Learning linguistic structure: Integrating big data analysis and visualization

Jackson Lee (PhD candidate in linguistics) **John Goldsmith** (faculty adviser) **Simon Jacobs** (consultant at RCC)
{jsllee, goldsmith, sdjacobs}@uchicago.edu

THE UNIVERSITY OF CHICAGO

Motivation

Linguistic structure is manifested at multiple and interconnected levels: letters/sounds, words, phrases, sentences, etc

Research program:
How can linguistic structure be learned?

Specific problems:

1. **Unsupervised learning**
   Focusing on word structure learning
2. **Visualization**
   Large datasets with millions of word tokens
   ⇒ Visualization for both data and analyses

Approach:
We are constructing *Linguistica 5*, a software with a graphical user interface that integrates linguistic data analysis and visualization.

Word pattern alignment

Current research: Word pattern alignment

Linguistica 4 (Goldsmith 2001) induces word patterns like these:

{jump, walk, ...} {Ø, ed, ing, s} {mov, not, ...} {e, ed, es, ing} etc.

Problem: alignment <u>across</u> word patterns?

How to induce this knowledge:

Ø	ed	ing	s
|	|	|	|
e	ed	ing	es

Solution: We need knowledge of *word categories*.

A graph-theoretic approach to unsupervised word category induction

Nodes: Words | *Edges:* Connect words that are distributionally similar based on n-grams.

⇒ Clustering of distributionally similar words, e.g., infinitives and auxiliary verbs in figure 1.

Inducing word categories: Network community detection, e.g., the Louvain algorithm (Blondel et al. 2008)

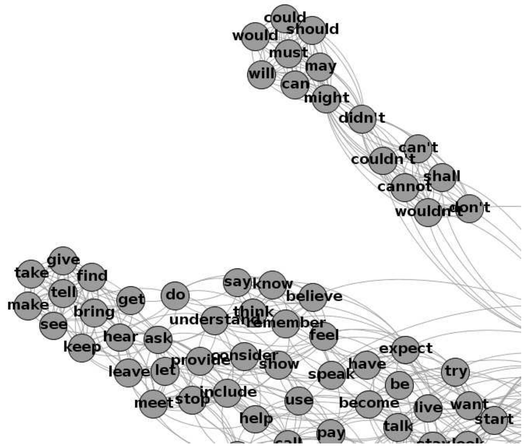


Figure 1

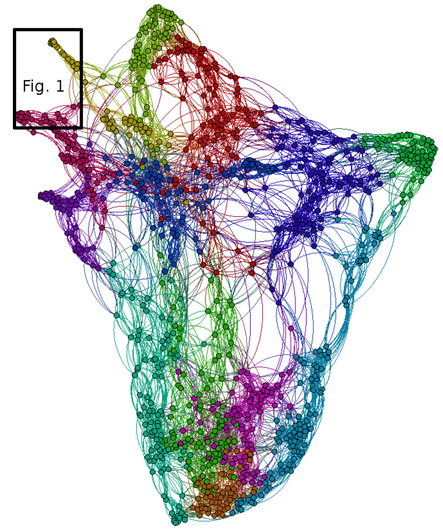
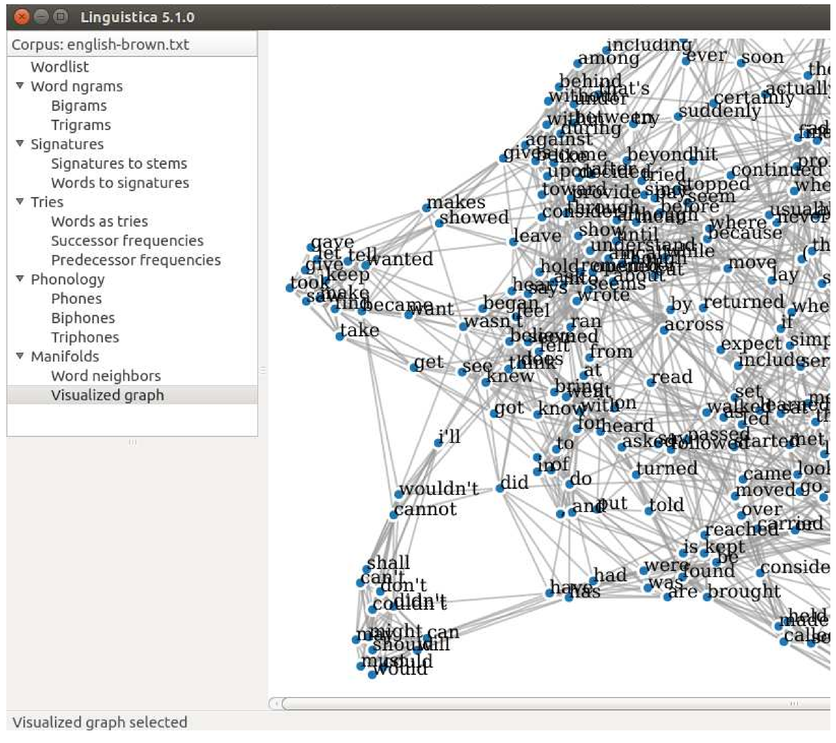


Figure 2

Linguistica 5

Code and datasets: `https://github.com/lxa2015/lxa-py`



Induced word patterns:

	Signature	Stem count	A few stems
1	NULL-s	2327	1970, 2-year-old, ...
2	's-NULL	813	1700, 1800, ...
3	NULL-ly	587	abnormal, absent-...
4	NULL-d-s	346	abuse, accentuate...
5	NULL-d	314	accede, accelerate...
6	ed-ing	197	accost, adjudg, ...
7	'-NULL	190	achaeans, adventi...
8	's-NULL-s	181	1890, 1920, ...
9	d-s	175	accompanie, acco...
10	ies-y	173	abilit, absurdit, ...
11	NULL-ed-ing-s	151	abound, administ...
12	NULL-ed	134	abolish, affix, ...
13	NULL-ed-ing	130	affront, applaud, ...
14	e-ed-es-ing	130	accus, achiev, ...
15	NULL-ing	105	abstain, accompan...
16	d-r	98	amble, betraye, ...
17	e-y	95	admirabl, agreeab...
18	e-ed-ing	88	appeas, beguil, ...

Word trigrams:

	Trigram			Frequency
1	,	and	the	662
2	one	of	the	403
3	the	united	states	328
4	,	however	,	321
5	,	in	the	266
6	.	s	.	266
7	,	he	said	257
8	as	well	as	238
9	u	.	s	235
10	,	it	is	234
11	,	and	he	225
12	of	course	,	220
13	,	of	course	189
14	some	of	the	179
15	the	u	.	176
16	out	of	the	174
17	the	fact	that	167
18	,	but	the	161