Exploring lexical dynamics using diachronic corpora and artificial language experiments

Andres Karjus
CUDAN lab, Tallinn University

In collaboration with: Kenny Smith, Richard A. Blythe, Simon Kirby (University of Edinburgh)

Colloquium for Computational Linguistics and Linguistics in Stuttgart
12.01.2021
- Started postdoc in: 
  [CUDAN](http://cudan.tlu.ee)

- PhD from: 
  [CENTRE for LANGUAGE EVOLUTION](http://example.com)
All living languages keep changing

- All the time
- Eventually diverge into different languages
- This is weird

- This research: focus on lexical change and competition therein
- What happens when new words are introduced into language?

- Massive centuries-spanning corpora open up an unprecedented avenue of possible investigations into language dynamics.
- Variant usage frequencies but also meaning (and change) using distributional semantics methods
In this talk

- Communicative need and lexical competition
  - The topical-cultural advection model
- Semantic similarity and colexification - and communicative need
- Future directions: complexity and informativeness
Some concepts

a semantic space

words

“competition”

lexifies a meaning
another meaning
another word

“colexification”
Complexity and informativeness

inverse of simplicity
relates to learning
cognitive cost

inverse of information loss
accuracy, expressivity
communicative cost

words

Complexity and informativeness

- inverse of simplicity relates to learning cognitive cost
- inverse of information loss
- accuracy, expressivity
- communicative cost

+complexity
-information loss
Complexity and informativeness

Inverse of simplicity relates to learning cognitive cost

Inverse of information loss

Accuracy, expressivity, communicative cost

-complexity

+information loss
Complexity and informativeness

Inverse of simplicity relates to learning cognitive cost

Inverse of information loss

Accuracy, expressivity

Communicative cost

+ Complexity
+ Information loss
The complexity-informativeness tradeoff and the optimal front

The complexity-informativeness tradeoff and the optimal front

The complexity-informativeness tradeoff and the optimal front

describes lexicons of kinship terms, colour, numeral systems, negation; similar optimization effects in artificial language experiments

The complexity-informativeness tradeoff and the optimal front

Communicative need modulates competition in language change

- Preprint: Karjus, Blythe, Kirby, Smith 2020

- As new words, e.g. neologisms & borrowings are selected for, what happens to their older synonyms? Does direct competition always follow local frequency changes?

- Hypothesis:
  - frequency increase in a word will lead to direct competition with (and possibly replacement of) near-synonym(s)
  - unless the lexical subspace experiences high communicative need
Communicative need modulates competition in language change

Figure 1: Example time series from the Corpus of Historical American English (COHA).
Communicative need modulates competition in language change
The corpora

- COHA&DTA: 10-year bins (5 for ERC, Czech, month for Twitter)
- Targets: min +2 log change, occurs min 100x & in
- A model of communicative need
- Need:
  - A model of competition
  - A model of communicative need
A model of communicative need

- Karjus, Blythe, Kirby, Smith 2020, Quantifying the dynamics of topical fluctuations in language. Language Dynamics and Change

- Idea: see how much the topic of a target word changes (weighted mean of the log frequency changes of the relevant topic (context) words of the target)

- Discourse topic prevalence ~ how much something needs to be talked about ~ communicative need

- Topics as the latent flow of language, dragging words along

- *advection* - the transfer of matter (or heat) by the flow of a fluid
A model of communicative need

- *advection* - the transfer of matter (or heat) by the flow of a fluid
Quantifying the dynamics of topical fluctuations in language

Increasing topics: words used more

Topics slowing down: words go out of usage
Advection a proxy to communicative need

- **microchip**
  - transistor, silicon, implant, fiberoptic, infringement, circuitry, semiconductor, converter, workstation, thickest, computing, smokestack, rile, fabrication, subversion, synthesizer, biotechnology, optics, energize, revolutionize, cerebral, ...

- **pantsuit**
  - highheeled, polyester, tailored, longsleeved, pajama, creamcolored, deadpan, beige, unimpressed, violet, quip, corduroy, incongruous, sportsman, devoted, safari, maternity, stylish, streamlined, bustle, joker, tuxedo, lustrous, petite, frosted, portly, navy, ...

A model of linguistic competition

- Meaning from word embeddings; **equalization range**: norm. cosine distance from target where the sum of (normalized) frequency decreases match the increase of the target
- Normalized corpus frequencies sum to 1
- Increase somewhere => decrease somewhere else
- A realistic model of language? Yes: time is finite and learning pressure biases for simpler lexicons. Can’t have infinitely many words.
- Semantics: inferred from LSA, trained for each target word based on (ppmi-weighted) co-occurrence matrix of the preceding time bin, fit target vector into this model – yields neighbours of the *position* where the new word will appear in
A model of linguistic competition

- Meaning from word embeddings; **equalization range**: norm. cosine distance from target where the sum of (normalized) frequency decreases match the increase of the target.
A model of linguistic competition

- Meaning from word embeddings; **equalization range**: norm. cosine distance from target where the sum of (normalized) frequency decreases match the increase of the target
A model of linguistic competition

- Meaning from word embeddings; **equalization range**: norm. cosine distance from target where the sum of (normalized) frequency decreases match the increase of the target
A model of linguistic competition

- Meaning from word embeddings; **equalization range**: norm. cosine distance from target where the sum of (normalized) frequency decreases match the increase of the target.
A model of linguistic competition

- Meaning from word embeddings; **equalization range**: norm. cosine distance from target where the sum of (normalized) frequency decreases match the increase of the target.
A model of linguistic competition
Both models based on lists of words, but decorrelated:

- advection: weighted list of associated, co-occurring words (1\textsuperscript{st} order similarity)
- competition: list of all words, ordered by embedding cosine similarity (2\textsuperscript{nd} order similarity), minus any words in the advection list for a given target

Necessary, but can weaken the competition model accuracy, if closest neighbours (~synonyms) also co-occur with target:

- airplane | aeroplane airship aerial propeller balloon engine machine submarine biplane wireless torpedo
Results

- Topical advection (proxy to communicative need) correlates with
- Equalization range (proxy to extent of competition)
- Lower communicative need: competition more likely
- High communicative need: similar words more likely to coexist

- Lower communicative need: competition more likely
Discussion

- Communicative need, after controlling for a slew of other lexicostatistical variables, describes a small amount of variance in competitive interactions.
- Small effect, but consistent across languages and genres.
- Presumably high communicative need facilitates the co-existence of similar words (more complex lexical subspace).
The complexity-informativeness tradeoff and the optimal front
Further evaluation

- But: direct synonym competition is very rare!
- Sample: COHA, equalization range <0.2 & number of losers <4 (n=52)
- near-synonym competition:
  - aeroplane → airplane, close-up → close shot, appropriation → funding, apartment+inn → motel
- some proper nouns
  - guerrilla → Taliban, Yugoslav → Algerian
- mostly contextual, in-topic replacements:
  - railway → airline, opera+concert → movie, atomic bomb → ballistic missile
- Still, advection predicts if replacement or not
Conceptual similarity and communicative need shape colexification: an experimental study (Karjus, Blythe, Kirby, Wang, Smith, in prep)

- Xu et al 2020, “Conceptual relations predict colexification across languages”, using 200+ languages
- Similar and associated senses (e.g. FIRE and FLAME) are more frequently colexified in world’s languages than unrelated or weakly associated meanings (like FIRE and SALT)
Conceptual similarity and communicative need shape colexification: an experimental study (Karjus, Blythe, Kirby, Wang, Smith, in prep)

- Xu et al 2020, “Conceptual relations predict colexification across languages”, using 200+ languages
- Similar and associated senses (e.g. FIRE and FLAME) are more frequently colexified in world’s languages than unrelated or weakly associated meanings (like FIRE and SALT)
- …but culture specific communicative needs should affect likelihood of colexification – e.g. if it is necessary for efficient communication to distinguish some similar meanings
- E.g. ICE and SNOW: less likely to be colexified in cold climates (Regier et al 2016)
Conceptual similarity and communicative need shape colexification: an experimental study

- What is the cognitive mechanism though that leads to this cross-linguistic tendency?
- Maybe we can test these two claims experimentally?
- 4 experiments: initial one with student sample, replication on Mechanical Turk, 2 more experiments with different conditions
- Dyadic communication game setup, 2 players, take turn sending and guessing messages (cf. Kirby et al 2008, Winters et al 2015)
- 135 rounds each (data from the first 1/3 of the game excluded)
- 10 meanings total
- 4 distractor meanings
- from Simlex999
- 6 target meanings
- 3 pairs

Baseline: pairs co-occur uniformly

Target condition: similar ones occur together more often!

7 signals
The game

Player 1

area  fashion

Communicate area using...

piti
wuli
liha
naru
mano
himu
qata
The game

Player 1

```
area  fashion
Communicate area using...
piti
wuli
liha
naru
mano
himu
qata
```

Player 1

```
Sent area using piti
stand by...
```

Player 2

```
area  fashion
Waiting for message...
```

Player 2

```
area  fashion
This means:
area
fashion
```
Expm no. 38, baseline condition, 96%, counts

7 signals
Analysis

- Exclude low-accuracy dyads (41 left)
- Iterate through each experiment, record each instance of colexification (same signal, different meaning) involving a target meaning; n=1218.
- Logistic mixed effects regression; control for dyads, meaning pairs. Are similar meanings less likely to be colexified in the target condition?
Results

- Yes ($p=0.001$). This includes interaction with round – some dyads change preferences over the course of the game.

- When no pressure to distinguish particular meanings (baseline condition), speakers prefer to colexify similar meanings (confirms Xu et al 2020)

- When need arises to distinguish similar meanings (target condition), speakers less likely to colexify them (confirms hypothesis that communicative needs may block colexification of related concepts)
Follow-up experiments

- Switch to Mechanical Turk: initial experiment was planned to be run the lab in spring 2020, but the apocalypse happened
- Experiment 2, replication on MTurk: exact same setup with 2 conditions; results of experiment 1 replicated.
- Lower accuracy: 79 dyads, could use data only from 53.
Follow-up experiments

- Experiment 3, target condition only: introduce similar-meaning pairs into the distractor set to make colexifying them more natural.
Follow-up experiments

- Experiment 4: no pressure to colexify (10 signals for 10 meanings). No effect, and participants make significantly more use of the bigger signal space. But: natural language does have pressure to simplify (can’t have infinite lexicons).
The complexity-informativeness tradeoff and the optimal front
The complexity-informative tradeoff and the optimal front

<table>
<thead>
<tr>
<th>last</th>
<th>last other</th>
<th>signal</th>
<th>complexity</th>
<th>ambiguity</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>x</td>
<td>x</td>
<td>+0</td>
<td>+1</td>
</tr>
<tr>
<td>x</td>
<td>y</td>
<td>x</td>
<td>+1</td>
<td>+0</td>
</tr>
<tr>
<td>x</td>
<td>x</td>
<td>y</td>
<td>+1</td>
<td>+1</td>
</tr>
<tr>
<td>x</td>
<td>y</td>
<td>y</td>
<td>+1</td>
<td>+2</td>
</tr>
<tr>
<td>x</td>
<td>y</td>
<td>z</td>
<td>+2</td>
<td>+1</td>
</tr>
</tbody>
</table>
Discussion

- Experimental results describe an individual-level lexical choice mechanism which produces results in line with typological colexification tendencies (Xu et al. 2020) as well as the communicative need hypothesis.

- Work in process: a model of lexical density (~extent of colexification) applied to embeddings trained on diachronic corpora.
Conclusions

- Converging typological, experimental and corpus evidence supports the argument for the role of communicative need from earlier cross-linguistic research.

- There are many reasons why languages change; one of them is adaption to the changing needs of their speakers.

- Future: apply the complexity-informativeness approach to products of cumulative cultural evolution other than language.

- Iron out the competition model, apply to data other than language.

- Other stuff: research into semantics-driven misunderstanding and semantic divergence on social media.
Appendix
1925–1944

lemonade
swell
basketball
bomber
habitual

1945–1964

lemonade
swell
basketball
fission
bomber
habitual
armistice
ceasefire
truce