Generative Lexicon Theory: Integrating Theoretical and Empirical Methods

James Pustejovsky
Brandeis University

Elisabetta Ježek
University of Pavia

July 11-15, 2016
NASSLLI 2016
Rutgers University

Course Outline

- July 11: Introduction to GL and Data Analytics
- July 12: Qualia Structure
- July 13: Event Structure
- July 14: Argument Structure
- July 15: Meaning Composition

Lecture 1- July 11

Introduction to Generative Lexicon

- Basic concepts in GL
  - Motivation
  - Notation and Language: typed feature structures
  - Meaning Composition in GL
- Polysemy and the Lexicon-Pragmatics Interface
- Evidence-based linguistics and data analytics

Lecture 2- July 12

Qualia Structure

- What is a Quale?
- What motivates Qualia?
- Default Qualia and context updating
- Methodology to identify Qualia
- Data for each Quale
- Qualia and Conventionalized Attributes
- Qualia for Verbs

Lab on Qualia identification and extraction

Lecture 3- July 13

Event Structure

- Events as Structured Objects
- Event Types
  - States
  - Transitions
  - Point Verbs
  - Processes
- Events as Labeled Transition Systems
- Dynamic Event Models

Lab on identification of event types

Lecture 4- July 14

Argument Structure

- Argument Types in GL
  - True Arguments
  - Shadow Arguments
  - Hidden Arguments
- Argument Structure Representation
- Arguments and Defaulting

Lab on hidden and shadow arguments

Lecture 5- July 15

Meaning composition

- Basic Assumptions
- Simple Function Application
- Coercion
- Subselection
- Co-composition

Lab or assignment on coercion

Lecture 1: Introduction to Generative Lexicon

- Language meaning is compositional.
- Compositionality is a desirable property of a semantic model.
- Many linguistic phenomena appear non-compositional.
- Generative Lexicon exploits richer representations and rules to enhance compositional mechanisms.
- Richer representations involve Principles of Decompositionality.
- Richer rules involve Coercion and Co-composition.
- Lexical Resources need to facilitate compositional processes.

Compositional Distinctions in Polysemy

- Inherent polysemy: where multiple interpretations of an expression are available by virtue of the semantics inherent in the expression itself.
- selectional polysemy: where any novel interpretation of an expression is available due to contextual influences, namely, the type of the selecting expression.

1. a. John bought the new Obama book. (pure selection)
   b. John doesn't agree with the new Obama book. (inherent)
2. a. Mary left after her cigarette. (selection as coercion)
   b. Mary left after her smoking a cigarette. (pure selection)

GL Type Structures

(1) a. Natural types:

- *Simple*: Natural kind concepts consisting of reference only to Formal or Constitutive qualia roles;
- *Functional*: Additional reference to Telic (purpose or function)

b. Artifactual types: Concepts making reference to Agentive (origin) for a specific Telic (purpose or function);
c. Complex types: Concepts integrating reference to a logical coherence relation between types from the other two levels.

Kinds of Compositionality

1. Weak Compositionality:
   If all you have for composition is function application, then you need to create as many lexical entries for an expression as there are environments it appears in.
2. True Compositionality: Enrich the mechanisms of making larger meanings by taking advantage of all espressions in the phrase; type coercion, qualia exploitation, co-composition.

Modes of Composition

(2) a. PURE SELECTION (Type Matching): the type a function requires is directly satisfied by the argument;
b. ACCOMMODATION: the type a function requires is inherited by the argument;
c. TYPE COERCION: the type a function requires is imposed on the argument type. This is accomplished by either:
i. *Exploitation*: taking a part of the argument's type to satisfy the function;
ii. *Introduction*: wrapping the argument with the type required by the function.

Notation and Language: typed feature structures

$$\begin{bmatrix} \alpha \\ \text{ARGSTR} = \begin{bmatrix} \text{ARG1} = x \\ \ldots \end{bmatrix} \\ \text{EVENTSTR} = \begin{bmatrix} \text{EVENT1} = e1 \\ \text{EVENT2} = e2 \end{bmatrix} \\ \text{QUALIA} = \begin{bmatrix} \text{CONST} = \textbf{what } x \textbf{ is made of} \\ \text{FORMAL} = \textbf{what } x \textbf{ is} \\ \text{TELIC} = e_2\textbf{: function of } x \\ \text{AGENTIVE} = e_1\textbf{: how } x \textbf{ came into being} \end{bmatrix} \end{bmatrix}$$

Polysemy in language

- What is the meaning of an individual word, out of context?
- Do words carry different meanings in a manner similar to the multiple interpretations that utterance may assume?
- Is there a sharp boundary between monosemy and polysemy in language?
- Is it possibile to maintain a distinction between lexical and pragmatic ambiguity?
- Evidence-based approach.

Context and meaning

- Words take on a different meaning **depending on the context** in which they are used.
- The couple at the next table was laughing.
- The next train is delayed.
- The coexistence of many possible meanings for a word is traditionally referred to as **polysemy**, and it is conceived as **a list of established senses** stored in the lexical entry.

Models of lexical semantics

- Traditional view.
- The denotation of a word may be single or multiple.
  - English *lamp*, denoting the device for giving light.
  - English *paper*, which denotes, inter alia, "the material used for writing" (recycled paper) and an "essay published in an academic journal" (a technical paper).
- A word with a single denotation is called *monosemous*, while a word with multiple denotations is referred to as *polysemous*.
- Polysemy is seen as **a checklist of senses**.
- Sense enumerative lexicons.

Models of lexical semantics

- Dynamic view.
  - Functional notion of polysemy.
  - The ability of lexical items to exhibit different (conceptually) related senses in different contexts, rather than a checklist of separate senses.
- Two major approaches.
  - **Meaning potential**: meaning is attached to units larger than words (.i.e. patterns: corpus linguistics and computational lexicography).
  - **Core meaning** and contextual operations of meaning adjustment.

Polysemy patterns

- At first, polysemy may appear to be an accidental phenomenon, especially when evaluated in relation to single words and in different languages.
- However, when we shift our attention from single words to the entire lexicon, it is possible to identify clear **polysemy patterns**, that is, systematic alternations of meaning that apply to classes of words instead of single words.
- **Regular polysemy** in the terminology introduced by Apresjan 1973 (cf. Dolling 2015 for a recent overview). Other terms are **systematic polysemy** and **logical polysemy**.

Systematic Polysemy

1. There's chicken in the salad.
2. We'll have a water and two beers.
3. Roser finished her thesis.
4. Mary began the novel.
5. Mary believes John's story.
6. Mary believes John.

Accounting for Missing Arguments

Fillmore (1985), Rappaport and Levin (1988), Jackendoff (1990), Levin (1993), Pustejovsky (1995), Goldberg (2002)

- John swept the dirt$_{material}$.
- John swept the room$_{region}$.

- The man shoveled the snow$_{material}$.
- The man shoveled the driveway$_{region}$.

- Mary translated the book. (the translation)
- They decorated the Christmas tree. (the decoration)
- Cathie sliced the bread. (slices)

Flexibility of Argument Interpretation 1/2

- That book bored me terribly.
- The movie frightened Mary.
- The newspaper article angered the Senator.

- The boy heard a cat.
- They heard a bang / rumor / rain.

- Mary believes the rumor.
- She never believes the newspaper.
- The student regrets his last homework assignment.

Flexibility of Argument Interpretation 2/2

- Mary began her beer / thesis / dinner / bath.
- John enjoyed his coffee / movie / a cigar.

- John knows that the earth is round.
- Mary knows what time it is.
- Mary knows the time.

- Mary told John where she lives.
- John told me how old he is.

- Mary told John her address.
- John told me his age.
- I just realized the time.

Metonymic Shifts

- The flight lasted three hours.
- The flight landed safely at about 9 a.m.
- I bought the flight for Christmas.

Hidden Events

- We canceled the taxi.
- From the house I heard the bell.
- We took a break before dessert.

Granularity of Senses

- John started the car.
- You should warm your car up in winter.
- Did you lock the car?
- The car screeched down the road.

Systematic polysemy

- container/content
  - I broke two *glasses*.
  - I drank two *glasses*.
- animal/food
  - The *rabbit* is under the car.
  - She served the *rabbit* with beans.
- process/result
  - The *building* was beginning to take place.
  - The *building* was burned down.

Systematic polysemy

- author/his work
  - *Freud* was born in 1856.
  - *Freud* is on the top shelf.
- institution/place/people
  - The *university* hired a new professor.
  - The *university* is close to the station.
  - This is a friendly *university*.
- event/food
  - It was a long *lunch*.
  - It was a heavy *lunch*.

Copredication evidence

- object and information
  - Jess almost dropped the *book*, then hastily replaced it on the shelf.
  - The author will be discussing her new *book*.
  - This is a bulky and demanding *book*.
- event and food
  - It was a long *lunch*.
  - It was a heavy *lunch*.
  - We had a quick and tasty *lunch* on the terrace.

Inherent polysemy and ontology

- Inherent polysemy can be seen as the linguistic correlate of **ontological complexity**.
- Each sense of an **inherently polysemous noun** denotes a single aspect of an entity which is inherently complex in its constitution.
- The basic idea is that no sense extension by way of metonymy applies in this case because we are still referring to the same object, while with metonymy, this is not the case.
- Cruse 2004 uses the term **facets** to refer to the "senses" of inherent polysemous words.

What counts as a copredication?

- Typically, copredication has been restricted to classic coordinative construction.
- Very few hits of coordination patterns "and" and "but" in corpora.
- Corpus work shows that several patterns are available.
- The book on the shelf is boring.
- The cat was climbing through the open window.
- Fr. La construction, qui a commencé hier, sera très jolie.
- 'The building, which started yesterday, will be very nice'. (Jacquey 2001, 155)

Problems with copredication test

- Copredication structures may involved 'coerced' artifactuals (corpus evidence in Pustejovsky and Jezek 2008, Jezek and Vieu 2015)
- It. Aprire il vino rosso con 30 minuti di anticipo.
- 'open the red wine 30 minutes in advance'.
- Sam grabbed and finished the sandwich in one minute.

Copredication in corpora

- Can we conceive a method to automatically extract inherent polysemous nouns from corpora, and distinguish them from coerced nouns?
- Variability of pairs of predicates in copredication contexts is the key to distinguish inherent polysemous nouns from nouns subject to coercion effects in the context of use (Jezek and Vieu 2015).
- Distributional and lexico-syntactic pattern-based methodology.

Experiment

- Choose a copredication pattern $[V[DetNAdj]]$ and an inherent polysemy pattern: phys•info.
- Identify predicates for the two aspects starting from a list of 10 seed nouns (from classical examples).
- Extract a list of candidate nouns that appear in copredication contexts by running queries of $[V[DetNAdj]]$ with all the context pairs <VPhys, AdjInfo> and <VPhys, AdjInfo>

Copredication in corpora

- We extracted 97 candidate nouns.
- We ranked 12 frequent candidates according to the variability of predicates.
- *articolo*, *lettera*, *pagina*, *libro*, *testo*, *documento*.
- *fenomeno*, *dichiarazione*, *ricerca*.

Evidence-based Language Analysis

- Linguistics is now both a theoretical and experimental discipline
- The scope of observed data for language study and theorizing is richer and broader than ever.
- Linguistic Corpora and captured media datasets will enable contextualized and embodied interpretation of linguistic utterances
- This will enable the development of more expressive and broader theories of language and communication

Methods in Linguistics

- Sapir, Bloomfield, Hockett, Wells (Structural analysis)
  Discovery procedure allows for the emergence of grammatical patterns and constructions in a dataset.
- Chomsky, Bar-Hillel (Transformational Grammars)
  Descriptive procedure allows for the generation of grammatical patterns.
- Chomsky (Generative Grammar 1962- present)
  Explanatory model allows for the generation of best grammatical patterns.

1950-1990 - The Absence of Data

- Chomsky liberated the field of linguistics in the 1950s
- Generation through recursive functions allows one to create your own corpus
- Experiment with new datasets that are not attested in actual "found data"

1990-2016 - The Absence of Theory

- Big Data and statistical modeling has largely dominated the fields of CL and AI, both theoretical and applied.
- Deep Learning seems positioned to obviate theory completely.
- This will not happen: machine learning and deep learning make theoretical assumptions in both the data preparation and feature selection and engineering phase of training.
- Theory is more relevant than ever before.

Corpora for Linguistic Research

- It is quite typical for researchers to use any collection of texts for linguistic analysis. Often proceed opportunistically: whatever data comes in handy is used.
- However, the term corpus usually implies the following characteristics:
  - sampling/representativeness
  - finite size
  - machine-readable form
  - a standard reference
  - (time-bound)

Limitations of Corpora

1. Existence in corpus ≠ grammatical.
   - **Response:** Intuition is necessary, but existence in corpora can point out new assumptions & reduce some biases next slide)
2. Finite corpus cannot capture all possible sentences.
   - **Response:** A corpus can supplement the sentences your brain can generate (& show appropriate context).
3. Grammaticality is not statistical.
   - **Response:** This point is arguable and grammaticality is not everything (cf. language *use*)
4. Corpora are observational, not experimental
   - **Response:** Both are worth investigating: controlled studies and real-world use.

Advantages

Corpus-based & Intuition-based approaches

Being empirical (i.e., using corpora [& experiments]) has advantages over intuition on its own:

- Intuition can be influenced by ideolect or dialect
  - corpus-based approach is free of overt judgments
- Intuition is based on a conscious monitoring of one's production
  - generated sentences may not be typical language use
- Intuition-based examples are difficult to verify

Additionally, corpus-based approaches can show differences that intuition cannot provide

Representativeness

Representativeness: the extent to which a sample includes the full range of variability in a population

- distinguishes corpora from archives
- allows findings to be generalized to a particular variety of language

A corpus is a sample of language use (i.e., from a particular population)

- balance: types of genres
- sampling: how the text is selected

Diachronic dimension

Should corpora be updated regularly?

- And if not, do they become un-representative?

Two general types of corpora:

- **sample corpus**: static corpus
- **monitor corpus**: dynamic corpus which grows

Multiple sample corpora can also provide a view of language change (e.g., Helsinki, LOB corpora)

Lexical Association Measures

- Pointwise Mutual Information:

$$\frac{\text{observed frequency}}{\text{expected frequency}} = PMI(A,B) = log_2 \frac{f_{AB}N}{f_A f_B}$$

- Association Score:

$$\text{AScore}(w_1, R, w_2) = log \frac{||w_1, R, w_2|| \cdot ||*,*,*||}{||w_1, R, *|| \cdot ||*,*,w_2||} \cdot log(||w_1, R, w_2|| + 1)$$

- t test:

$$\frac{\text{observed frequency - expected frequency}}{\sqrt{\text{expected frequency}}}$$

Corpus Analysis Toolkit – SketchEngine

- Corpus creating, loading, handling environment
- Allows extensive querying over the corpus and results of analysis
- Performs statistics and analytics over corpora
- https://www.sketchengine.co.uk/

Corpus Analysis Toolkit – SketchEngine

SketchEngine | English Web 2013 (enTenTen13) | Prof. James Pustejovsk

- Home
- Search
- Word list
- Word sketch
- Thesaurus
- Sketch diff
- Corpus info
- My jobs
- User guide

- Save
- Change options
- Cluster
- Sort by freq
- Hide gramrels
- More data
- Less data

Menu position

read *(verb)*

English Web 2013 (enTenTen13) freq = 9,135,047 (401.91 per million)

modifiers of "read"		
	1,198,223	0.13
aloud +	19,777	9.01
read aloud		
somewhere +	13,428	8.14
I read somewhere that		
carefully +	16,764	8.04
carefully read		
please +	14,495	7.88
please read		
about +	65,599	7.75
read about		
ever +	38,541	7.54
've ever read		
recently +	21,144	7.36
I recently read		
widely +	10,650	7.30
widely read		
first +	18,088	7.26
I first read		
just +	101,055	7.08
just read		
here +	39,368	7.04
read here		
actually +	28,477	6.89
actually read		
below +	7,064	6.85
read below		
never +	29,207	6.35
never read		
then +	28,750	6.35

objects of "read"		
	3,463,987	0.38
book +	341,688	10.82
article +	232,787	10.57
post +	108,274	9.61
blog +	90,960	9.45
review +	85,830	9.36
rest +	65,241	9.02
. Read the rest of this entry		
story +	72,130	8.78
bible +	33,203	8.22
read the Bible		
comment +	39,167	8.19
report +	35,099	7.91
storey +	29,895	7.84
text +	27,936	7.83
page +	30,690	7.70
paper +	26,296	7.67
novel +	23,057	7.67
word +	35,736	7.66
label +	22,937	7.65
letter +	26,827	7.64
newspaper +	21,613	7.61
chapter +	18,873	7.35
magazine +	16,351	7.18
news +	19,248	7.18
email +	17,319	7.09
print +	14,975	7.07
read the fine print		
message +	19,379	7.02

subjects of "read"		
	707,028	0.08
anyone +	24,123	8.30
anyone reading		
continue +	4,469	7.68
... Continue reading the story "		
fun +	5,040	7.44
a fun read		
i +	35,227	7.27
i read		
ive +	3,945	7.26
Ive read		
reader +	4,812	7.09
everyone +	10,555	6.88
statement +	4,007	6.87
statement read .		
sign +	3,428	6.85
sign reads		
min +	2,586	6.85
mins reading time)		
people +	49,234	6.72
people read		
student +	10,469	6.61
students read		
someone +	9,520	6.60
time +	13,379	6.58
time reading		
headline +	2,149	6.55
headline reads		
hour +	3,438	6.52
hours reading		

"read" and/or ...		
	409,895	0.04
write +	68,986	11.34
read and write		
hear +	22,214	10.19
understand +	18,873	9.64
to read and understand		
watch +	11,027	9.27
listen +	10,004	9.13
study +	7,983	8.93
read and study		
see +	11,968	8.73
comment +	5,168	8.54
reading and commenting		
learn +	8,273	8.42
read and learn		
follow +	5,154	8.35
to read and follow		
speak +	5,126	8.22
discuss +	4,704	8.16
read and discuss		
enjoy +	4,837	8.00
read and enjoy		
interpret +	3,391	7.90
to read and interpret		
review +	3,856	7.86
to read and review		
sit +	4,026	7.81
to sit and read		
respond +	3,094	7.72
read and respond to		

prepositional phrases		
	615,601	
"read" about ...	136,992	0.01
"read" in ...	114,644	0.01
"read" through ...	54,402	0.01
"read" on ...	51,404	0.01
"read" from ...	36,399	0.00
"read" by ...	32,859	0.00
"read" for ...	30,399	0.00
"read" to ...	21,754	0.00
"read" at ...	21,572	0.00
"read" of ...	18,864	0.00
"read" with ...	18,105	0.00
"read" as ...	16,589	0.00
"read" like ...	15,533	0.00
"read" between ...	9,952	0.00
"read" over ...	8,752	0.00
"read" into ...	7,976	0.00
"read" during ...	2,639	0.00
"read" before ...	1,986	0.00
"read" so ...	1,355	0.00
"read" than ...	1,301	0.00
"read" without ...	1,232	0.00
"read" after ...	935	0.00
"read" below ...	751	0.00
"read" within ...	713	0.00
"read" until ...	600	0.00