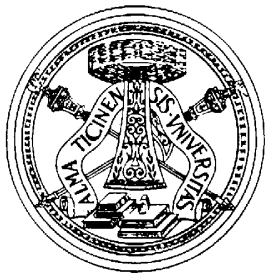
Using TPAS for Linguistic Studies of Italian Verbs

Elisabetta Jezek

Dipartimento di Studi Umanistici
Università di Pavia

PHRASALEX: Phraseological Approaches to Learner's Lexicography
Università di Modena e Reggio-Emilia
19-20 September 2019



Outline of the talk

- Why a semantic resource of predicate-argument structures for linguistic analysis and NLP?
- Features of the resource.
- Why is T-PAS useful? Coercions, semantic clustering, and verb classes.
- Concluding observations and ongoing work.

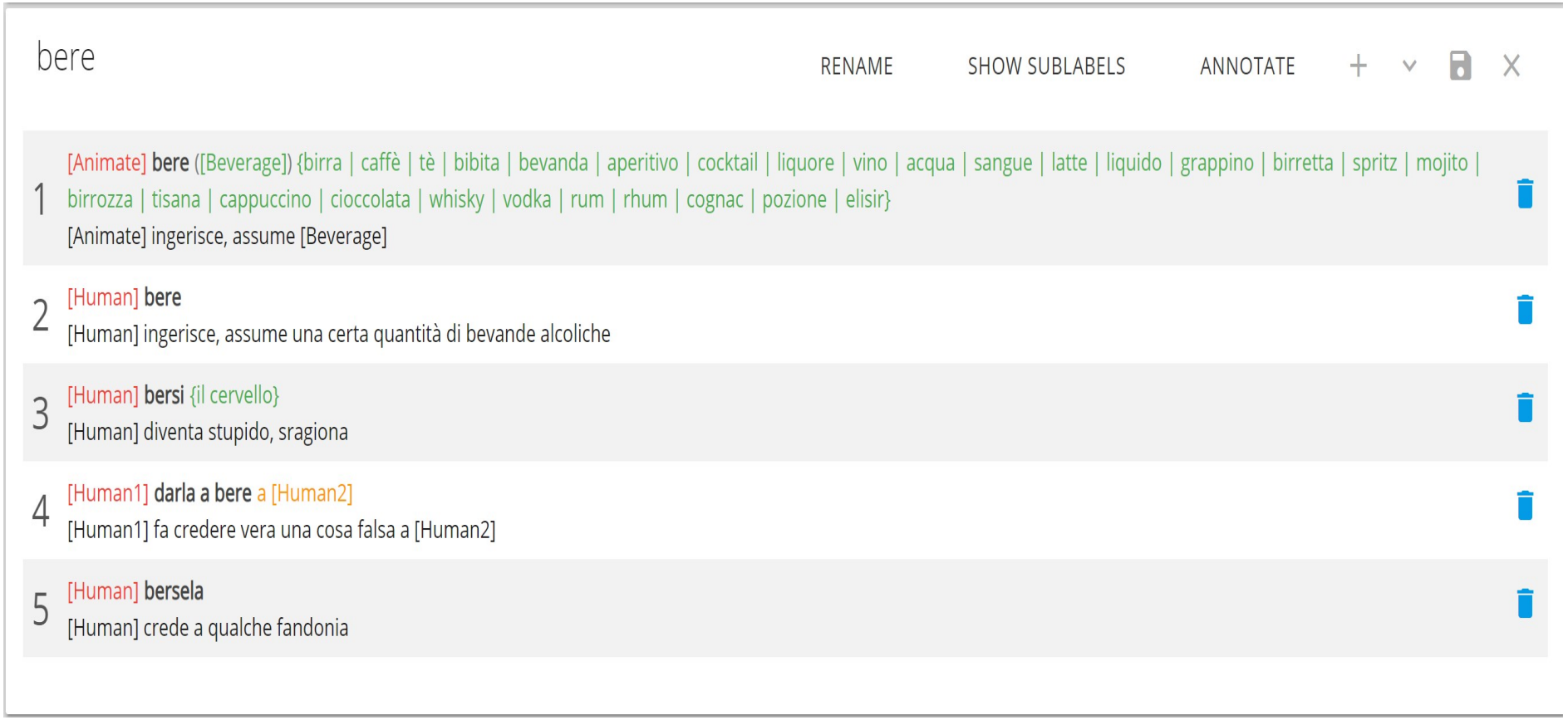
Why a semantic resource

- Several resources containing predicate-argument structures or classes for verbs focus on - or start from - the **subcategorization** (syntactic) properties, and $\rightarrow$ then - if so - move on to include semantics.
- We claim that the complex interplay between syntax and semantics in verbs **cannot be fully unraveled through this methodology alone**.

Methodology for T-PAS

- In T-PAS we propose to start from the semantic properties, particularly the **semantic types** that characterize each argument position of verbs.
- Our theoretical stance is that the semantics of the argument(s) **co-determines the meaning of the verb**.
- We collect **semantic structures** from corpora by generalizing over contexts of use, where the verb meaning is disambiguated.
- A Typed Predicate Argument Structure (T-pas) is a corpus-derived predicate argument structure with the specification of the expected **semantic type** for each argument position.

T-pass for *bere* ‘to drink’



T-PAS first version

- Inventory of **4241 Typed Predicate Argument Structures (T-pass)** for **1000 average polysemy** Italian verbs.
- Random extraction of 1000 verbs from **"lemmi ad alta disponibilità"** of Sabatini Coletti 2008:
- 10% 2-sense verbs;
- 60% 3-5-sense verbs;
- 30% 6-11-sense verbs.

T-PAS first version

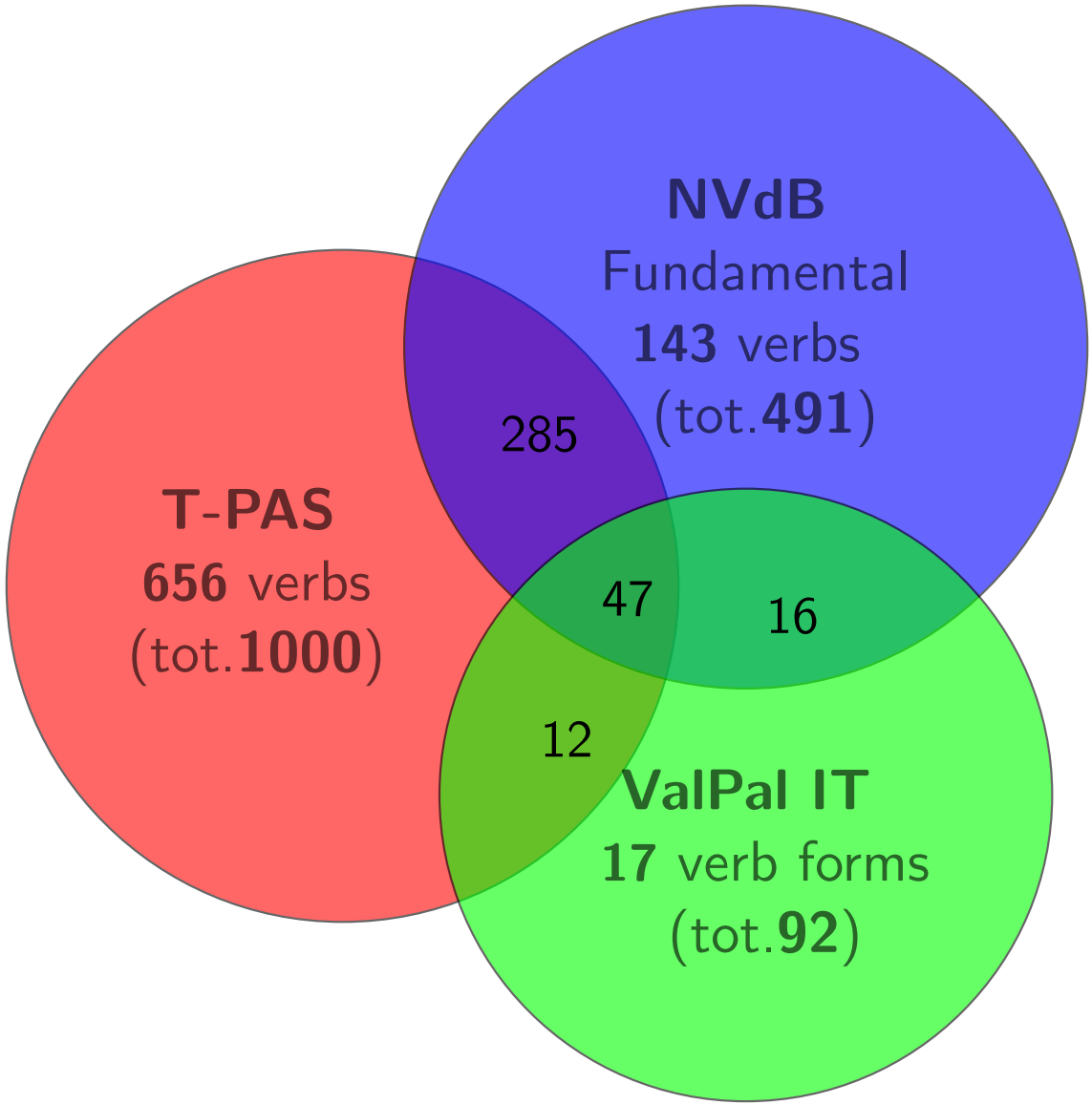
- Developed at the Dept. of Humanities and the Language Centre of the University of Pavia with the technical support of Lexical Computing / Sketch Engine, Brno (CZ); in collaboration with Fondazione Bruno Kessler (FBK), Trento.
- ITWaC corpus, obtained by crawling texts from the Italian domain in the web using medium frequency vocabulary as seeds (1,585,620,279 tokens, Baroni et al., 2009).
- Reduced version for TPAS: 935,698,409 tokens, 751,542,948 words, 33,659,325 sentences.
- For linguistic research, semantic processing tasks, and possibly pedagogical purposes.

- Jezek, Magnini, Feltracco, Bianchini, Popescu. 2014. *T-PAS: A resource of Typed Predicate Argument Structures for linguistic analysis and semantic processing.* LREC 2014.

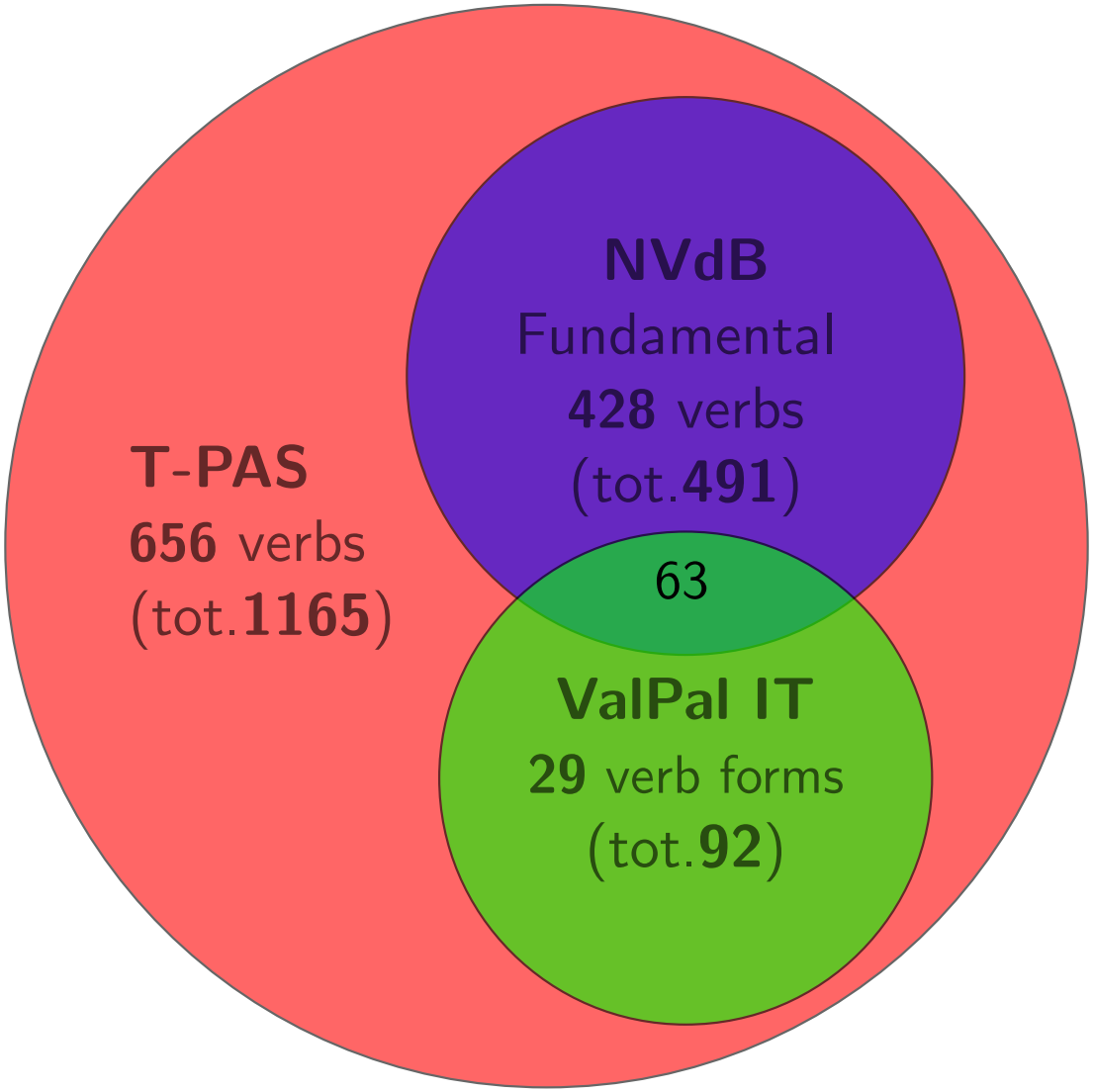
T-PAS: enrichment

- more verbs!
- → Cover the remaining **fundamental verb lemmas** ("verbi fondamentali") of Nuovo Vocabolario di Base (De Mauro 2016).
- "**Fundamental lemmas** are the words that in all languages tend to cover on average about 90 percent of the occurrences of words in texts and discourse."
- → cover ValPal IT (Italian Valency Patterns Leipzig Online Database, Cennamo and Fabrizio 2013) - part of the PRIN project "Argument Structure in Flux".

T-PAS version 1



The Enrichment



Let's have a look

verbs	T-PAS	T-PAS Enriched	number of t-pas	Polysemy Sab.Col.	NVDB transitivity	class (Jezek 2003)	VALPAL IT coding frame
abbandonare		x	2	high	tr.	10	
abbassare	x	x	8	high	tr.	10	
abbracciare		x	4	high	tr.	10	1>V.subj[1]>2
abitare	x	x	3	mid	intr. - tr.	9	1>V.subj[1] (>LOC2)
abituare		x	5	low	tr.	10	
accadere	x	x	1	low	intr.	3	
accendere		x	6	high	tr.	10	
accettare	x	x	6	high	tr.	10	
accogliere	x	x	5	high	tr.	1	
accompagnare		x	7	high	tr.	10	
accorgersi		x	1	low	pronom.intr.	4	
accusare	x	x	6	high	tr.	10	
acquistare	x	x	4	high	tr. - intr.	9	
addossare		x	3	high	tr.	10	
adorare	x	x	3	mid	tr.	1	
adottare	x	x	3	mid	tr.	1	
affermare	x	x	4	mid	tr.	10	
affidare	x	x	2	mid	tr.	10	
affrontare	x	x	4	high	tr.	10	

Top level of the T-PAS System of Semantic Types

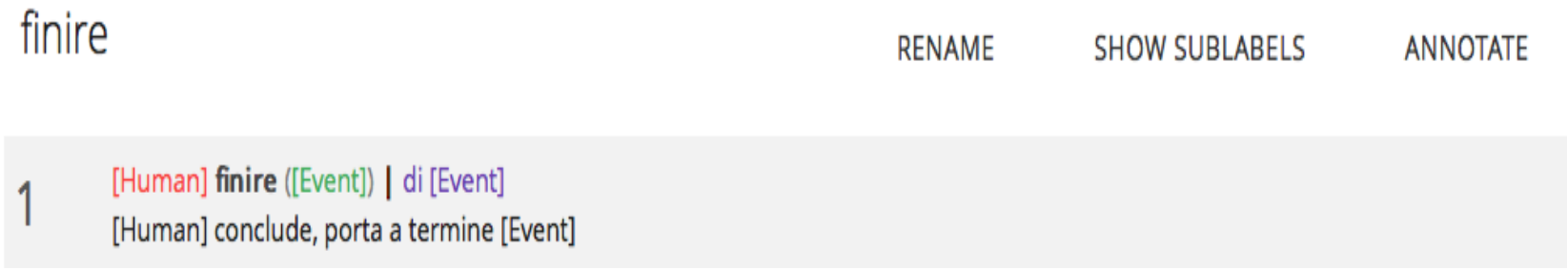
with a selection of leaf types

- ANYTHING
  - ENTITY
    - PHYSICAL ENTITY
      - INANIMATE
        - ARTIFACT
        - STUFF
        - LIGHT SOURCE [LOCATION, INANIMATE]
      - ANIMATE
        - HUMAN
        - HUMAN GROUP
        - ANIMAL
        - ANIMAL GROUP
      - BODY
      - PART OF BODY
      - PLANT
      - LOCATION
    - ABSTRACT ENTITY
      - INSTITUTION [ABSTRACT ENTITY, HUMAN GROUP]
      - INFORMATION SOURCE
        - DOCUMENT [ARTIFACT, INFORMATION SOURCE]
      - ……
  - EVENTUALITY
    - EVENT
    - STATE
  - PROPERTY
    - COLOUR
    - ROLE
    - WEIGHT
    - ……

Jezek E. 2019. *Sweetening Ontology Cont'd*. Joint Ontology Workshop

Syntax and Semantics in T-PAS

T-PASs are semantically determined. Different syntactic realizations are encoded as alternating subcategorizations within the same T-PAS.



- Finisce l'allenamento, la gente l'applaude.
  'He finishes the training, the people applaud him'
- Non faccio in tempo a finire di bere la mia birra che viene un mio amico.
  'Before I make it to finish my beer, a friend arrives.'

Theoretical and Methodological Tools

- Corpus Pattern Analysis procedure (CPA), Hanks 2004, 2013.
- Semantic **compositional mechanisms** active in V-ARG combination in addition to simple composition: **co-composition** and **type-coercion** (Pustejovsky 1995).

Reference points for the T-PAS project

- LexIt (Lenci et al. 2012), a resource providing automatically acquired distributional information about Italian verbs, adjectives and nouns.
- Differently from T-PAS, LexIt does not convey an inventory of patterns and the categories used for classifying the semantics of arguments are not corpus-driven.
- Inventory of senses such as MultiWordNet (Pianta et al. 2002), Senso Comune (Oltramari et al. 2013) and Verb Atlas (Navigli et al 2019) are resources to which T-PAS can be successfully linked with the goal of populating the former with corpus-driven pattern-based sense distinctions for verbs.

Semantic Type mismatches

- In a number of studies we have shown that the paradigmatic sets of words that populate specific argument slots within the same verb sense **do not map neatly onto the "expected" type** (selected by V) (Pustejovsky and Jezek 2008).
- **Mismatches** between **pattern** type (expected by V) and **instance type** (inherent in N) within the same grammatical relation.
- The phenomenon is pervasive (Jezek and Quochi, 2010, Pustejovsky et al. SemEval Task 7, 2010) and - we discovered - **spread over specific Levin's 1993 and VerbNet classes**
- Aspectual verbs, communication verbs, perception verbs, directed motion verbs, verbs of motion using a vehicle.

Type mismatch: Aspectual verbs

[[Human]-subj] interrompe [[Event]-obj]

- Arriva Mirko e interrompe la conversazione.
  'Mirko arrives and interrupts the conversation' (matching)
- Il presidente interrompe l'oratore.
  'The president interrupts the speaker' (HUMAN as EVENT)

Type mismatch: Communication verbs

[[Human]-subj] annuncia [[Event]-obj]

- Lo speaker annuncia la partenza.
  'The speaker announces the departure' (matching)
- Il maggiordomo annuncia gli invitati.
  'The butler announces the guests' (HUMAN as EVENT)
- L'altoparlante annunciava l'arrivo del treno.
  'The loudspeaker announces the arrival of the train'
  (ARTIFACT as HUMAN)
- Una telefonata anonima avvisa la polizia.
  'An anonymous telephone call alerted the police' (EVENT as HUMAN)

Type mismatch: Directed motion verbs

[[Human]-subj] raggiunge [[Location]-obj]

- Abbiamo raggiunto l'isola alle 5.
  'We reached the island at 5' (matching)
- Ho raggiunto il semaforo e ho svoltato a destra.
  'I reached the traffic light and turned right' (ARTIFACT as LOCATION)

Type mismatch: Perception Verbs

[[Human]-subj] ascolta [[Sound]-obj]

- Rilassarsi ascoltando il rumore della pioggia.
  'Relax while listening to the sound of rain' (matching)
- Ascoltava la radio con la cuffia.
  'He listened to the radio with his earphones' (ARTIFACT as SOUND)
- Rimasi a lungo ad ascoltare il suo respiro.
  'I stayed for a long while listening to his breath' (EVENT as SOUND)
- Non ho potuto ascoltare tutti i colleghi
  'I could not listen to all colleagues' (HUMAN as SOUND)

Sounds

sul suo divano e, attraverso il soffitto,	ascoltare	1	il rumore di passi dei bambini e la musica dei
; con oltre 75 minuti di durata, il cd permette di	ascoltare	1	la voce degli interpreti presenti al Comunale
prima per la sua costituzione fraudolenta . </s><s>	Ascoltò		la mia lunga arringa , la mia paziente analisi
ılla sua moto ferma sul marciapiede . </s><s> Mangia e	ascolta		la musica attraverso un walkman . </s><s> CENTRALE
scorrere il quadrante sul righello potrà	ascoltare	1	il rumore degli attrezzi per giocare ad imitare
su Giorgio Perlasca " Il silenzio del Giusto " ,	ascolterà		la testimonianza di Franco Per lasca curerà la
: si piazzò accanto ai binari della ferrovia, e	ascoltò	1	il suono emesso da un vagone pieno di musicisti,
oratoria , E infine , quasi tutti , come vuoi che	ascoltino		lo stormire delle foglie o il gorgoglio del
, la sua ironia , il suo sense of humour . </s><s> Se	ascolti		molti degli album hip hop che escono oggi
ederli dal vivo per rendersene conto . </s><s> O basta	ascoltare		questo loro nuovo disco . </s><s> Il titolo fa pensare ad
innovativi » . </s><s> In precedenza l' assemblea aveva	ascoltato		l' intervento di Rocco Giordano , coordinatore
che di questioni / fatti politici , ma mi piace	ascoltare		gli altri che parlano di politica Pertanto se
, Fidenza e San Secondo avranno l' occasione di	ascoltare		padre Alex Zanotelli , ex direttore della
</s><s> Intanto l' assessorato dal 4 marzo è pronto ad	ascoltare		con molta attenzione il dibattito dal quale ci
ovare . </s><s> Se io fossi una persona che non ha mai	ascoltato		Patty Smith , e vedessi un moderatore che dice "
sua Autobiografia che ad un tratto , mentre	ascoltava		la Messa e pregava per la mamma , una misteriosa
la ripetizione chi non parla ha il compito di	ascoltare		chi sta parlando , aiutarlo se si trova in
di perdermi ... </s><s> Mi basta chiudere gli occhi e	ascoltarlo		cantare , per cadere giù in fondo ai miei
che provi mentre leggi un grande libro , o mentre	ascolti		un album fantastico o quando guardi un film
ın languido ballo cheek to cheek . </s><s> Ti metti ad	ascoltare	1	la sua vellutata voce da crooner e percepisci
CD 139 L' autoradio VDO Dayton CD 139 permette di	ascoltare		CD / CD - RW e MP 3 con una qualità sonora
delle "fan" che rendevano impossibile	ascoltare	1	il suono che producevano. </s><s> Erano inoltre
vederli durante le pause pubblicitarie mentre	ascoltano		alcune persone del pubblico : quando la

Interpreted as Sounds

sul suo divano e, attraverso il soffitto,	ascoltare	1	il rumore di passi dei bambini e la musica dei
; con oltre 75 minuti di durata, il cd permette di	ascoltare	1	la voce degli interpreti presenti al Comunale
prima per la sua costituzione fraudolenta . </s><s>	Ascoltò	1.a	la mia lunga arringa , la mia paziente analisi
ılla sua moto ferma sul marciapiede . </s><s> Mangia e	ascolta	1.a	la musica attraverso un walkman . </s><s> CENTRALE
scorrere il quadrante sul righello potrà	ascoltare	1	il rumore degli attrezzi per giocare ad imitare
su Giorgio Perlasca " Il silenzio del Giusto " ,	ascolterà	1.a	la testimonianza di Franco Per lasca curerà la
: si piazzò accanto ai binari della ferrovia, e	ascoltò	1	il suono emesso da un vagone pieno di musicisti,
oratoria , E infine , quasi tutti , come vuoi che	ascoltino	1.a	lo stormire delle foglie o il gorgoglio del
, la sua ironia , il suo sense of humour . </s><s> Se	ascolti	1.a	molti degli album hip hop che escono oggi
ederli dal vivo per rendersene conto . </s><s> O basta	ascoltare	1.a	questo loro nuovo disco . </s><s> Il titolo fa pensare ad
innovativi » . </s><s> In precedenza l' assemblea aveva	ascoltato	1.a	l' intervento di Rocco Giordano , coordinatore
che di questioni / fatti politici , ma mi piace	ascoltare	1.a	gli altri che parlano di politica Pertanto se
, Fidenza e San Secondo avranno l' occasione di	ascoltare	1.a	padre Alex Zanotelli , ex direttore della
</s><s> Intanto l' assessorato dal 4 marzo è pronto ad	ascoltare	1.a	con molta attenzione il dibattito dal quale ci
ovare . </s><s> Se io fossi una persona che non ha mai	ascoltato	1.a	Patty Smith , e vedessi un moderatore che dice "
sua Autobiografia che ad un tratto , mentre	ascoltava	1.a	la Messa e pregava per la mamma , una misteriosa
la ripetizione chi non parla ha il compito di	ascoltare	1.a	chi sta parlando , aiutarlo se si trova in
di perdermi ... </s><s> Mi basta chiudere gli occhi e	ascoltarlo	1.a	cantare , per cadere giù in fondo ai miei
che provi mentre leggi un grande libro , o mentre	ascolti	1.a	un album fantastico o quando guardi un film
un languido ballo cheek to cheek . </s><s> Ti metti ad	ascoltare	1.a	la sua vellutata voce da crooner e percepisci
CD 139 L' autoradio VDO Dayton CD 139 permette di	ascoltare	1.a	CD / CD - RW e MP 3 con una qualità sonora
delle "fan" che rendevano impossibile	ascoltare	1	il suono che producevano. </s><s> Erano inoltre
vederli durante le pause pubblicitarie mentre	ascoltano	1.a	alcune persone del pubblico : quando la

Semantic Types and Lexical Sets

Ponti, Jezek, Magnini 2016 D-SALT ESSLLI 2016, Clic-It 2016; Jezek, Ponti, Magnini 2019 ISA-15

- In a different study we used distributional techniques to verify whether the fillers of the argument positions constitute **homogeneous groupings from a semantic point of view**.

Case study

We focus on a list of 21 Italian **Causative/Incoative** verbs, inspired by Haspelmath et al 2014's scale of spontaneous occurrence; see also Samardžić and Merlo 2012): *rompere* (break), *riempire* (fill), *seccare* (dry), *affondare* (sink), *finire* (finish), *suonare* (ring, play) ...

- Un uomo ha suonato {il campanello | la campana.}
  A man rang {the door bell | the bell.}
- {Il campanello | la campana } ha suonato.
  {The door bell | the bell } rang.

A single verb in the lexicon that is polysemous between a sense that **encodes the cause**, and a sense that **lexicalizes the result** (Levin 1993, Jezek 2003).

Motivation

- We focus on this verb class because one can expect that verb alternations **preserve** semantic selection.
- Thus alternations provide an opportunity to test to what extent semantic selection can be characterized by distributional semantics, **S/O alternation** in particular.

Harvest fillers

- We harvest the list of fillers for the argument positions of the target verbs from the freely available ItWaC corpus.
- We focus on Subject of transitive (A), **Subject of intransitive (S)**, **Object (0)**.
- A sample of about 2,000,000 sentences was extracted, enriched through the Mate-tools parser and filtered by max sentences length 99 tokens.
- We obtained a database of fillers structured as a list for the different argument positions (A, S, O) of each verb.

Fillers as vectors

- We then transform each of the collected fillers into **vectors** into vectors using the Word2Vec Continuous Bag of Words (CBOW) Distributional Semantic Model (Mikolov et al, 2013).
- This results into a vector model of **fillers only**.
- DSMs are based on the assumption that the meaning of words can be inferred by the neighbouring words (**distributional hypothesis**, Harris 1954, Firth 1957); they represent the similarity of meaning between words as geometric distance in multi-dimensional **vector spaces**.

Vector Quantization

- Vectors are then fed to an algorithm (*k-means*) to obtain **clusters**.
- **Vector quantization** is the operation of defining k clusters in a model and then assessing vector membership - abiding the rule that every vector is assigned to one and only one cluster.
- The best trade-off is reached by **minimising cluster numbers** and **maximizing their internal similarity**.

Vector Quantization

- We obtain 377 clusters using the Elbow method.
- Each cluster is identified by a number, and nothing else.
- In order to characterize them semantically, for each cluster we identify the filler with the vector closest to the cluster **centroid** (**the prototype**), and manually assign a tentative **semantic label**.

A Proof of Concept

cluster	prototype	closeness	label
24	storie	0.708	stories
58	album	0.789	discs
143	clarinetto	0.736	instruments
157	brontolio	0.807	sounds
193	rockabilly	0.758	performers
257	palermo	0.686	towns
278	cornamuse	0.726	instruments
285	aggettivo	0.736	language
325	scemenza	0.801	judgements

Table 1: 9 clusters that are relevant for the It. verb *suonare* 'to ring (a bell)', 'to play (an instrument)'. Cluster centroids (prototypes), closeness (cosine distance) and manually assigned semantic labels.

List of fillers of suonare_object in cluster **157**

157 chiasso muto grida tuoni tuono brontolio folla cicale agitarsi fischi strepiti baccano clacson fracasso rintocchi fruscio grido frastuono eco silenzio brusio sbadiglio campanello ululati ululato sordo urlando campanacci orecchio gemito gemiti andirivieni sommesso ronzio timpani ruggito ruggiti squillo fragore fragori gorgoglio mormorio rumore rumori urla urlo rantolo stridore udito miagolio tonfo tumulto spari rincorrersi voce sospiri detonazioni singhiozzo campane fischio risa udiva muezzin sirena sirene lamento lamenti tam battito battiti

Relevant clusters for *suonare* and their intersections

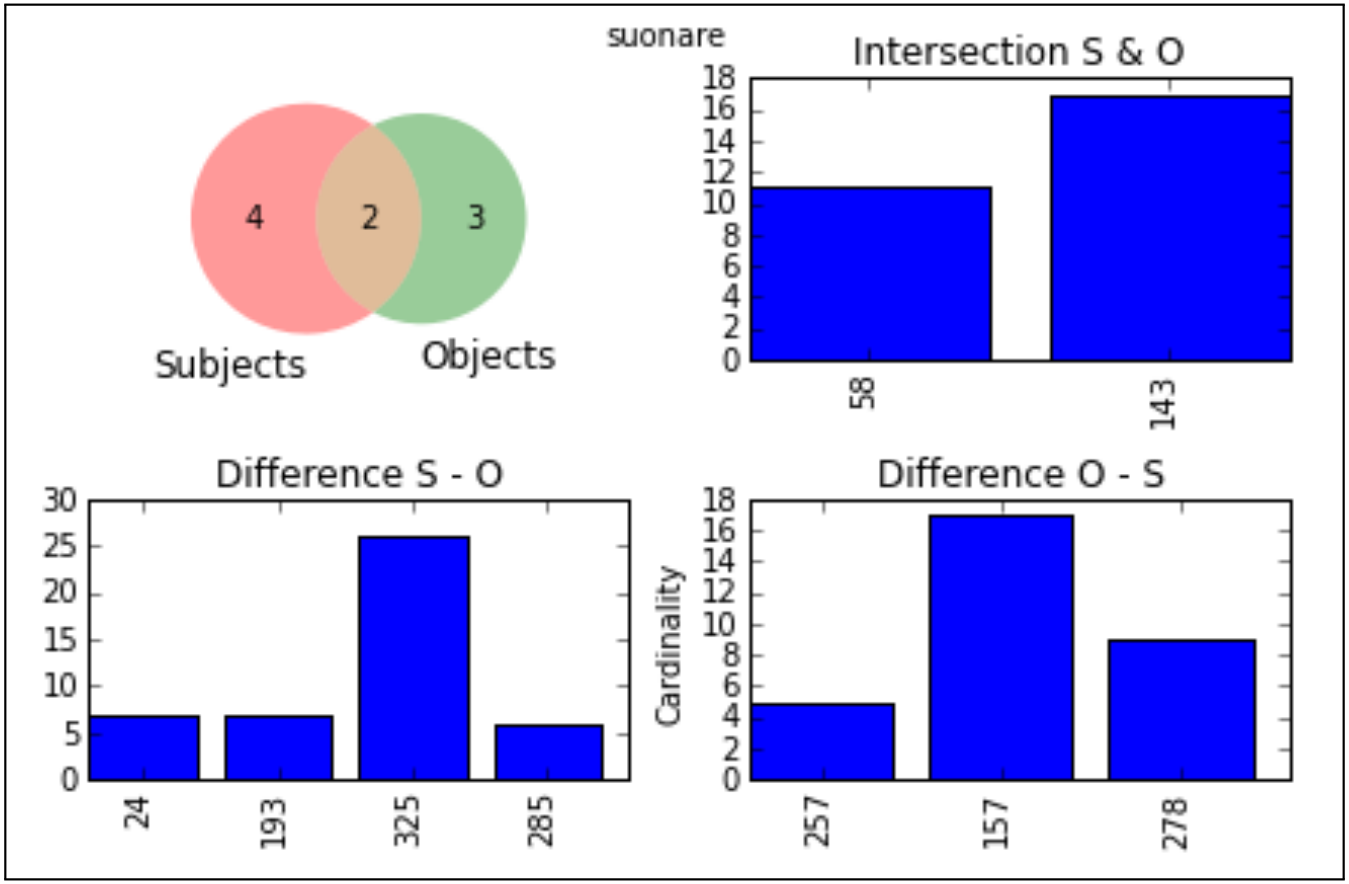


Figure 1: Intersection of fillers (S and O), Subject only (Difference S-O) and Object only (O-S) for the 9 relevant clusters of *suonare*. The three blue bar plots represent the number of fillers of the verb that are members of the cluster.

Qualitative analysis

- The intersections between S and O is dominated by cluster **143** (*Musical Instruments*) and cluster **58** (*Discs*).
- In Subject only, we find cluster **24** (*stories*) and **285** (*language*) related to the meaning 'to seem':
- "Questa storia (mi) suona strana"
  'this story sounds strange (to me)'.
- In Subject only, we also register cluster **193** (*performers*), due to the monoargumental use:
- "Mentre il chitarrista suonava..."
  'while the guitarist was playing...'.

Qualitative analysis

- In Object only (O-S) we find cluster **157** *Sounds*; however, most of its members denote sound-producing artefacts (*campanello* 'doorbell', *clacson* 'car horn').
- Cluster **278** *Instruments* has as members several words denoting musical compositions, such as *cantilena* 'singsong', *litanie* 'litanies', *ritornello* 'refrain' and *canzonetta* 'jingle'.
- → The **centroid** appears not to be **the best indicator** of the semantics of the cluster.
- Cluster **257** (*towns* is due to a systematic parsing error (object instead of locative complement).

Evaluation

- We evaluate the obtained clusters against the **human judgments** on verb semantic selection present in the T-PAS resource.

T-pass for *suonare*

suonare

RENAME SHOW SUBLABELS

1	[Human] **suonare** [Musical Instrument] [[Human]] trae il suono da, pratica [[Musical Instrument]]
2	[Human] **suonare** [Human] esegue musica, si esibisce in esecuzioni musicali, trae il suono da uno strumento musicale
3	[Human] **suonare** [Musical Composition] {canzone | brano | pezzo | concerto} [[Human]] esegue [[Musical Composition]] con strumenti musicali
4	[Human] **suonare** [Sound Maker] {campane | campanello | gong} [[Human]] fa risuonare, attiva, aziona [[Sound Maker]] come segnale
5	[Human] **suonare** [[Human]] segnala il suo arrivo azionando il campanello
6	[Musical Instrument | Sound Maker] **suonare** [[Musical Instrument | Sound Maker]] emette suoni
7	[Time Point] **suonare** Un suono indica lo scoccare di [[Time Point]]
8	[Human] **suonare** {l' allarme} [[Human]] da {l'allarme}
9	[Human1] **suonarle** a [Human2] [[Human1]] picchia, da botte a [[Human2]]
10	[Anything] **suonare** [Property | Anything] [[Anything]] sembra, appare [[Property]] | [[come Anything]]

Evaluation

- In T-PAS, for the O of *suonare* there are three judgments for three distinct verb senses:
- 'produce music from Musical Instrument';
- 'cause Sound Maker to produce a sound';
- 'play Musical Composition'.

Evaluation

- For the S of *suonare* we find:
- HUMAN 'performs music, performs in musical performances, draws sound from a musical instrument';
- ANYTHING 'seems, appears' ...:
- SOUND MAKER | MUSICAL INSTRUMENT 'emits sound'.

Evaluation

- There is **matching** between the judgment MUSICAL INSTRUMENT in both O and S with cluster **143**, which corresponds to the **dominant intersective cluster**.
- Several words corresponding to the judgment MUSICAL COMPOSITION are found in the **second intersective cluster 58** (canzone 'song', brano 'piece of music'), whose prototype is *album*.
- The apparent mismatch between judgment and prototype can be explained by assuming that albums and similar objects **are used metonymically** for the musical compositions recorded on them.
- Metonymical displacements obfuscate the semantics of the clusters (as predicted in Pustejovsky and Jezek 2008).

Evaluation

- The judgment SOUND MAKER is not directly matched by any cluster, although - as previously observed - cluster **157** *Sounds* contains several word of such kind (*campanello* 'doorbell').
- The evaluation confirms that the **centroid** is **not a good predictor** of the cluster semantics.

Evaluation

- For the Subject only, cluster **193** is matched by the judgement HUMAN for the sense 'play music' (note that the cluster label *Performers* is more fine-grained).
- The remaining clusters (**24**, **325**, **285**) are all matched by the semantic type ANYTHING for the sense 'to seem, to appear'.

Conclusions

- We provide evidence that there are exceptions to the perfect overlap between S and O with causative / inchoative verbs, and that these exceptions are not isolated, but rather are organized semantically themselves.
- We observe a fair matching between human judgments based on manual clustering and automatically obtained clusters.
- The evaluation suggests that clustering is **a viable methodology** to model semantic selection.
- The centroid is **not systematically the best predictor** of the cluster semantics.
- **Metonymic uses** may lead to incorrect semantic analysis.

Ongoing work

- Public graphical interface in 2020.
- CROATPAS first version to be presented at Clic-IT 2019. Marini and Jezek, Paper accepted!
- Multilingual pattern linking based on Italian, Croatian and English (in collaboration with Research Group of Information and Language Processing University of Wolverhampton).

THANK YOU FOR LISTENING!

THANK YOU FOR LISTENING!

The Lexicon

An Introduction

By **Elisabetta Ježek**, Associate Professor of Linguistics, University of Pavia

Oxford Textbooks in Linguistics
9780199601530| Hardback| £65.00 | January 2016
9780199601547| Paperback| £24.99 | January 2016