will change several distances, not one;

Measuring correlation

Testing significance

- pairwise distances are not independent: changing one character in a word will change several distances, not one;
- Spearman correlation must be additionally tested for significance;

Measuring correlation

Testing significance

- pairwise distances are not independent: changing one character in a word will change several distances, not one;
- Spearman correlation must be additionally tested for significance;
- we use Mantel permutation test [Mantel, 1967].

Measuring correlation

Testing significance

- pairwise distances are not independent: changing one character in a word will change several distances, not one;
- Spearman correlation must be additionally tested for significance;
- we use Mantel permutation test [Mantel, 1967].

- Mantel test randomly shuffles the values in one of the two sets;
- does it x times;

Measuring correlation

Testing significance

- pairwise distances are not independent: changing one character in a word will change several distances, not one;
- Spearman correlation must be additionally tested for significance;
- we use Mantel permutation test [Mantel, 1967].

- Mantel test randomly shuffles the values in one of the two sets;
- does it x times;
- x correlation values are computed for x 'possible lexicons'.

Measuring correlation

Testing significance

- pairwise distances are not independent: changing one character in a word will change several distances, not one;
- Spearman correlation must be additionally tested for significance;
- we use Mantel permutation test [Mantel, 1967].

- Mantel test randomly shuffles the values in one of the two sets;
- does it x times;
- x correlation values are computed for x ‘possible lexicons’.
- How many random lexicons produced higher correlation than the real one?

Testing significance

- pairwise distances are not independent: changing one character in a word will change several distances, not one;
- Spearman correlation must be additionally tested for significance;
- we use Mantel permutation test [Mantel, 1967].

- Mantel test randomly shuffles the values in one of the two sets;
- does it *x* times;
- *x* correlation values are computed for *x* ‘possible lexicons’.
- How many random lexicons produced higher correlation than the real one?
- If the real data does contain systematicity, the random lexicons will very rarely exhibit the same.

Results

Our results: Mantel test with 1 000 random permutations

Our results: Mantel test with 1 000 random permutations

Dataset	Spearman correlation	Mantel test upper-tail p-value
Mono	0.0310	0.001
Bi_NoDim	0.0519	0.001
Bi	0.0586	0.001
All	0.0800	0.001

Correlations between edit distances and semantic distances

Results

Our results: Mantel test with 1 000 random permutations

Dataset	Spearman correlation	Mantel test upper-tail p-value
Mono	0.0310	0.001
Bi_NoDim	0.0519	0.001
Bi	0.0586	0.001
All	0.0800	0.001

Correlations between edit distances and semantic distances

- $p = 0.001$ means that none of the 1 000 random lexicons exhibited correlation more or equal to the real one.

Results

Our results: Mantel test with 1 000 random permutations

Dataset	Spearman correlation	Mantel test upper-tail p-value
Mono	0.0310	0.001
Bi_NoDim	0.0519	0.001
Bi	0.0586	0.001
All	0.0800	0.001

Correlations between edit distances and semantic distances

- $p = 0.001$ means that none of the 1 000 random lexicons exhibited correlation more or equal to the real one.
- The correlations are extremely significant (though low).

Our results: Mantel test with 1 000 random permutations

Dataset	Spearman correlation	Mantel test upper-tail p-value
Mono	0.0310	0.001
Bi_NoDim	0.0519	0.001
Bi	0.0586	0.001
All	0.0800	0.001

Correlations between edit distances and semantic distances

- $p = 0.001$ means that none of the 1 000 random lexicons exhibited correlation more or equal to the real one.
- The correlations are extremely significant (though low).
- The **Mono** correlation is twice higher than 0.016 reported in [Monaghan et al., 2014] for the set of English mono-morphemic words.

Localizing systematicity

- Why this highly significant correlation is so low?

Localizing systematicity

- Why this highly significant correlation is so low?
- Can it be ‘localized’ in some parts of the lexicon?

Localizing systematicity

- Why this highly significant correlation is so low?
- Can it be 'localized' in some parts of the lexicon?

We split the **Mono** dataset into subsets corresponding to the initial two-character sequences (arguably, phonaesthemes):

Localizing systematicity

- Why this highly significant correlation is so low?
- Can it be 'localized' in some parts of the lexicon?

We split the **Mono** dataset into subsets corresponding to the initial two-character sequences (arguably, phonaesthemes):

- nouns starting with 'ст-',

Localizing systematicity

- Why this highly significant correlation is so low?
- Can it be 'localized' in some parts of the lexicon?

We split the **Mono** dataset into subsets corresponding to the initial two-character sequences (arguably, phonaesthemes):

- nouns starting with 'ст-',
- nouns starting with 'ха-'
- etc...

Localizing systematicity

- Why this highly significant correlation is so low?
- Can it be 'localized' in some parts of the lexicon?

We split the **Mono** dataset into subsets corresponding to the initial two-character sequences (arguably, phonaesthemes):

- nouns starting with 'ст-',
- nouns starting with 'ха-'
- etc...
- this gave us 321 subsets.

Localizing systematicity

- Why this highly significant correlation is so low?
- Can it be 'localized' in some parts of the lexicon?

We split the **Mono** dataset into subsets corresponding to the initial two-character sequences (arguably, phonaesthemes):

- nouns starting with 'ст-',
- nouns starting with 'ха-'
- etc...
- this gave us 321 subsets.

Filtered out:

Localizing systematicity

- Why this highly significant correlation is so low?
- Can it be 'localized' in some parts of the lexicon?

We split the **Mono** dataset into subsets corresponding to the initial two-character sequences (arguably, phonaesthemes):

- nouns starting with 'ст-',
- nouns starting with 'ха-'
- etc...
- this gave us 321 subsets.

Filtered out:

- 159 subsets containing less than 3 nouns;

Localizing systematicity

- Why this highly significant correlation is so low?
- Can it be 'localized' in some parts of the lexicon?

We split the **Mono** dataset into subsets corresponding to the initial two-character sequences (arguably, phonaesthemes):

- nouns starting with 'ст-',
- nouns starting with 'ха-'
- etc...
- this gave us 321 subsets.

Filtered out:

- 159 subsets containing less than 3 nouns;
- 18 subsets with no variance in pairwise edit distances (for example, all distances equal to 1).

Localizing systematicity

- Why this highly significant correlation is so low?
- Can it be ‘localized’ in some parts of the lexicon?

We split the **Mono** dataset into subsets corresponding to the initial two-character sequences (arguably, phonaesthemes):

- nouns starting with ‘ст-’,
- nouns starting with ‘ха-’
- etc...
- this gave us 321 subsets.

Filtered out:

- 159 subsets containing less than 3 nouns;
- 18 subsets with no variance in pairwise edit distances (for example, all distances equal to 1).

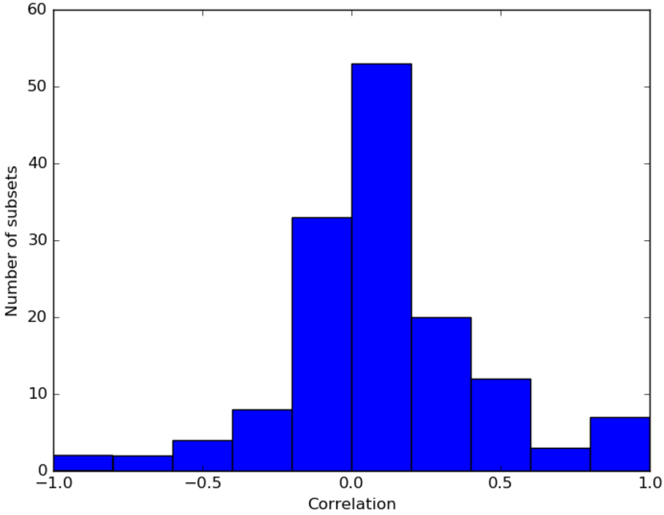
144 ‘valid subsets’ in the end: calculated correlations separately for each of them.

Localizing systematicity

Grouping by initial characters reveals local areas of high systematicity:

Localizing systematicity

Grouping by initial characters reveals local areas of high systematicity:



*Correlations distribution in the subsets of the **Mono** dataset*

Localizing systematicity

Direction of correlation

- In many cases, the correlation ρ was high, but not statistically significant;

Localizing systematicity

Direction of correlation

- In many cases, the correlation ρ was high, but not statistically significant;
  - For example, 'тв-' subset ('тварь', 'твердь', 'твист'): $\rho = 1$, $p = 0.17$.

Localizing systematicity

Direction of correlation

- In many cases, the correlation ρ was high, but not statistically significant;
    - For example, 'тв-' subset ('тварь', 'твердь', 'твист'): $\rho = 1$, $p = 0.17$.
- This is especially true for negative correlations (difficult to interpret anyway).

Localizing systematicity

Direction of correlation

- In many cases, the correlation ρ was high, but not statistically significant;
  - For example, 'тв-' subset ('тварь', 'твердь', 'твист'): $\rho = 1$, $p = 0.17$.
- This is especially true for negative correlations (difficult to interpret anyway).

Can we prove this is not a simple fluctuation?

Localizing systematicity

Direction of correlation

- In many cases, the correlation ρ was high, but not statistically significant;
  - For example, 'тв-' subset ('тварь', 'твердь', 'твист'): $\rho = 1$, $p = 0.17$.
- This is especially true for negative correlations (difficult to interpret anyway).

Can we prove this is not a simple fluctuation?

- Comparison with randomly generated subsets of comparable sizes:

Localizing systematicity

Direction of correlation

- In many cases, the correlation ρ was high, but not statistically significant;
  - For example, 'тв-' subset ('тварь', 'твердь', 'твист'): $\rho = 1$, $p = 0.17$.
- This is especially true for negative correlations (difficult to interpret anyway).

Can we prove this is not a simple fluctuation?

- Comparison with randomly generated subsets of comparable sizes:
  - random subsets follow normal distribution of correlations, concentrate around zero, no outliers;

Localizing systematicity

Direction of correlation

- In many cases, the correlation ρ was high, but not statistically significant;
  - For example, 'тв-' subset ('тварь', 'твердь', 'твист'): $\rho = 1$, $p = 0.17$.
- This is especially true for negative correlations (difficult to interpret anyway).

Can we prove this is not a simple fluctuation?

- Comparison with randomly generated subsets of comparable sizes:
  - random subsets follow normal distribution of correlations, concentrate around zero, no outliers;
  - the initial phonaesthemes based subsets break the normal distribution, introducing strong skew towards high values;

Localizing systematicity

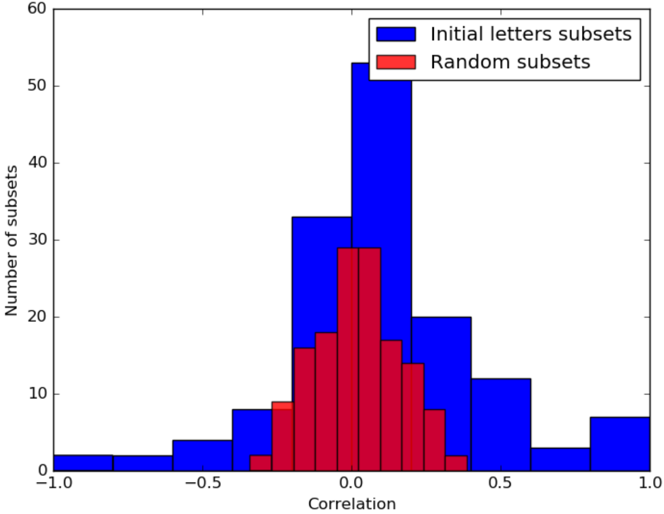
Direction of correlation

- In many cases, the correlation ρ was high, but not statistically significant;
  - For example, 'тв-' subset ('тварь', 'твердь', 'твист'): $\rho = 1$, $p = 0.17$.
- This is especially true for negative correlations (difficult to interpret anyway).

Can we prove this is not a simple fluctuation?

- Comparison with randomly generated subsets of comparable sizes:
  - random subsets follow normal distribution of correlations, concentrate around zero, no outliers;
  - the initial phonaesthemes based subsets break the normal distribution, introducing strong skew towards high values;
  - connection between the form and the meaning is at least partly conditioned by the initial characters.

Localizing systematicity



Correlations distribution in the subsets of the ***Mono*** *dataset*

Localizing systematicity

Top subsets by the correlation ρ ($p < 0.05$):

Localizing systematicity

Top subsets by the correlation ρ ($p < 0.05$):

Initial	ρ	p	Subset size	Examples
ха-	**0.57**	**0.011**	9	хай, хам, харч, хадж...
дж-	0.43	0.047	7	джей, джим, джин...
ше-	0.39	0.015	9	шелк, шерсть, шейх, шельф...
фо-	0.35	0.019	9	фон, фонд, фок, форс...
ва-	0.33	0.017	10	вал, вальс, вар, вамп...
ло-	0.32	**0.011**	13	лов, лоб, лог, лорд, лось...
ле-	0.27	0.012	14	лесть, лещ, лед, лев...
ка-	0.26	0.029	16	кайф, казнь, кадр, кант, кат...
ку-	0.25	0.012	17	куб, культ, курд, кус, куст...
гл-	0.37	0.055	8	глубь, глушь, гладь, глаз...

Discussion

What does that mean?

- the principle of the arbitrariness of linguistic sign in general still holds;

Discussion

What does that mean?

- the principle of the arbitrariness of linguistic sign in general still holds;
- however, there are regular exceptions, manifested throughout the lexicon;

Discussion

What does that mean?

- the principle of the arbitrariness of linguistic sign in general still holds;
- however, there are regular exceptions, manifested throughout the lexicon;
- most of the correlations can probably be explained with rigorous diachronic research:
  - words in the pairs can be cognates, etc..

What does that mean?

- the principle of the arbitrariness of linguistic sign in general still holds;
- however, there are regular exceptions, manifested throughout the lexicon;
- most of the correlations can probably be explained with rigorous diachronic research:
  - words in the pairs can be cognates, etc..
- still, these 'pockets of sound symbolism' [Gutierrez et al., 2016] deserve a deeper analysis.

Discussion

Instead of conclusion

- Graphic form and semantics of Russian nouns do correlate in the present state of language.

Instead of conclusion

- Graphic form and semantics of Russian nouns do correlate in the present state of language.
- $\rho = 0.03$, as calculated on a set of 1 729 mono-syllabic nouns.

Discussion

Instead of conclusion

- Graphic form and semantics of Russian nouns do correlate in the present state of language.
- $\rho = 0.03$, as calculated on a set of 1 729 mono-syllabic nouns.
- This is higher than the reported value for English (0.016).

Instead of conclusion

- Graphic form and semantics of Russian nouns do correlate in the present state of language.
- $\rho = 0.03$, as calculated on a set of 1 729 mono-syllabic nouns.
- This is higher than the reported value for English (0.016).
- In some local lexical subsets, this correlation is even stronger, up to 0.3 and even 0.57 (statistically significant).

Discussion

Instead of conclusion

- Graphic form and semantics of Russian nouns do correlate in the present state of language.
- $\rho = 0.03$, as calculated on a set of 1 729 mono-syllabic nouns.
- This is higher than the reported value for English (0.016).
- In some local lexical subsets, this correlation is even stronger, up to 0.3 and even 0.57 (statistically significant).

The datasets and calculated pairwise distances:
`http://ltr.uio.no/~andreku/arbitrariness/`

Q&A

Arbitrariness of Linguistic Sign Questioned:
Correlation between Word Form and Meaning in Russian

Thank you for your attention!
Questions are welcome.

Andrey Kutuzov
andreku@ifi.uio.no

Dialogue'17

May 31, Moscow, Russia

References I

Bergen, B. K. (2004).
The psychological reality of phonaesthemes.
Language, pages 290–311.

Blasi, D. E., Wichmann, S., Hammarström, H., Stadler, P. F., and Christiansen, M. H. (2016).
Sound–meaning association biases evidenced across thousands of languages.
Proceedings of the National Academy of Sciences, page 201605782.

Gutierrez, E., Levy, R., and Bergen, B. (2016).
Finding non-arbitrary form-meaning systematicity using string-metric learning for kernel regression.
In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2379–2388. Association for Computational Linguistics.

References II

Leviant, I. and Reichart, R. (2015).
Separated by an un-common language: Towards judgment language informed vector space modeling.
arXiv preprint arXiv:1508.00106.

Mantel, N. (1967).
The detection of disease clustering and a generalized regression approach.
Cancer research, 27(2 Part 1):209–220.

Mikhalev, A. (2008).
Psycholinguistic problems of phonaesthemes.
Language being of humans and ethnic groups: cognitive and psycholinguistic aspects, (14):140–148.

References III

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013).
Distributed representations of words and phrases and their compositionality.
Advances in Neural Information Processing Systems 26.

Monaghan, P., Shillcock, R. C., Christiansen, M. H., and Kirby, S. (2014).
How arbitrary is language?
Phil. Trans. R. Soc. B, 369(1651):20130299.

Zhuravlev, A. (1991).
Sound and meaning.
Prosveschenie, 160:1.