Generative Lexicon: Integrating Theoretical and Distributional Methods

James Pustejovsky
Brandeis University

Elisabetta Ježek
University of Pavia

July 17-21, 2017
1. Introduction to GL and Distributional Analysis
ESSLLI 2017
University of Toulouse

Course Outline

- July 17: Introduction to GL and Distributional Analysis
- July 18: Qualia Structure
- July 19: Event Structure
- July 20: Argument Structure
- July 21: Meaning Composition and Co-composition

Lecture 1: July 17

Introduction to GL and Distributional Analysis

- Basic concepts in GL
  - Notation and Language: typed feature structures
  - Qualia Structure
  - Events and their participants
  - Meaning Composition in GL
- Distributed meaning: Spreading the semantic load
- Polysemy in language
  - Types of contextual variations
  - Detecting copredications in corpora
- Evidence-based linguistics and distributional analysis

Lecture 2: July 18

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

Lab on Qualia identification in corpora using SkE

Lecture 3: July 19

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 type properties in corpora

Lecture 4: July 20

Argument Structure

- Argument Types in GL
  - True Arguments
  - Default Argument
  - Shadow Arguments
  - Hidden Arguments
- Argument Structure Representation
  - Semantic Types and Lexical Sets
  - Distributional Approch to Semantic Types
- Arguments and Defaulting
- Dynamic Argument Structure

Lecture 5: July 21

Meaning Composition and Co-composition in GL

- Basic Assumptions
- Simple Function Application
- Coercion
- Subselection
- Co-composition
- Lexical Information and Pragmatics

Studies in evidence-based coercion

Lecture 1: Introduction to GL and Distributional Analysis

- Language meaning is compositional.

Lecture 1: Introduction to GL and Distributional Analysis

- Language meaning is compositional.
- Compositionality is a desirable property of a semantic model.

Lecture 1: Introduction to GL and Distributional Analysis

- Language meaning is compositional.
- Compositionality is a desirable property of a semantic model.
- Many linguistic phenomena appear non-compositional.

Lecture 1: Introduction to GL and Distributional Analysis

- 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.

Lecture 1: Introduction to GL and Distributional Analysis

- 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.

Lecture 1: Introduction to GL and Distributional Analysis

- 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.

Lecture 1: Introduction to GL and Distributional Analysis

- 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.

Some Factors Contributing to Meaning

- Change of being or state of an individual (e.g., location, size, constitution, or other relation);

Some Factors Contributing to Meaning

- Change of being or state of an individual (e.g., location, size, constitution, or other relation);
- Causation and encoding of agency;

Some Factors Contributing to Meaning

- Change of being or state of an individual (e.g., location, size, constitution, or other relation);
- Causation and encoding of agency;
- Specific attributes of the arguments of the verb;

Some Factors Contributing to Meaning

- Change of being or state of an individual (e.g., location, size, constitution, or other relation);
- Causation and encoding of agency;
- Specific attributes of the arguments of the verb;
- Specification of manner and means of an activity;

Some Factors Contributing to Meaning

- Change of being or state of an individual (e.g., location, size, constitution, or other relation);
- Causation and encoding of agency;
- Specific attributes of the arguments of the verb;
- Specification of manner and means of an activity;
- Temporal or spatial constraints on the event;

Some Factors Contributing to Meaning

- Change of being or state of an individual (e.g., location, size, constitution, or other relation);
- Causation and encoding of agency;
- Specific attributes of the arguments of the verb;
- Specification of manner and means of an activity;
- Temporal or spatial constraints on the event;
- Intentionality of the actor;

Some Factors Contributing to Meaning

- Change of being or state of an individual (e.g., location, size, constitution, or other relation);
- Causation and encoding of agency;
- Specific attributes of the arguments of the verb;
- Specification of manner and means of an activity;
- Temporal or spatial constraints on the event;
- Intentionality of the actor;
- Specification of an instrument involved;

Some Factors Contributing to Meaning

- Change of being or state of an individual (e.g., location, size, constitution, or other relation);
- Causation and encoding of agency;
- Specific attributes of the arguments of the verb;
- Specification of manner and means of an activity;
- Temporal or spatial constraints on the event;
- Intentionality of the actor;
- Specification of an instrument involved;
- Mention of the psychological state of the participants;

Some Factors Contributing to Meaning

- Change of being or state of an individual (e.g., location, size, constitution, or other relation);
- Causation and encoding of agency;
- Specific attributes of the arguments of the verb;
- Specification of manner and means of an activity;
- Temporal or spatial constraints on the event;
- Intentionality of the actor;
- Specification of an instrument involved;
- Mention of the psychological state of the participants;
- Determination of the medium of the situation or event.

Decomposition Strategies: the early history

1. Nida (1949), Hjelmslev (1961), Jakobson (1973)

Decomposition Strategies: the early history

1. Nida (1949), Hjelmslev (1961), Jakobson (1973)
2. Katz and Fodor (1963g), Katz and Postal (1964)

Decomposition Strategies: the early history

1. Nida (1949), Hjelmslev (1961), Jakobson (1973)
2. Katz and Fodor (1963g), Katz and Postal (1964)
3. Tesnière (1959), Fillmore (1965), Lakoff (1965/1970), Gruber (1965), Jackendoff (1972)

Decomposition Strategies: the early history

1. Nida (1949), Hjelmslev (1961), Jakobson (1973)
2. Katz and Fodor (1963g), Katz and Postal (1964)
3. Tesnière (1959), Fillmore (1965), Lakoff (1965/1970), Gruber (1965), Jackendoff (1972)
4. Bierwisch (1967), Garvin (1967)

Decomposition Strategies: the early history

1. Nida (1949), Hjelmslev (1961), Jakobson (1973)
2. Katz and Fodor (1963g), Katz and Postal (1964)
3. Tesnière (1959), Fillmore (1965), Lakoff (1965/1970), Gruber (1965), Jackendoff (1972)
4. Bierwisch (1967), Garvin (1967)
5. Dowty (1972, 1979)

Lehrer's Semantic Field of Factors

<table>
<tr><td colspan="10">cook$_1$</td><td>bake$_1$</td></tr>
<tr><td colspan="11">cook$_2$</td></tr>
<tr><td rowspan="3">steam</td><td colspan="3">boil$_1$</td><td colspan="2">fry</td><td colspan="3">broil</td><td></td><td>roast</td><td>bake$_2$</td></tr>
<tr><td colspan="3">simmer</td><td rowspan="2">boil$_2$</td><td>sauté</td><td>deep-fry
French-fry</td><td>grill</td><td colspan="2">barbecue
charcoal</td></tr>
<tr><td>poach</td><td>stew</td><td>braise</td></tr>
</table>

Table 2.3: *Lexical field of cooking words*, Lehrer (1974)

Lehrer's Semantic Field of Factors

	Water	Oil or fat	Vapor	Amount of liquid	Kind of source of heat	Cooking action	Special utensil	Special additional purpose	Cooking speed
boil$_1$	+	-	-						
boil$_2$	+	-	-			[Vigorous]			
simmer	+	-	-			[Gentle]			
stew	+	-	-			[Gentle]		[To soften]	[Slow]
poach	+	-	-			[Gentle]		[To preserve shape]	
braise	+	+	-	[Small]			(Pot with lid)		
steam	+	-	+				(Rack, sieve)		
fry	-	+	*				(Frying pan)		
sauté	-	+	*	[Small]					[Fast]
French fry deep fry	-	+	*	[Large]					
broil	-	-	*	*	[Radiant]	*			
grill	-	-	*	*	[Radiant]	*	(Grill, griddle)		
barbecue charcoal	-	-	*	*	[Radiant (Hot coals)]	*			
bake	-	-	*	*	[Conducted]	*	(Oven)		
roast	-	-	*	*	[Radiant or conducted]	*			

Table 2.5: *Lexical field of cooking words, componential analysis*, Lehrer (1974)

Generative Lexicon is a Syntax for Decomposition

- Addresses the generative expressiveness of language

Generative Lexicon is a Syntax for Decomposition

- Addresses the generative expressiveness of language
- Semantic load of compositionality is distributed across the phrase

Generative Lexicon is a Syntax for Decomposition

- Addresses the generative expressiveness of language
- Semantic load of compositionality is distributed across the phrase
- Lexical meaning is fundamentally decompositional, i.e. based on the idea that words encode complex concepts that may be decomposed into simpler notions

Generative Lexicon is a Syntax for Decomposition

- Addresses the generative expressiveness of language
- Semantic load of compositionality is distributed across the phrase
- Lexical meaning is fundamentally decompositional, i.e. based on the idea that words encode complex concepts that may be decomposed into simpler notions
- Decompositional processes are compositional in nature

Spreading the Semantic Load of Composition

Pustejovsky 1995

- Problem: Lexical Polysemy is a product of sense overloading:

Spreading the Semantic Load of Composition

Pustejovsky 1995

- Problem: Lexical Polysemy is a product of sense overloading:
  - Mary opened the door / a book / the bottle.
  - John fried an omelette / an egg.
  - The woman finished her coffee / a sandwich / the movie.
- Solution: (a) Distribute the components of meaning; (b) underspecify functional behavior:

Spreading the Semantic Load of Composition

Pustejovsky 1995

- Problem: Lexical Polysemy is a product of sense overloading:
  - Mary opened the door / a book / the bottle.
  - John fried an omelette / an egg.
  - The woman finished her coffee / a sandwich / the movie.
- Solution: (a) Distribute the components of meaning; (b) underspecify functional behavior:
  - Enrich the semantics of nouns: e.g., what doors, books, and bottles are for;
  - Encode what one does with an object: e.g., coffee, sandwich, movie;
  - Generalize the meaning of verbs: e.g., frying is a manner of cooking.

Spreading the Semantic Load of Composition

Given a phrase $[_D A\ B\ C]$:

Spreading the Semantic Load of Composition

Given a phrase $[_D A\ B\ C]$:

- Traditional frameworks identify one function within a phrase:

Spreading the Semantic Load of Composition

Given a phrase $[_D A\ B\ C]$:

- Traditional frameworks identify one function within a phrase:
  - Selectional information is encoded in one word, e.g., $\lambda x \lambda y.B(y,x)$

Spreading the Semantic Load of Composition

Given a phrase $[_D A\ B\ C]$:

- Traditional frameworks identify one function within a phrase:
  - Selectional information is encoded in one word, e.g., $\lambda x \lambda y.B(y,x)$
  - Everything else is either an argument or a modifier, e.g., A, C

Spreading the Semantic Load of Composition

Given a phrase $[_D A\ B\ C]$:

- Traditional frameworks identify one function within a phrase:
  - Selectional information is encoded in one word, e.g., $\lambda x \lambda y.B(y,x)$
  - Everything else is either an argument or a modifier, e.g., A, C
- Generative Lexicon allows all words to act functionally:

Spreading the Semantic Load of Composition

Given a phrase $[_D A\ B\ C]$:

- Traditional frameworks identify one function within a phrase:
  - Selectional information is encoded in one word, e.g., $\lambda x \lambda y.B(y,x)$
  - Everything else is either an argument or a modifier, e.g., A, C
- Generative Lexicon allows all words to act functionally:
  - $\lambda x \lambda y.B(y,x)$
  - $\lambda z.A(z)$
  - $\lambda w.C(w)$

Sense Enumerative Lexicon (SEL)

Sense Enumerative Lexicon (SEL)

1. Different senses of a word are separate lexical entries

Sense Enumerative Lexicon (SEL)

1. Different senses of a word are separate lexical entries
2. Different senses behave differently in composition

Sense Enumerative Lexicon (SEL)

1. Different senses of a word are separate lexical entries
2. Different senses behave differently in composition

1. Lexical entry encodes both syntactic and semantic properties semantics

Sense Enumerative Lexicon (SEL)

1. Different senses of a word are separate lexical entries
2. Different senses behave differently in composition

1. Lexical entry encodes both syntactic and semantic properties
   semantics

   love: love(θ_1,θ_2)

Sense Enumerative Lexicon (SEL)

1. Different senses of a word are separate lexical entries
2. Different senses behave differently in composition

1. Lexical entry encodes both syntactic and semantic properties

   semantics

   love: love(θ_1,θ_2)

   θ_1: HUMAN, θ_2: HUMAN

Sense Enumerative Lexicon (SEL)

1. Different senses of a word are separate lexical entries
2. Different senses behave differently in composition

1. Lexical entry encodes both syntactic and semantic properties
   semantics

   love: love(θ_1,θ_2)

   θ_1: HUMAN, θ_2: HUMAN

1. John loves Mary.

Sense Enumerative Lexicon (SEL)

1. Different senses of a word are separate lexical entries
2. Different senses behave differently in composition

1. Lexical entry encodes both syntactic and semantic properties
   semantics

   love: love(θ_1,θ_2)

   θ_1: HUMAN, θ_2: HUMAN

1. John loves Mary.
2. Apply love(θ_1,θ_2) to Mary

Sense Enumerative Lexicon (SEL)

1. Different senses of a word are separate lexical entries
2. Different senses behave differently in composition

1. Lexical entry encodes both syntactic and semantic properties
   semantics

   love: love(θ_1,θ_2)

   θ_1: HUMAN, θ_2: HUMAN

1. John loves Mary.
2. Apply love(θ_1,θ_2) to Mary
3. $\Longrightarrow$ love(θ_1,Mary)

Sense Enumerative Lexicon (SEL)

1. Different senses of a word are separate lexical entries
2. Different senses behave differently in composition

1. Lexical entry encodes both syntactic and semantic properties
   semantics

   love: love(θ_1,θ_2)

   θ_1: HUMAN, θ_2: HUMAN

1. John loves Mary.
2. Apply love(θ_1,θ_2) to Mary
3. $\Longrightarrow$ love(θ_1,Mary)
4. Apply love(θ_1,Mary) to John

Sense Enumerative Lexicon (SEL)

1. Different senses of a word are separate lexical entries
2. Different senses behave differently in composition

1. Lexical entry encodes both syntactic and semantic properties semantics

   love: love(θ_1,θ_2)

   θ_1: HUMAN, θ_2: HUMAN

1. John loves Mary.
2. Apply love(θ_1,θ_2) to Mary
3. $\Longrightarrow$ love(θ_1,Mary)
4. Apply love(θ_1,Mary) to John
5. $\Longrightarrow$ love(John,Mary)

Lexical ambiguity is handled through separate lexical entries 1/2

1. **bake** (change-of-state): bake(θ_1,θ_2)

Lexical ambiguity is handled through separate lexical entries 1/2

1. **bake** (change-of-state): bake(θ_1,θ_2)
2. **bake** (create) : bake(θ_1,θ_3)

Lexical ambiguity is handled through separate lexical entries 1/2

1. **bake** (change-of-state): bake(θ_1,θ_2)
2. **bake** (create) : bake(θ_1,θ_3)

1. John baked a potato.

Lexical ambiguity is handled through separate lexical entries 1/2

1. **bake** (change-of-state): bake(θ_1,θ_2)
2. **bake** (create) : bake(θ_1,θ_3)

1. John baked a potato.
2. Apply bake(θ_1,θ_2) to a potato

Lexical ambiguity is handled through separate lexical entries 1/2

1. **bake** (change-of-state): bake(θ_1,θ_2)
2. **bake** (create) : bake(θ_1,θ_3)

1. John baked a potato.
2. Apply bake(θ_1,θ_2) to a potato
3. $\Longrightarrow$ bake(θ_1,a_potato)

Lexical ambiguity is handled through separate lexical entries 1/2

1. **bake** (change-of-state): bake(θ_1,θ_2)
2. **bake** (create) : bake(θ_1,θ_3)

1. John baked a potato.
2. Apply bake(θ_1,θ_2) to a potato
3. $\Longrightarrow$ bake(θ_1,a_potato)
4. Apply bake(θ_1,a_potato) to John

Lexical ambiguity is handled through separate lexical entries 1/2

1. **bake** (change-of-state): bake(θ_1,θ_2)
2. **bake** (create) : bake(θ_1,θ_3)

1. John baked a potato.
2. Apply bake(θ_1,θ_2) to a potato
3. $\Longrightarrow$ bake(θ_1,a_potato)
4. Apply bake(θ_1,a_potato) to John
5. $\Longrightarrow$ bake(John,a_potato)

Lexical ambiguity is handled through separate lexical entries 2/2

1. **bake** (change-of-state): bake(θ_1,θ_2)

Lexical ambiguity is handled through separate lexical entries 2/2

1. **bake** (change-of-state): bake(θ_1,θ_2)
2. **bake** (create) : bake(θ_1,θ_3)

Lexical ambiguity is handled through separate lexical entries 2/2

1. **bake** (change-of-state): bake(θ_1,θ_2)
2. **bake** (create) : bake(θ_1,θ_3)

1. John baked a cake.

Lexical ambiguity is handled through separate lexical entries 2/2

1. **bake** (change-of-state): bake(θ_1,θ_2)
2. **bake** (create) : bake(θ_1,θ_3)

1. John baked a cake.
2. Apply bake(θ_1,θ_3) to a cake

Lexical ambiguity is handled through separate lexical entries 2/2

1. **bake** (change-of-state): bake(θ_1,θ_2)
2. **bake** (create) : bake(θ_1,θ_3)

1. John baked a cake.
2. Apply bake(θ_1,θ_3) to a cake
3. $\Longrightarrow$ bake(θ_1,a_cake)

Lexical ambiguity is handled through separate lexical entries 2/2

1. **bake** (change-of-state): bake(θ_1,θ_2)
2. **bake** (create) : bake(θ_1,θ_3)

1. John baked a cake.
2. Apply bake(θ_1,θ_3) to a cake
3. $\Longrightarrow$ bake(θ_1,a_cake)
4. Apply bake(θ_1,a_cake) to John

Lexical ambiguity is handled through separate lexical entries 2/2

1. **bake** (change-of-state): bake(θ_1,θ_2)
2. **bake** (create) : bake(θ_1,θ_3)

1. John baked a cake.
2. Apply bake(θ_1,θ_3) to a cake
3. $\Longrightarrow$ bake(θ_1,a_cake)
4. Apply bake(θ_1,a_cake) to John
5. $\Longrightarrow$ bake(John,a_cake)

Integrating GL with Distributional Semantics

1. Context vectors for words encode syntagmatic relations between words

$$w_i \quad \textbf{Verb} \quad w_j$$

Integrating GL with Distributional Semantics

1. Context vectors for words encode syntagmatic relations between words

$$w_i \quad \textbf{Verb} \quad w_j$$

2. Abstractions or clusters over contexts give rise to paradigmatic relations between words

Integrating GL with Distributional Semantics

1. Context vectors for words encode syntagmatic relations between words

$$w_i \quad \textbf{Verb} \quad w_j$$

2. Abstractions or clusters over contexts give rise to paradigmatic relations between words

$$\begin{cases} v_1 \\ v_2 \\ \dots \\ v_n \end{cases} \quad \textbf{Verb} \quad \begin{cases} w_1 \\ w_2 \\ \dots \\ w_n \end{cases}$$

Compositional Distinctions in Polysemy

- Inherent polysemy: where multiple interpretations of an expression are available by virtue of the semantics inherent in the expression itself.

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.

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)

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:

GL Type Structures

(2) a. Natural types:

- *Simple*: Natural kind concepts consisting of reference only to Formal or Constitutive qualia roles;

GL Type Structures

(3) 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)

GL Type Structures

(4) 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);

GL Type Structures

(5) 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.

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

(6) a. PURE SELECTION (Type Matching): the type a function requires is directly satisfied by the argument;

Modes of Composition

(7) 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;

Modes of Composition

(8) 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:

Modes of Composition

(9) 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;

Modes of Composition

(10) 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?

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?

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?

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?

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 lexical meaning

- Words are able to take on a different meaning **depending on the context** in which they are used.

Context and lexical meaning

- Words are able to take on a different meaning **depending on the context** in which they are used.
- The couple at the next table was laughing.

Context and lexical meaning

- Words are able to 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.

Context and lexical meaning

- Words are able to 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 next costumer, the next slide ...

Context and lexical meaning

- The coexistence of many possible meanings for a word is traditionally referred to as **polysemy**, and it is conceived as **a list of pre-defined established senses** stored in the lexical entry.

Context and lexical meaning

- The coexistence of many possible meanings for a word is traditionally referred to as **polysemy**, and it is conceived as **a list of pre-defined established senses** stored in the lexical entry.
- **checklist theory** of lexical meaning.

Context and lexical meaning

- The coexistence of many possible meanings for a word is traditionally referred to as **polysemy**, and it is conceived as **a list of pre-defined established senses** stored in the lexical entry.
- **checklist theory** of lexical meaning.
- **sense enumeration** lexicon.

Context and lexical meaning

- The coexistence of many possible meanings for a word is traditionally referred to as **polysemy**, and it is conceived as **a list of pre-defined established senses** stored in the lexical entry.
- **checklist theory** of lexical meaning.
- **sense enumeration** lexicon.
- This is the standard way dictionaries and **resources** used for NLP tasks (i.e. WordNet for word sense detection etc.) are put together.

Types of contextual variation

Properties of objects coming into the foreground in the context.

Types of contextual variation

Properties of objects coming into the foreground in the context.

- This car *weighs* over 2,000 lbs.

Types of contextual variation

Properties of objects coming into the foreground in the context.

- This car *weighs* over 2,000 lbs.
- John started the car.

Types of contextual variation

Properties of objects coming into the foreground in the context.

- This car *weighs* over 2,000 lbs.
- John started the car.
- You should warm your car up in winter.

Types of contextual variation

Properties of objects coming into the foreground in the context.

- This car *weighs* over 2,000 lbs.
- John started the car.
- You should warm your car up in winter.
- Did you lock the car?

Types of contextual variation

Properties of objects coming into the foreground in the context.

- This car *weighs* over 2,000 lbs.
- John started the car.
- You should warm your car up in winter.
- Did you lock the car?
- The car screeched down the road.

Types of contextual variation

Properties of objects coming into the foreground in the context.

- This car *weighs* over 2,000 lbs.
- John started the car.
- You should warm your car up in winter.
- Did you lock the car?
- The car screeched down the road.

Pustejovsky and Jezek 2012 *Introducing Qualia Stucture*

Types of contextual variation

Contextual coercions

Types of contextual variation

Contextual coercions

- The flight lasted three hours.

Types of contextual variation

Contextual coercions

- The flight lasted three hours.
- The flight landed safely at about 9 a.m.

Types of contextual variation

Contextual coercions

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

Types of contextual variation

Contextual coercions

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

Types of contextual variation

Contextual coercions

- The flight lasted three hours.
- The flight landed safely at about 9 a.m.
- I bought the flight for Christmas.
- You reached the house.
- Do you want the whole house waken up?

Types of contextual variation

Contextual coercions

- The flight lasted three hours.
- The flight landed safely at about 9 a.m.
- I bought the flight for Christmas.
- You reached the house.
- Do you want the whole house waken up?
- The rest of the house was sleeping.

Types of contextual variation

Contextual coercions

- The flight lasted three hours.
- The flight landed safely at about 9 a.m.
- I bought the flight for Christmas.
- You reached the house.
- Do you want the whole house waken up?
- The rest of the house was sleeping.

Pustejovsky and Jezek 2008, Jezek and Quochi 2010, Pustejovsky et al. 2010 SemEval-2010 Task 7 Argument Selection and Coercion.

Types of contextual variations

Hidden Events

Types of contextual variations

Hidden Events

- We canceled the taxi.

Types of contextual variations

Hidden Events

- We canceled the taxi.
- From the house I heard the bell.

Types of contextual variations

Hidden Events

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

Types of contextual variations

Hidden Events

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

Types of contextual variations

Hidden Events

- We canceled the taxi.
- From the house I heard the bell.
- We took a break before dessert.
- They finished the beer.
- They finished their cake.

Types of contextual variations

Hidden Events

- We canceled the taxi.
- From the house I heard the bell.
- We took a break before dessert.
- They finished the beer.
- They finished their cake.
- Any chocolate? Not after that cake!

Types of contextual variations

Hidden Events

- We canceled the taxi.
- From the house I heard the bell.
- We took a break before dessert.
- They finished the beer.
- They finished their cake.
- Any chocolate? Not after that cake!
- I prefer cake to biscuits.

Types of contextual variations

Hidden Events

- We canceled the taxi.
- From the house I heard the bell.
- We took a break before dessert.
- They finished the beer.
- They finished their cake.
- Any chocolate? Not after that cake!
- I prefer cake to biscuits.
- We skipped the cake and settled for another coffee.

Types of contextual variations

Hidden Events

- We canceled the taxi.
- From the house I heard the bell.
- We took a break before dessert.
- They finished the beer.
- They finished their cake.
- Any chocolate? Not after that cake!
- I prefer cake to biscuits.
- We skipped the cake and settled for another coffee.

Pustejovsky and Anick 1988 (later "Qualia roles"; data from Pustejovsky and Jezek 2012).

Types of contextual variations

Systematic polysemies (Apresjan 1973)

Types of contextual variations

Systematic polysemies (Apresjan 1973)

- container/content

Types of contextual variations

Systematic polysemies (Apresjan 1973)

- container/content
  - I broke two glasses.

Types of contextual variations

Systematic polysemies (Apresjan 1973)

- container/content
  - I broke two glasses.
  - I drank two glasses.

Types of contextual variations

Systematic polysemies (Apresjan 1973)

- container/content
  - I broke two glasses.
  - I drank two glasses.
- institution/place/people

Types of contextual variations

Systematic polysemies (Apresjan 1973)

- container/content
  - I broke two glasses.
  - I drank two glasses.
- institution/place/people
  - The university hired a new professor.

Types of contextual variations

Systematic polysemies (Apresjan 1973)

- container/content
  - I broke two glasses.
  - I drank two glasses.
- institution/place/people
  - The university hired a new professor.
  - The university is close to the station.

Types of contextual variations

Systematic polysemies (Apresjan 1973)

- container/content
  - I broke two glasses.
  - I drank two glasses.
- institution/place/people
  - The university hired a new professor.
  - The university is close to the station.
  - This is a friendly university.

Types of contextual variations

Systematic polysemies (Apresjan 1973)

- container/content
  - I broke two glasses.
  - I drank two glasses.
- institution/place/people
  - The university hired a new professor.
  - The university is close to the station.
  - This is a friendly university.
- process/result

Types of contextual variations

Systematic polysemies (Apresjan 1973)

- container/content
  - I broke two glasses.
  - I drank two glasses.
- institution/place/people
  - The university hired a new professor.
  - The university is close to the station.
  - This is a friendly university.
- process/result
  - The building was beginning to take place.

Types of contextual variations

Systematic polysemies (Apresjan 1973)

- container/content
  - I broke two glasses.
  - I drank two glasses.
- institution/place/people
  - The university hired a new professor.
  - The university is close to the station.
  - This is a friendly university.
- process/result
  - The building was beginning to take place.
  - The building was burned down.

Types of contextual variations

Systematic polysemies (Apresjan 1973)

- container/content
  - I broke two glasses.
  - I drank two glasses.
- institution/place/people
  - The university hired a new professor.
  - The university is close to the station.
  - This is a friendly university.
- process/result
  - The building was beginning to take place.
  - The building was burned down.

Types of contextual variation

Inherent polysemies (Pustejovsky 1995)

Types of contextual variation

Inherent polysemies (Pustejovsky 1995)

- object and information

Types of contextual variation

Inherent polysemies (Pustejovsky 1995)

- object and information
  - Jess almost dropped the book, then hastily replaced it on the shelf.

Types of contextual variation

Inherent polysemies (Pustejovsky 1995)

- object and information
  - Jess almost dropped the book, then hastily replaced it on the shelf.
  - The author will be discussing her new book.

Types of contextual variation

Inherent polysemies (Pustejovsky 1995)

- 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*.

Types of contextual variation

Inherent polysemies (Pustejovsky 1995)

- 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

Types of contextual variation

Inherent polysemies (Pustejovsky 1995)

- 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.

Types of contextual variation

Inherent polysemies (Pustejovsky 1995)

- 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.

Types of contextual variation

Inherent polysemies (Pustejovsky 1995)

- 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.

Types of contextual variation

Inherent polysemies (Pustejovsky 1995)

- 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.

Cruse 1995's *nouns with facets*, Asher's 2011 *dual aspect nouns*.

What counts as a copredication?

- Typically, copredication has been restricted to classic **coordinative construction**.

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; when present, they do not frequently mix different aspect but tends to predicate on a single aspect.

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; when present, they do not frequently mix different aspect but tends to predicate on a single aspect.
  - It. La costruzione fu lenta e paziente.

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; when present, they do not frequently mix different aspect but tends to predicate on a single aspect.
  - It. La costruzione fu lenta e paziente.
    'The construction was slow and patient.'
  - It. La costruzione era solida e stabile.

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; when present, they do not frequently mix different aspect but tends to predicate on a single aspect.
  - It. La costruzione fu lenta e paziente.
    'The construction was slow and patient.'
  - It. La costruzione era solida e stabile.
    'The building was solid and stable.'

Jezek and Melloni 2011.

What counts as a copredication?

- Corpus work shows that several patterns are available **beyond coordination**.
  - The book on the shelf is boring.

What counts as a copredication?

- Corpus work shows that several patterns are available **beyond coordination**.
  - The book on the shelf is boring.
  - The cat was climbing through the open window.

What counts as a copredication?

- Corpus work shows that several patterns are available **beyond coordination**.
  - 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.

What counts as a copredication?

- Corpus work shows that several patterns are available **beyond coordination**.
  - 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
  - It. Una volta completata, la traduzione si può caricare in una sezione apposita del sito.

What counts as a copredication?

- Corpus work shows that several patterns are available **beyond coordination**.
  - 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
  - It. Una volta completata, la traduzione si può caricare in una sezione apposita del sito.
    'Once completed, the translation may be uploaded in a special section of the site'.
    Jezek and Melloni 2011, 27

Experimental setting

- We conduct a research to automatically extract copredication contexts from corpora and identify a list of candidate inherently polysemous nouns.

Experimental setting

- We conduct a research to automatically extract copredication contexts from corpora and identify a list of candidate inherently polysemous nouns.
- We use the tagged Italian ItTenTen10 (2,5 Gigawords) corpus queried through Sketch Engine and its API.

Experimental setting

- We conduct a research to automatically extract copredication contexts from corpora and identify a list of candidate inherently polysemous nouns.
- We use the tagged Italian ItTenTen10 (2,5 Gigawords) corpus queried through Sketch Engine and its API.
- We focus on the copredication pattern $[V[Det\_N\_Adj]]$ and on the inherent polysemy scheme physical_object•information.

Experimental setting

- We conduct a research to automatically extract copredication contexts from corpora and identify a list of candidate inherently polysemous nouns.
- We use the tagged Italian ItTenTen10 (2,5 Gigawords) corpus queried through Sketch Engine and its API.
- We focus on the copredication pattern $[V[Det\_N\_Adj]]$ and on the inherent polysemy scheme physical_object•information.
  - It. Consultare un libro voluminoso.

Experimental setting

- We conduct a research to automatically extract copredication contexts from corpora and identify a list of candidate inherently polysemous nouns.
- We use the tagged Italian ItTenTen10 (2,5 Gigawords) corpus queried through Sketch Engine and its API.
- We focus on the copredication pattern $[V[Det\_N\_Adj]]$ and on the inherent polysemy scheme physical_object•information.
  - It. Consultare un libro voluminoso.
    'Consult a bulky book'
  - It. ...bruciavano i libri controversi.

Experimental setting

- We conduct a research to automatically extract copredication contexts from corpora and identify a list of candidate inherently polysemous nouns.
- We use the tagged Italian ItTenTen10 (2,5 Gigawords) corpus queried through Sketch Engine and its API.
- We focus on the copredication pattern $[V[Det\_N\_Adj]]$ and on the inherent polysemy scheme physical_object•information.
  - It. Consultare un libro voluminoso.
    'Consult a bulky book'
  - It. ...bruciavano i libri controversi.
    '...they burned the controversial books'.

Distributional analysis of copredication in Jezek and Vieu 2014, Vieu, Jezek and VanDeCruys 2015.

Predicate selection

- Effective copredication contexts extraction require identification of predicates that select the different aspects of the noun (i.e. adjective and verbs selecting for either the **information** or the **physical_object** aspect to be tested).

Predicate selection

- Effective copredication contexts extraction require identification of predicates that select the different aspects of the noun (i.e. adjective and verbs selecting for either the **information** or the **physical_object** aspect to be tested).
- Here, we need 4 lists: **VPhys**, **AdjInfo**, **VInfo**, **AdjPhys**.

Predicate selection

- Effective copredication contexts extraction require identification of predicates that select the different aspects of the noun (i.e. adjective and verbs selecting for either the **information** or the **physical_object** aspect to be tested).
- Here, we need 4 lists: **VPhys**, **AdjInfo**, **VInfo**, **AdjPhys**.
- Manual extraction of predicates is costly and time-consuming.

Predicate selection

- Effective copredication contexts extraction require identification of predicates that select the different aspects of the noun (i.e. adjective and verbs selecting for either the **information** or the **physical_object** aspect to be tested).
- Here, we need 4 lists: **VPhys**, **AdjInfo**, **VInfo**, **AdjPhys**.
- Manual extraction of predicates is costly and time-consuming.
- We exploit distributional semantics to extract predicates.

Predicate selection

- Effective copredication contexts extraction require identification of predicates that select the different aspects of the noun (i.e. adjective and verbs selecting for either the **information** or the **physical_object** aspect to be tested).
- Here, we need 4 lists: **VPhys**, **AdjInfo**, **VInfo**, **AdjPhys**.
- Manual extraction of predicates is costly and time-consuming.
- We exploit distributional semantics to extract predicates.
- We use a model that relies on latent dimensions computed by non-negative matrix factorization.

Semi-automatic predicate extraction

- Latent semantic distributional model to semi-automatically extract predicates from corpus.

Semi-automatic predicate extraction

- Latent semantic distributional model to semi-automatically extract predicates from corpus.
  Hypothesis: the **latent dimensions** obtained hint at particular predication contexts, such as `Phys`, `Info` or both.
- Method: extension of non-negative matrix factorization (Van de Cruys, 2008)

Semi-automatic predicate extraction

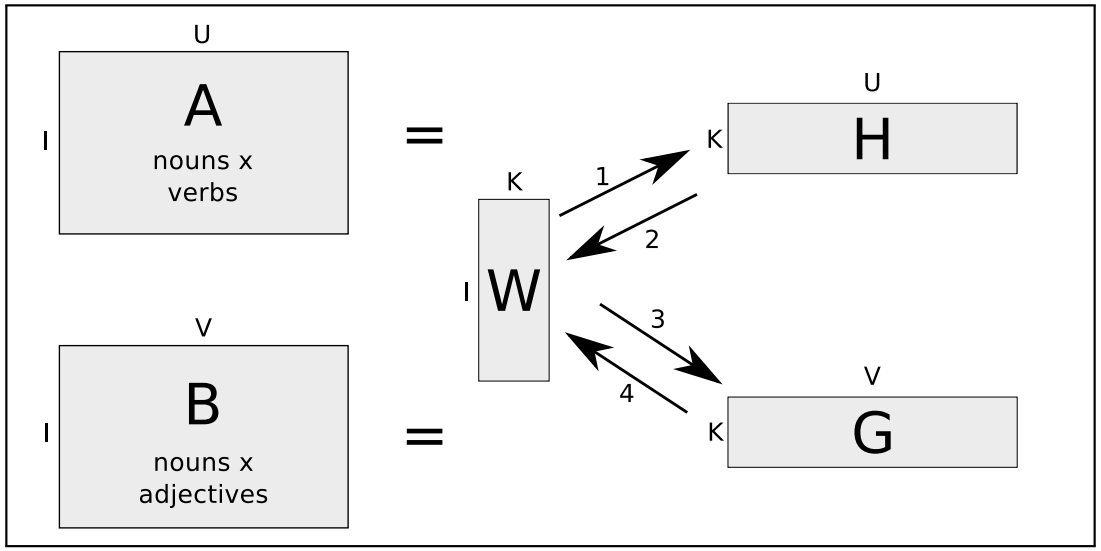
- Latent semantic distributional model to semi-automatically extract predicates from corpus.
  Hypothesis: the **latent dimensions** obtained hint at particular predication contexts, such as `Phys`, `Info` or both.
- Method: extension of non-negative matrix factorization (Van de Cruys, 2008)
  - NMF: Non-negative matrix **A** is factorized into two other non-negative matrices:

$$A_{i \times j} \approx W_{i \times k} H_{k \times j} \tag{1}$$

where k is much smaller than i, j (Lee, 2001)

Semi-automatic predicate extraction

- Extension of NMF to jointly induce latent factors for 3 modes (N, V, Adj) in the co-predication pattern [V [Det N Adj]] by interleaved factorizations



- Exploits syntactic constraints to obtain semantic similarity, not just topic similarity.

Extracting the latent dimensions

- Extract co-occurrence frequencies from freely available ItWaC corpus (Baroni et al, 2009), using most frequently occurring 1K verbs, 4K nouns and 2K adjectives.

Extracting the latent dimensions

- Extract co-occurrence frequencies from freely available ItWaC corpus (Baroni et al, 2009), using most frequently occurring 1K verbs, 4K nouns and 2K adjectives.
- Set the number of latent dimensions to k 100.

Extracting the latent dimensions

- Extract co-occurrence frequencies from freely available ItWaC corpus (Baroni et al, 2009), using most frequently occurring 1K verbs, 4K nouns and 2K adjectives.
- Set the number of latent dimensions to k 100.
- Select dimensions and take their first 20 items.

Extracting the latent dimensions

- Extract co-occurrence frequencies from freely available ItWaC corpus (Baroni et al, 2009), using most frequently occurring 1K verbs, 4K nouns and 2K adjectives.
- Set the number of latent dimensions to k 100.
- Select dimensions and take their first 20 items.
  Yields 4 lists of predicates VPhys, AdjInfo, VInfo, and AdjPhys
  Two alternative methods:
  - Manually review the 15 dimensions most associated with the 10 Info•Phys **seed nouns** to pick between 2 and 5 dimensions for each list (DimsN).
  - Automatically pick the 5 dimensions most associated with 10 **seed predicates** (manually chosen) for each list (DimspP).

Verbs	Adjectives	Nouns
narrare (narrate)	*antico* (ancient)	*leggenda* (legend)
raccontare (tell)	*greco* (Greek)	*favola* (fable)
imparare (learn)	*volgare* (vulgar)	*fiaba* (fable)
conoscere (know)	*latino* (Latin)	*storia* (story)
inventare (invent)	*crudele* (cruel)	*latino* (Latin)
evocare (evoke)	*medievale* (medieval)	*greco* (Greek)
apprendere (learn)	*saggio* (wise)	*dialetto* (dialect)
credere (believe)	*triste* (sad)	*mito* (myth)
sognare (dream)	*medioevale* (medieval)	*antico* (ancient)
insegnare (teach)	*romantico* (romantic)	*mestiere* (profession)
recitare (recite)	*napoletano* (Neapolitan)	*eroe* (hero)
sapere (know)	*italico* (Italic)	*poesia* (poetry)
tradurre (translate)	*eroico* (heroic)	*lingua* (tongue)
parlare (talk)	*nobile* (noble)	*poeta* (poet)
amare (love)	*parlato* (spoken)	*danza* (dance)
ispirare (inspire)	*indiano* (Indian)	*arabo* (Arab)
dipingere (paint)	*popolare* (popular)	*comico* (comic)
adorare (adore)	*orientale* (eastern)	*accento* (accent)
diventare (become)	*moderno* (modern)	*spagnolo* (Spaniard)
vivere (live)	*cinese* (Chinese)	*dramma* (drama)

A dimension picked for both VInfo and AdjInfo

Verbs	Adjectives	Nouns
compilare (compile)	*cartaceo* (of paper)	*fotocopia* (photocopy)
allegare (attach)	*elettronico* (electronic)	*copia* (copy)
spedire (send)	*allegato* (attached)	*certificato* (certificate)
corredare (equip)	*inviato* (sent)	*documento* (document)
inviare (send)	*apposito* (specific)	*ricevuta* (receipt)
inoltrare (forward)	*modulistico* (of form)	*modulo* (form)
stampare (print)	*leggibile* (readable)	*questionario* (questionnaire)
copiare (copy)	*firmato* (signed)	*autocertificazione* (self-certification)
archiviare (file)	*informatico* (of computer)	*pdf* (pdf)
ricevere (receive)	*digitale* (digital)	*documentazione* (documentation)
modulare (modulate)	*valido* (valid)	*informazione* (information)
certificare (certify)	*scaricabile* (downloadable)	*E-mail* (e-mail)
trasmettere (transmit)	*On-line* (on-line)	*dato* (datum)
consegnare (deliver)	*telematico* (telematic)	*posta* (mail)
depositare (deposit)	*redatto* (written)	*verbale* (report)
reperire (find)	*disponibile* (available)	*originale* (original)
redigere (write)	*lino* (of linen)	*scheda* (card)
sottoscrivere (sign)	*postale* (postal)	*certificazione* (certificate)
pervenire (reach)	*reperibile* (available)	*autenticazione* (authentication)
munire (provide)	*identificativo* (identifying)	*formato* (format)

A dimension picked for both VPhys and AdjPhys

Results of predicate extraction

- We started with 10 seed nouns (*articolo* 'article', *diario* 'diary', *documento* 'document', *etichetta* 'label', *fumetto* 'comic', *giornale* 'newspaper', *lettera* 'letter', *libro* 'book', *racconto* 'short novel', *romanzo* 'novel').

Results of predicate extraction

- We started with 10 seed nouns (*articolo* 'article', *diario* 'diary', *documento* 'document', *etichetta* 'label', *fumetto* 'comic', *giornale* 'newspaper', *lettera* 'letter', *libro* 'book', *racconto* 'short novel', *romanzo* 'novel').
- We applied an extension of NMF to jointly induce latent factors for three different modes (N, V, and A).

Results of predicate extraction

- We started with 10 seed nouns (*articolo* 'article', *diario* 'diary', *documento* 'document', *etichetta* 'label', *fumetto* 'comic', *giornale* 'newspaper', *lettera* 'letter', *libro* 'book', *racconto* 'short novel', *romanzo* 'novel').
- We applied an extension of NMF to jointly induce latent factors for three different modes (N, V, and A).
- We produced matrixes with the pairwise co-occurrence frequencies for the different modes and then **interleaved** then.

Results of predicate extraction

- We started with 10 seed nouns (*articolo* 'article', *diario* 'diary', *documento* 'document', *etichetta* 'label', *fumetto* 'comic', *giornale* 'newspaper', *lettera* 'letter', *libro* 'book', *racconto* 'short novel', *romanzo* 'novel').
- We applied an extension of NMF to jointly induce latent factors for three different modes (N, V, and A).
- We produced matrixes with the pairwise co-occurrence frequencies for the different modes and then **interleaved** then.
- With the aid of manual selection, we obtained 4 lists containing from 37 **AdjPhys** to 91 info **AdjInfo**, which we used as fillers for the pattern to extract copredications contexts.

Results of predicate extraction

- We started with 10 seed nouns (*articolo* 'article', *diario* 'diary', *documento* 'document', *etichetta* 'label', *fumetto* 'comic', *giornale* 'newspaper', *lettera* 'letter', *libro* 'book', *racconto* 'short novel', *romanzo* 'novel').
- We applied an extension of NMF to jointly induce latent factors for three different modes (N, V, and A).
- We produced matrixes with the pairwise co-occurrence frequencies for the different modes and then **interleaved** then.
- With the aid of manual selection, we obtained 4 lists containing from 37 **AdjPhys** to 91 info **AdjInfo**, which we used as fillers for the pattern to extract copredications contexts.

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.

Context-sensitive model of lexical semantics

- **Semantic flexibility** is a property of natural language.
- The meaning of each word is expected to vary from occurrence to occurrence as a function of the interaction with the other words it combines with, and of the situation of utterance.
- **Functional notion** of polysemy.
- Generative Lexicon proposes that context-sensitivity is not confined to words with functional roles (traditionally verbs and adjectives), but extends e.g. to nouns (Pustejovsky's Qualia theory, Lecture 2).
- Generative Lexicon spreads the semantic role to all items in the lexicon.

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"

1990-2017 - The Absence of Theory

- Big Data and statistical modeling have 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

Corpus Analysis Toolkit – SketchEngine

SketchEngine | English Web 2013 (enTenTen13) | Elisabet

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

- Clustering
- Save

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

Lemma	Score	Freq
write	0.515	9,209,491
learn	0.467	8,933,221
think	0.455	21,650,907
speak	0.441	4,829,660
talk	0.439	5,399,739
know	0.438	29,229,869
give	0.436	22,289,988
hear	0.434	6,078,870
present	0.429	3,390,837
tell	0.429	11,428,601
say	0.423	39,839,420
understand	0.422	6,215,980
want	0.413	21,635,056
receive	0.410	7,099,321
create	0.410	11,463,028
provide	0.410	16,231,340
see	0.410	29,777,567
need	0.407	21,424,720
like	0.406	8,924,162
make	0.405	51,256,256
show	0.404	10,492,472
mean	0.403	9,118,407
feel	0.403	12,463,798