CIP: Corpus di Italiano Parlato

Struttura e raccolta dati

Caterina Mauri caterina.mauri@unibo.it

Eugenio Goria eugenio.goria@unibo.it

Overview

1. Il progetto LEAdHoC: *Linguistic expression of ad hoc categories*
2. Il corpus - la raccolta dati
3. Corpus design
4. Possibili integrazioni: verso il *CIP* (*Corpus di Italiano Parlato*)
5. Alcuni esempi

LEAdhoC: Linguistic expression of ad hoc categories

SIR 2014 - n. RBSI14IIG0

Coordinatore: *Caterina Mauri*

Assegnisti di ricerca: *Ilaria Fiorentini, Eugenio Goria*

www.leadhoc.org

Inizio progetto: settembre 2015

Fine progetto: settembre 2019

LEAdhoC: Linguistic expression of ad hoc categories

SIR 2014 - n. RBSI14IIG0

Coordinatore: *Caterina Mauri*

Assegnisti di ricerca: *Ilaria Fiorentini, Eugenio Goria*

www.leadhoc.org

Inizio progetto: settembre 2015

Fine progetto: settembre 2019

… cosa sono le categorie ad hoc?

Categorie *ad hoc*: cosa sappiamo?

Categorie *ad hoc*: cosa sappiamo?

Categorie stabili

E.g. parole o sintagmi brevi: **[mobili], [vestiti], [cibo per gatti] [cibo vegetariano]…**

- ✓ Etichette linguistiche brevi e convenzionali
- ✓ Fissate nella memoria a lungo termine
- ✓ Indipendenti dal contesto

Categorie *ad hoc*: cosa sappiamo?

Categorie stabili

E.g. parole o sintagmi brevi: **[mobili], [vestiti], [cibo per gatti] [cibo vegetariano]…**

- ✓ Etichette linguistiche brevi e convenzionali
- ✓ Fissate nella memoria a lungo termine
- ✓ Indipendenti dal contesto

Categorie *ad hoc*

E.g. espressioni complesse: **[attività turistiche da fare a Roma], [cose da mettere in valigia prima di partire per l'Alaska per una vacanza di un mese]**

- ✓ No etichette linguistiche convenzionali
- ✓ Non fissate nella memoria
- ✓ Dipendono dal contesto per essere costruite e interpretate

Categorie *ad hoc*: cosa sappiamo?

Categorie stabili

E.g. parole o sintagmi brevi: **[mobili], [vestiti], [cibo per gatti] [cibo vegetariano]…**

- ✓ Etichette linguistiche brevi e convenzionali
- ✓ Fissate nella memoria a lungo termine
- ✓ Indipendenti dal contesto

Categorie *ad hoc*

E.g. espressioni complesse: **[attività turistiche da fare a Roma], [cose da mettere in valigia prima di partire per l'Alaska per una vacanza di un mese]**

- ✓ No etichette linguistiche convenzionali
- ✓ Non fissate nella memoria
- ✓ Dipendono dal contesto per essere costruite e interpretate

- **Barsalou (1983, 1991, 2003, 2010)**: categorie costruite estemporaneamente, per il raggiungimento di obiettivi specifici – evidenza sperimentale
- **Croft & Cruse (2004):** *construal* di categorie

Categorie *ad hoc*: cosa sappiamo?

- La costruzione di categorie ad hoc è un **processo cognitivo universale e onnipresente, strettamente dipendente dalla abilità linguistica di esprimerlo** – tipicamente tramite la menzione esplicita di *uno o più esemplari della categoria*

Categorie *ad hoc*: cosa sappiamo?

- La costruzione di categorie ad hoc è un **processo cognitivo universale e onnipresente, strettamente dipendente dalla abilità linguistica di esprimerlo** – tipicamente tramite la menzione esplicita di *uno o più esemplari della categoria*

- Barsalou (2010): "although ad hoc categories are ubiquitous in our everyday cognition, they have been subject to **relatively little research, especially as far as their linguistic realizations are concerned**."

Categorie *ad hoc*: cosa sappiamo?

- La costruzione di categorie ad hoc è un **processo cognitivo universale e onnipresente, strettamente dipendente dalla abilità linguistica di esprimerlo** – tipicamente tramite la menzione esplicita di *uno o più esemplari della categoria*

- Barsalou (2010): "although ad hoc categories are ubiquitous in our everyday cognition, they have been subject to **relatively little research, especially as far as their linguistic realizations are concerned**."

✧ **Questo progetto intende colmare tale lacuna,**
- **prendendo in esame un ampio campione di lingue**
- **analizzando come la costruzione di categorie *ad hoc* viene realizzata nel discorso**

Linee di ricerca e obiettivi

Linee di ricerca e obiettivi (1)

1) Quali sono i tipi di costruzioni attestate nelle lingue del mondo per esprimere le categorie ad hoc? Lingue diverse si comportano diversamente!

Derivazione (e non solo)
It. *bambin-***ame**
bambini **annessi&connessi**
bambini **e via dicendo**
Bambini, giochi, urla **eccetera**

Plurali speciali
1) Dogon (Niger-Congo)
isu **mbe**
pesce **PL**
'pesce e cose così'

Connettivi
2) Giap. *[Biiru-***ya** *sake-o]*
Birra, saké ecc.

Raddoppiamento
3) Turco *bira* **m***ira*

Linee di ricerca e obiettivi (1) e (2)

1) Quali sono i tipi di costruzioni attestate nelle lingue del mondo per esprimere le categorie ad hoc? Lingue diverse si comportano diversamente!

Derivazione (e non solo)
It. *bambin-***ame**
bambini **annessi&connessi**
bambini **e via dicendo**
Bambini, giochi, urla **eccetera**

Plurali speciali
1) Dogon (Niger-Congo)
isu **mbe**
pesce **PL**
'pesce e cose così'

Connettivi
2) Giap. *[Biiru-***ya** *sake-o]*
Birra, saké ecc.

Raddoppiamento
3) Turco *bira* **m***ira*

2) Qual è l'origine di queste strutture? Quali risorse grammaticali sono più frequentemente impiegate per veicolare questa funzione?

Linee di ricerca e obiettivi (1) e (2)

- **Linea di ricerca 1):**

INDAGINE TIPOLOGICA- analisi della variazione interlinguistica attestata, sulla base di un campione tipologico di 150 lingue, per identificare *pattern* di variazione e tendenze universali

- **Linea di ricerca 2):**

ANALISI DIACRONICA- identificazione delle sorgenti diacroniche e dei percorsi di mutamento che conducono allo sviluppo di strategie per l'espressione di categorie ad hoc.

Mauri, C. & Sansò, A. Forthcoming. Special issue di *Folia Linguistica Diachronica* su "Linguistic strategies for the construction of ad hoc categories: synchronic and diachronic perspectives"

Mauri, C. 2017. Building and interpreting ad hoc categories: a linguistic analysis". In J. Blochowiak, C. Grisot, S. Durrleman-Tame and C. Laenzlinger (eds.) *Formal models in the study of language,* Berlin: Springer

Linee di ricerca e obiettivi (3)

Linee di ricerca e obiettivi (3)

3) Quando e **perché** questo processo di categorizzazione viene attivato nella conversazione spontanea? In base a quali proprietà i parlanti selezionano alcuni esemplari come più funzionali di altri per risalire alla categoria ad hoc?

- **Linea di ricerca 3):**

ANALISI DI CORPORA – valutazione delle proprietà testuali e pragmatiche delle categorie ad hoc nel discorso, per comprendere gli scopi del parlante nel costruirle e il loro ruolo nella gestione del topic discorsivo

Linee di ricerca e obiettivi (3)

3) Quando e **perché** questo processo di categorizzazione viene attivato nella conversazione spontanea? In base a quali proprietà i parlanti selezionano alcuni esemplari come più funzionali di altri per risalire alla categoria ad hoc?

- **Linea di ricerca 3):**

ANALISI DI CORPORA – valutazione delle proprietà testuali e pragmatiche delle categorie ad hoc nel discorso, per comprendere gli scopi del parlante nel costruirle e il loro ruolo nella gestione del topic discorsivo

METODOLOGIA:

- ✓ **Costruzione e analisi di un corpus di Italiano parlato**, per:
  - Individuare tutte le costruzioni attestate nella variazione intra-linguistica
  - Monitorare i contesti d'uso, le funzioni pragmatiche e discorsive

Linee di ricerca e obiettivi (3)

3) Quando e **perché** questo processo di categorizzazione viene attivato nella conversazione spontanea? In base a quali proprietà i parlanti selezionano alcuni esemplari come più funzionali di altri per risalire alla categoria ad hoc?

- **Linea di ricerca 3):**

ANALISI DI CORPORA – valutazione delle proprietà testuali e pragmatiche delle categorie ad hoc nel discorso, per comprendere gli scopi del parlante nel costruirle e il loro ruolo nella gestione del topic discorsivo

METODOLOGIA:

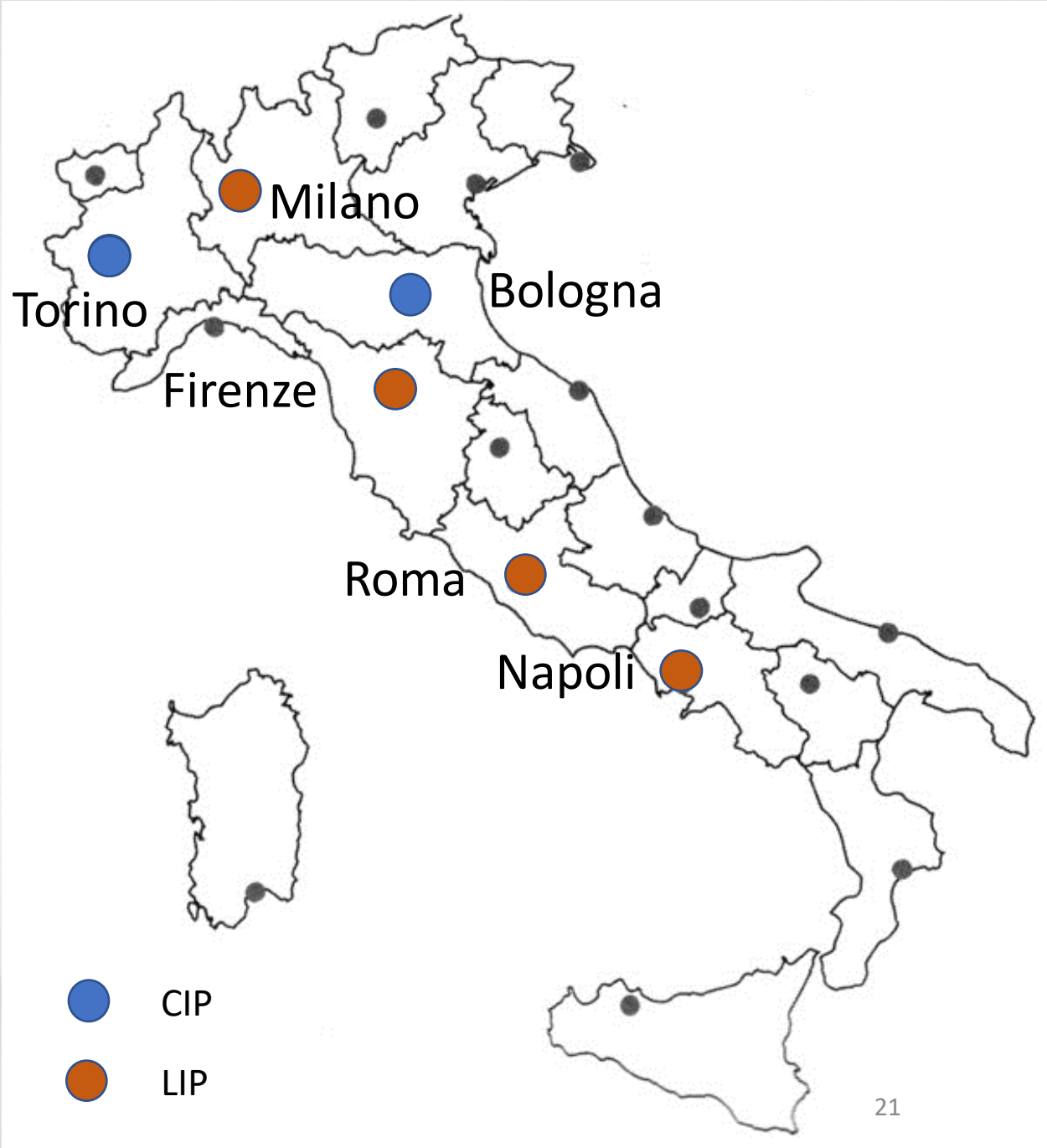
- ✓ **Costruzione e analisi di un corpus di Italiano parlato**, per:
  - Individuare tutte le costruzioni attestate nella variazione intra-linguistica
  - Monitorare i contesti d'uso, le funzioni pragmatiche e discorsive

Corpus design

La costruzione della base di dati

Punti di indagine

- Torino e Bologna
- Presenza di standard regionali (Cerruti, Regis 2014)
- Punti non considerati nel LIP
- Italiano parlato all'interno dell'università



Caratteristiche del corpus

- 35 ore per città (70 totali)
- Circa 700 mila parole (stimate)
- Adozione di uno standard per la gestione dei dati e per la raccolta dei metadati
  - Liberatoria
  - Metadati
  - Uso dell'intermediario («friend of a friend», Tagliamonte 2006)
- Minimizzazione della variabile diastratica: italiano di parlanti colti
- Massimizzazione della variabile diafasica: vari tipi di contesto interazionale

Gestione dei dati

- Possibilità di «scissione» della liberatoria
  - Consenso all'utilizzo dei dati
  - Consenso alla diffusione
- Adozione di un sistema di metadati
  - Età
  - Luogo di nascita
  - Luogo scuole superiori
  - Eventuali altre occupazioni
  - Lingue straniere

Caratteristiche degli informanti

- 2 profili principali
  - Docenti universitari
  - Studenti universitari
- Caratteristiche diastratiche comuni
  - Titolo di studio come principale indicatore di classe sociale
  - Sesso/genere e età come ulteriori variabili diastratiche
  - Possibilità di studi in diacronia apparente (Sankoff 2006)
- Soggetti ad alta mobilità
  - Dinamiche di contatto orizzontale, italiano composito (Canepari 1983, Cerruti 2009, 2013)

Le situazioni osservate

	SITUAZIONE / TASK	DURATA (lordo)	NUMERO	TOTALE	RICERCATORE	RELAZIONE PARLANTI
TIPO A – scambio bidirezionale con presa di parola libera	1. Ricevimento Studenti	30min	6	180	Assente	Asimmetrica
	2. Focus Group	90 min	5	450	Intermediario	Simmetrica
	3. Conversazione libera	30 min	6	180	Assente	Simmetrica
TIPO C – scambio bidirezionale con presa di parola **non** libera	1. Esami universitari	60 min	3	180	Assente	Asimmetrica
TIPO D – scambio unidirezionale in presenza del destinatario	1. Lezioni universitarie	90 min	8	720	Assente	Asimmetrica
	2. Storytelling / intervista semistrutturata	30 min	13	390	Intermediario	Simmetrica

Avanzamento del lavoro

		TORINO		BOLOGNA	
A1	**6**	0	00:00:00	6	02:00:37
A2	**5**	2	01:02:40	0	00:00:00
A3	**6**	0	00:00:00	2	01:26:39
C1	**3**	7	03:17:46	0	00:00:00
D1	**8**	9	09:38:10	1	01:36:24
D2	**13**	0	00:00:00	2	00:32:34
TOTALE			13:58:36		05:36:14

Febbraio 2017
- 20 ore non trascritte

Aprile 2017
- 40-50 ore totali
- Inizio trascrizioni: dottorandi, collaboratori esterni, assegnisti

Settembre 2017
- Completamento della raccolta dati
- Elaborazione di un'interfaccia per la consultazione online
- Inserimento graduale dei materiali allineati con l'audio

???
- POS *tagging* automatico

Verso il Corpus di Italiano Parlato

- Da un corpus di progetto alla creazione di una risorsa condivisa
- possibili direzioni:
  1. Ripetizione della raccolta dati in altri punti di inchiesta utilizzando la stessa metodologia
     - Università dell'Insubria (Andrea Sansò)
  2. Applicazione delle stesse categorie di analisi ad altri contesti di indagine
     - Italiano popolare parlato dei semi-colti
  3. Adattamento di registrazioni già esistenti alla griglia in uso
     - Università di Salerno (Miriam Voghera)

Alcuni esempi

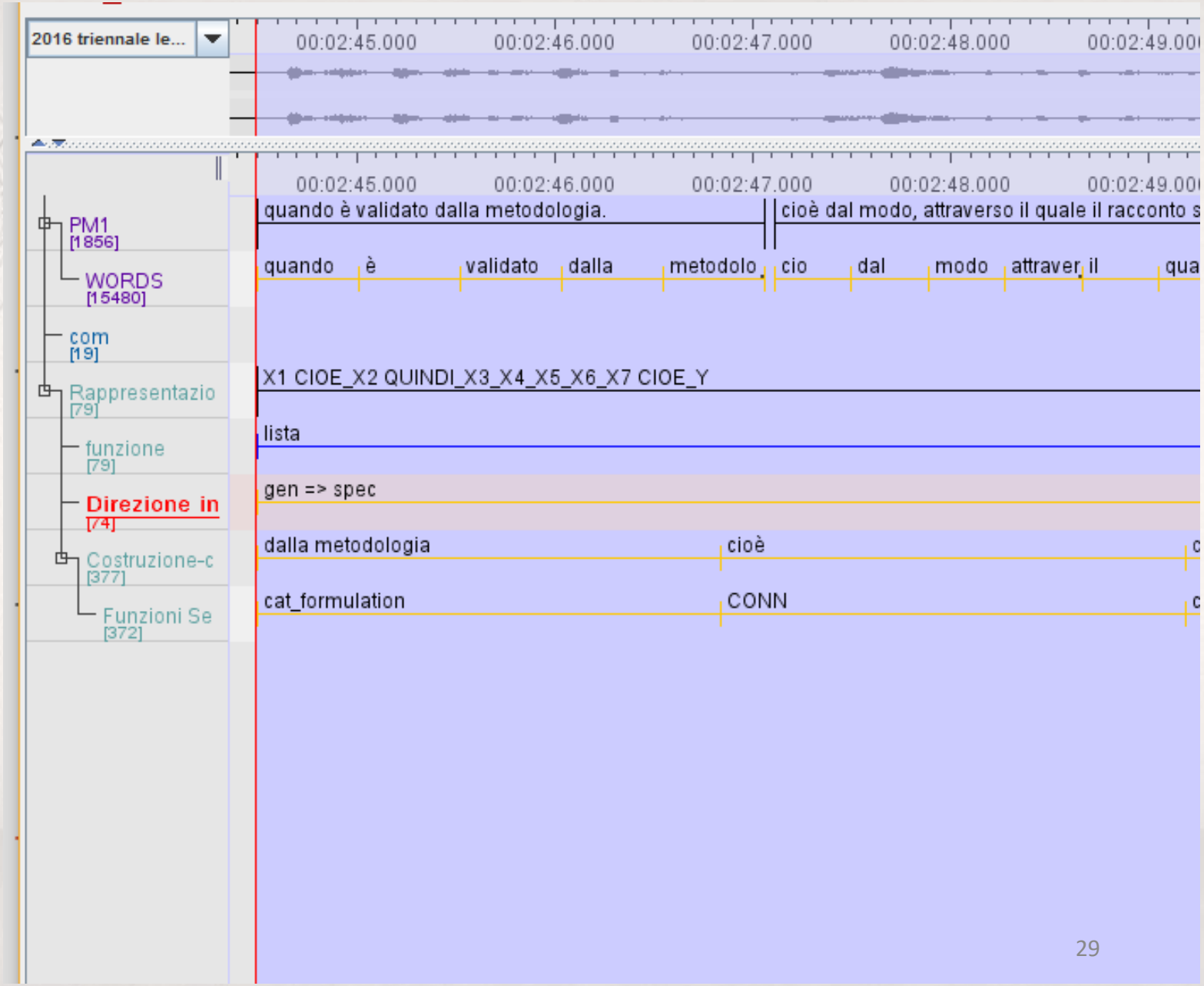
allineamento con l’audio

trascrizione

tokenizzazione

etichettatura manuale delle costruzioni

annotazione per funzioni



Annotazione e analisi dei dati

- Individuazione degli elementi e delle costruzioni che fungono da trigger del processo di categorizzazione

> Esistono altre *x* caratterizzate dalla proprietà P

  - General extenders (Channel 1994, Ovestreet 1999)
  - Esemplificazione (Barotto 2016)
  - Costruzioni similative
  - Morfologia derivazionale
  - Liste non esaustive (Jefferson 1996, Selting 2007, Bonvino et al. 2009)
  - Connettivi non esaustivi (Mauri 2014)
- Direzione dell'astrazione
  - Generico > Specifico (top down)
  - Specifico > Generico (bottom up)

Esempio 1

```
ma ci sono anche documenti che provengono direttamente da quel passato più lontano.
ad esempio (.) i dati dell'archeologi::a,
le iscrizioni di cui:: parlere:mo,
la numismatica le::: mh:: le monete,
```

Formulazione della categoria *ad hoc*

documenti che provengono direttamente da quel passato più lontano

trigger (esemplificatore)

ad esempio

lista

Es.1 i dati dell' archeologia

Es.2 le iscrizioni di cui parleremo

Es.3 la numismatica

le monete

Esempio 2

eh ma c'è uno storico che dialoga con le fonti,
e parla (.) con (.) e:h il proprio lettore i propri ascoltatori insomma le persone (.) i destinatari della comunicazione (.) storica.

ma c'è uno storico che dialoga con le fonti

e parla

con

il proprio lettore

i propri ascoltatori

Es.1

Es.2

lista

insomma

le persone

i destinatari della comunicazione storica

trigger

Formulazione della categoria *ad hoc*

Conclusioni

Obiettivo a breve termine

- Ottenere una panoramica sulle principali strategie che in italiano permettono di creare categorie nel discorso
- Valutare eventuali dinamiche di variazione in rapporto con i parametri considerati all'interno del *corpus*

Obiettivo a medio termine

- Creazione di una risorsa di libero accesso per lo studio dell'italiano di parlanti colti

Obiettivo a lungo termine

- Integrazione della risorsa con altri dati mediante il coordinamento con altri ricercatori

GRAZIE