Investigating Lexical Sets through Distributional and Ontological Approaches

Bernardo Magnini
Fondazione Bruno Kessler, Trento, Italy
Email: magnini@fbk.eu

Joint work with:
Elisabetta Jezek, Anna Feltracco, Lorenzo Gatti and Edoardo Maria Ponti

Florence, June 9th 2017

WHAT ARE LEXICAL SETS?

Lexical sets are paradigmatic sets of words which occupy the same argument position of a verb, as found in a corpus. (cf. Hanks, 1996 and Jezek and Hanks, 2015)[1]

to read

-> Subject *reads* Object

-> Object *{book, letter, newspaper, report, paper, word, article, story, papers, time, text, mind, page, novel, magazine, poem, passage, ..}* [2]

[1] Hanks P., 1996. Contextual dependencies and lexical sets. *The International Journal of Corpus Linguistics*, 1(1).
Jezek E. and Hanks P., 2010, "What lexical sets tell us about conceptual categories." Lexis 4.7: 22.

[2] Lemmas are extracted from the BNC Corpus, using SketchEngine (Kilgarriff, A. et al., 2004, "Itri-04-08 the sketch engine." Information Technology 105: 116.)

Lexical sets change from verb to verb

- to read – OBJ: *{book, letter, newspaper, report, paper, word, article, story, papers, time, text, mind, page, novel, magazine, poem, passage, bible, ..}*
- to publish – OBJ: *{report, book, article, paper, result, work, letter, study, document, ..}*
- to write – OBJ: *{letter, book, article, poem, report, song, name, program, story, word, ..}*
- to send – OBJ: *{letter, message, copy, child, man, troops, money, ..report, .. food,..}*
- to devour – OBJ: *{book, meal, animal, plant, child, Mariana, buffalo, carcass, .. food,.. }*
- to eat – OBJ: *{food, meal meat, fish, breakfast, sandwich, lunch, dinner, bread, diet, ..}*

Lemmas are extracted from the BNC Corpus, using SketchEngine (Kilgarriff, A. et al.,2004, "Itri-04-08 the sketch engine." Information Technology 105: 116.)

Lexical sets change from verb to verb

- to read – OBJ: *{book, letter, newspaper, report, paper, word, article, story, papers, time, text, mind, page, novel, magazine, poem, passage, bible, ..}*
- to publish – OBJ: *{report, book, article, paper, result, work, letter, study, document,..}*
- to write – OBJ: *{letter, book, article, poem, report, song, name, program, story, word, ..}*
- to send – OBJ: *{letter, message, copy, child, man, troops, money, ..report, .. food,..}*
- to devour – OBJ: *{book, meal, animal, plant, child, Mariana, buffalo, carcass, .. food,.. }*
- to eat – OBJ: *{food, meal meat, fish, breakfast, sandwich, lunch, dinner, bread, diet, ..}*

Lemmas are extracted from the BNC Corpus, using SketchEngine (Kilgarriff, A. et al.,2004, "Itri-04-08 the sketch engine." Information Technology 105: 116.)

Different senses of a verb have different lexical sets

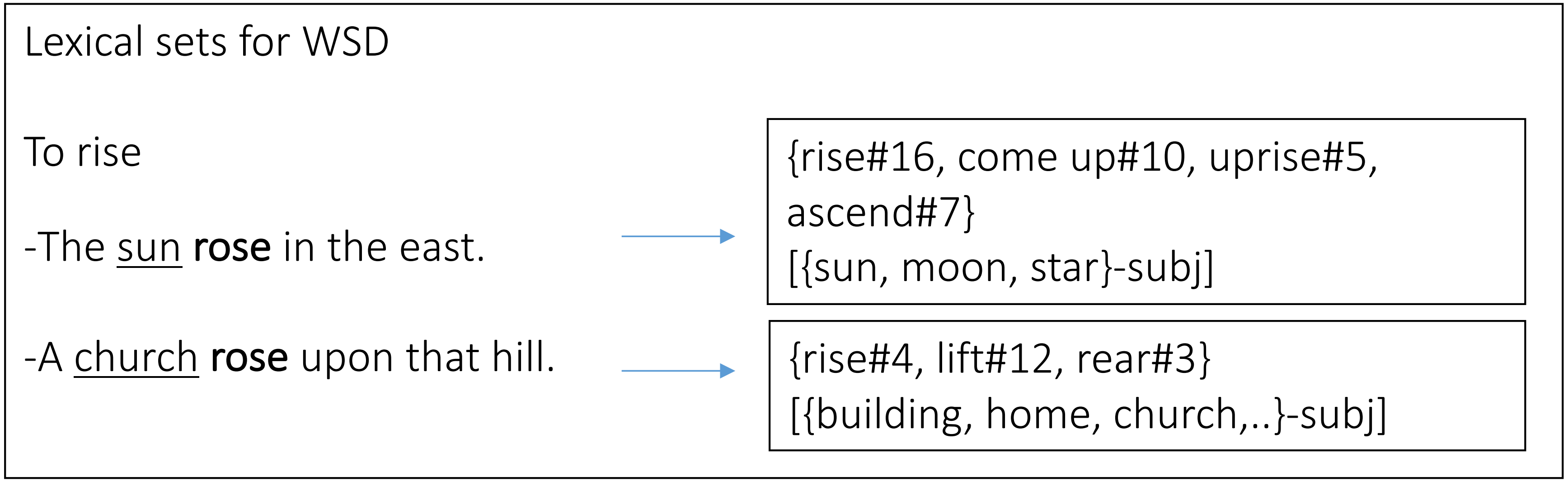
Subject of ‘to rise’ for different senses of the verb:

- rise, rise up, rear: *{building, home, church,..}*
- rise, come up, uprise: *{sun, moon}*
- rise, go up, increase (in value): *{turnover, price, share, rate, unemployment, profit, income, figure, temperature, cost, level, ..}*
- rise, come up, move up: *{smoke, ..}*

WHY LEXICAL SETS

- Verbs' selectional preferences
- Word Sense Disambiguation

if lexical sets are associated to verb senses -> verb meaning can be induced



Lexical sets for WSD

To rise

-The sun **rose** in the east. → {rise#16, come up#10, uprise#5, ascend#7} [{sun, moon, star}-subj]

-A church **rose** upon that hill. → {rise#4, lift#12, rear#3} [{building, home, church,..}-subj]

WHY LEXICAL SETS

- Verbs' selectional preferences
- Word Sense Disambiguation

  if lexical sets are associated to verb senses -> verb meaning can be induced
- Semantic Role Labeling -> to automatically annotate roles

Lexical sets for SRL

To rise

- The *land* was silent when the *sun* **rose** in the east.

Propbank Rise.01 :

Arg1: *Logical subject, patient, thing rising*

Candidate: "land" and "sun"

[{building, home, church, sun, moon, star}-subj]

no "land" -> Arg1: sun

OUTLINE

- Collecting lexical sets with an ontological approach
  - Using lexical resources: T-PAS and WordNet
  - Baseline and LEA algorithm
  - Results: better precision
- Investigating the internal structure of lexical sets with a distributional approach
  - Lexical sets as vectors
  - Do lexical set elements distribute uniformly in the space, or rather gather near or far the prototype?

ONTOLOGICAL APPROACH

GOAL: Building lexical sets for argument positions of Italian verbs at sense level [1]

WE NEED:

- a repository of verbs with the specification of the argument structure for each sense of the verb
- a repository of sentences associated to each verb sense from which the members of the lexical sets can be extracted

[1] Feltracco, Gatti, Magnolini, Magnini, Jezek: Using WordNet to Build Lexical Sets for Italian Verbs, Proceedings of the Eighth Global WordNet Conference, 2016.

METHODOLOGY

- We use the T-PAS resource [1] , a repository of verb frames for Italian in which :
  - the expected semantic type for each argument slot is specified (e.g. Human, Food, Event, Location, Artifact, …)
  - each frame is related to sentences in a corpus in which the verb is annotated

- In these sentences, we automatically annotate the sets of fillers for the argument slots of the selected verb -> the Baseline Algorithm and the Lea Algorithm

- Both algorithms use a mapping from Semantic types to MultiWordNet synsets [2]

T-PAS resource + MultiWordNet + Sentence Annotation -> Lexical Set

[1] Jezek E. et al., 2014, "T-PAS: a resource of corpus-derived Typed Predicate Argument Structures for linguistic analysis and semantic processing" In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14),* Reykjavik, Iceland.

[2] Pianta E. et al., 2002. "MultiWordNet: developing an aligned multilingual database". In *Proceedings of the 1st international conference on global WordNet*, volume 152, pages 55–63.

T-PAS: Typed Predicate Argument Structures

T-PAS is a repository of corpus-derived verb patterns for Italian with specification of the expected semantic type for each argument slot.
T-PASs are acquired following Corpus Patten Analysis methodology (Hanks, 2004).

Hanks P., 2004. “Corpus pattern analysis”. In *Proceedings of the Eleventh EURALEX International Congress*, Lorient, France, Universite de Bretagne-Sud;

T-PAS: Typed Predicate Argument Structures

T-PAS is a repository of corpus-derived verb patterns for Italian with specification of the expected semantic type for each argument slot.
T-PASs are acquired following Corpus Patten Analysis methodology (Hanks, 2004).

repository of T-PAS

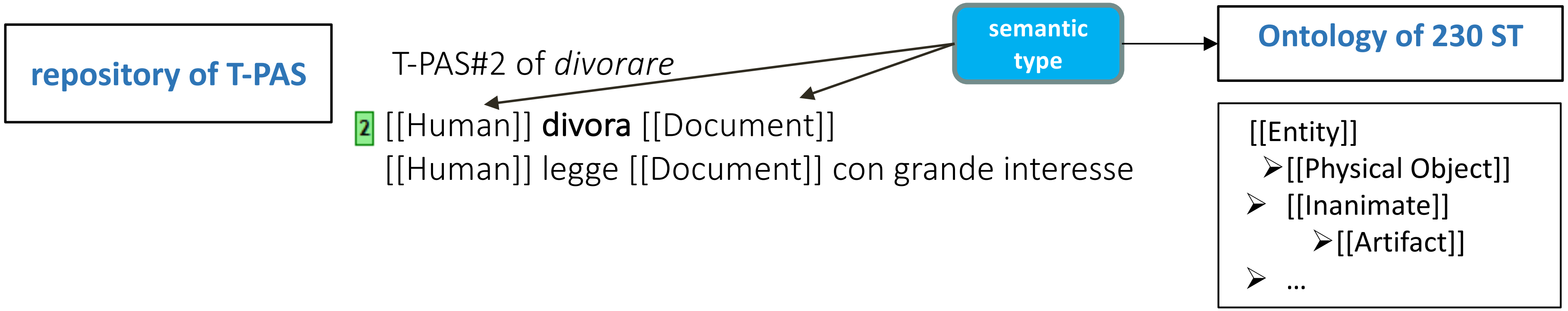
T-PAS#2 of *divorare (devour)*

2 [[Human]] **divora** [[Document]]
[[Human]] legge [[Document]] con grande interesse

Hanks P., 2004. “Corpus pattern analysis”. In *Proceedings of the Eleventh EURALEX International Congress*, Lorient, France, Universite de Bretagne-Sud;

T-PAS: Typed Predicate Argument Structures

T-PAS is a repository of corpus-derived verb patterns for Italian with specification of the expected semantic type for each argument slot.
T-PASs are acquired following Corpus Patten Analysis methodology (Hanks, 2004).



repository of T-PAS

T-PAS#2 of *divorare*

2 [[Human]] **divora** [[Document]]
[[Human]] legge [[Document]] con grande interesse

semantic type

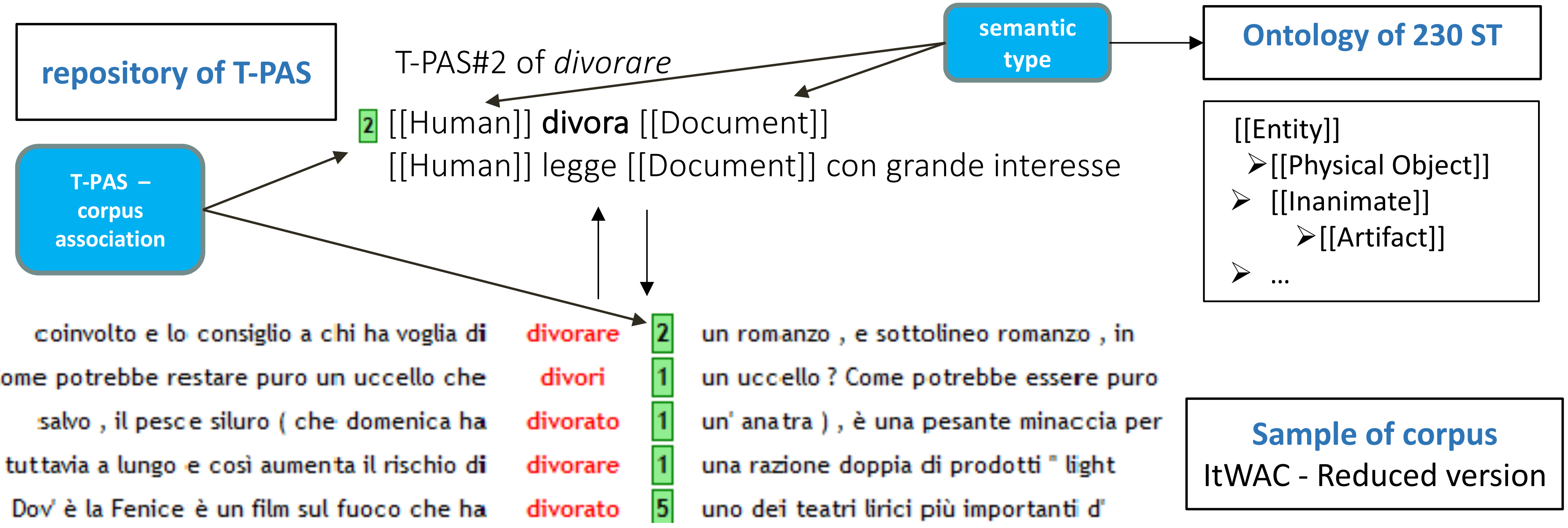
Ontology of 230 ST

[[Entity]]
- ➢[[Physical Object]]
- ➢ [[Inanimate]]
  - ➢[[Artifact]]
- ➢ …

Hanks P., 2004. “Corpus pattern analysis”. In *Proceedings of the Eleventh EURALEX International Congress*, Lorient, France, Universite de Bretagne-Sud;

T-PAS: Typed Predicate Argument Structures

T-PAS is a repository of corpus-derived verb patterns for Italian with specification of the expected semantic type for each argument slot.
T-PASs are acquired following Corpus Patten Analysis methodology (Hanks, 2004).



T-PAS#2 of *divorare*

2 [[Human]] **divora** [[Document]]
[[Human]] legge [[Document]] con grande interesse

coinvolto e lo consiglio a chi ha voglia di	divorare	2	un romanzo , e sottolineo romanzo , in
Come potrebbe restare puro un uccello che	divori	1	un uccello ? Come potrebbe essere puro
salvo , il pesce siluro (che domenica ha	divorato	1	un' anatra) , è una pesante minaccia per
tuttavia a lungo e così aumenta il rischio di	divorare	1	una razione doppia di prodotti " light
Dov' è la Fenice è un film sul fuoco che ha	divorato	5	uno dei teatri lirici più importanti d'

Visit tpas.fbk.eu and download T-PAS

Hanks P., 2004. “Corpus pattern analysis”. In *Proceedings of the Eleventh EURALEX International Congress*, Lorient, France, Universite de Bretagne-Sud;

SENTENCE ANNOTATION AND LEXICAL SET BUILDING

Input data from T-PAS

repository of T-PASs

T-PAS#2 of *preparare*
[[Human]] **prepara** [[Food | Drug]]
Eng.: [[Human]] **prepare** [[Food | Drug]]

Sentences

"La nonna, prima di infornare le patate, **prepara** una torta"
Eng. "The grandmother, before baking the potatoes, **prepares** a cake"

SENTENCE ANNOTATION AND LEXICAL SET BUILDING

Input data from T-PAS

repository of T-PASs

T-PAS#2 of *preparare*
[[Human]] **prepara** [[Food | Drug]]
Eng.: [[Human]] **prepare** [[Food | Drug]]

Sentences

"La nonna, prima di infornare le patate, **prepara** una torta"

Eng. "The grandmother, before baking the potatoes, **prepares** a cake"

Sentence annotation = annotate lexical items corresponding to Semantic type

[[Human]] – subj = ? [[Food]] – obj = ? [[Drug]] – obj = ?

SENTENCE TAGGING AND LEXICAL SET BUILDING

Input data from T-PAS

repository of T-PASs

T-PAS#2 of *preparare*
[[Human]] **prepara** [[Food | Drug]]
Eng.: [[Human]] **prepare** [[Food | Drug]]

Sentences

"La nonna, prima di infornare le patate, **prepara** una torta"
Eng. "The grandmother, before baking the potatoes, **prepares** a cake"

Sentence tagging = annotate lexical items corresponding to Semantic types

[[Human]] – subj = ?   [[Food]] – obj = ?   [[Drug]] – obj = ?

For all the sentences
=
Lexical set

THE BASELINE ALGORITHM

to identify possible candidate members:

[[Human]] – subj = ? [[Food]] – obj = ? [[Drug]] – obj = ?

1) uses TextPro 2.0[1] for PoS-tagging and lemmatization
2) check if each lemma is in MultiWordNet
3) use the Semantic type – synsets mapping

Automatic Semantic Type-Synsets mapping
[[Human]] -> human#n
[[Food]] -> food#n
[[Drug]] -> drug#n

if the lemma belongs (is an **hyponym)** to a corresponding mapped synset then the lemma is included in the lexical set

[1] Pianta E. et al., 2008. The TextPro Tool Suite. In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco.

BASELINE

T-PAS#2 of *preparare*

[[Human]] **prepara** [[Food | Drug]]

Eng.: [[Human]] **prepare [[Food | Drug]]**

[[Human]] – subj = ? [[Food]] – obj = ? [[Drug]] – obj = ?

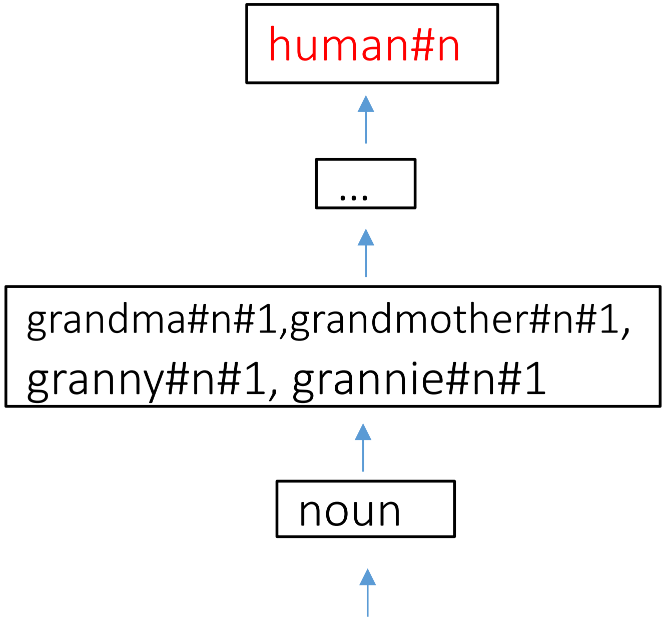
"La nonna, prima di infornare le patate, **prepara** una torta"

Eng. "the grandmother, before baking the potatoes, **prepares** a cake"

BASELINE

T-PAS#2 of *preparare*
[[Human]] **prepara** [[Food | Drug]]
Eng.: [[Human]] **prepare** [[Food | Drug]]

[[Human]] – subj = ?   [[Food]] – obj = ?   [[Drug]] – obj = ?



human#n

...

grandma#n#1,grandmother#n#1, granny#n#1, grannie#n#1

noun

"La nonna, prima di infornare le patate, **prepara** una torta"
Eng. "the grandmother, before baking the potatoes, **prepares** a cake"

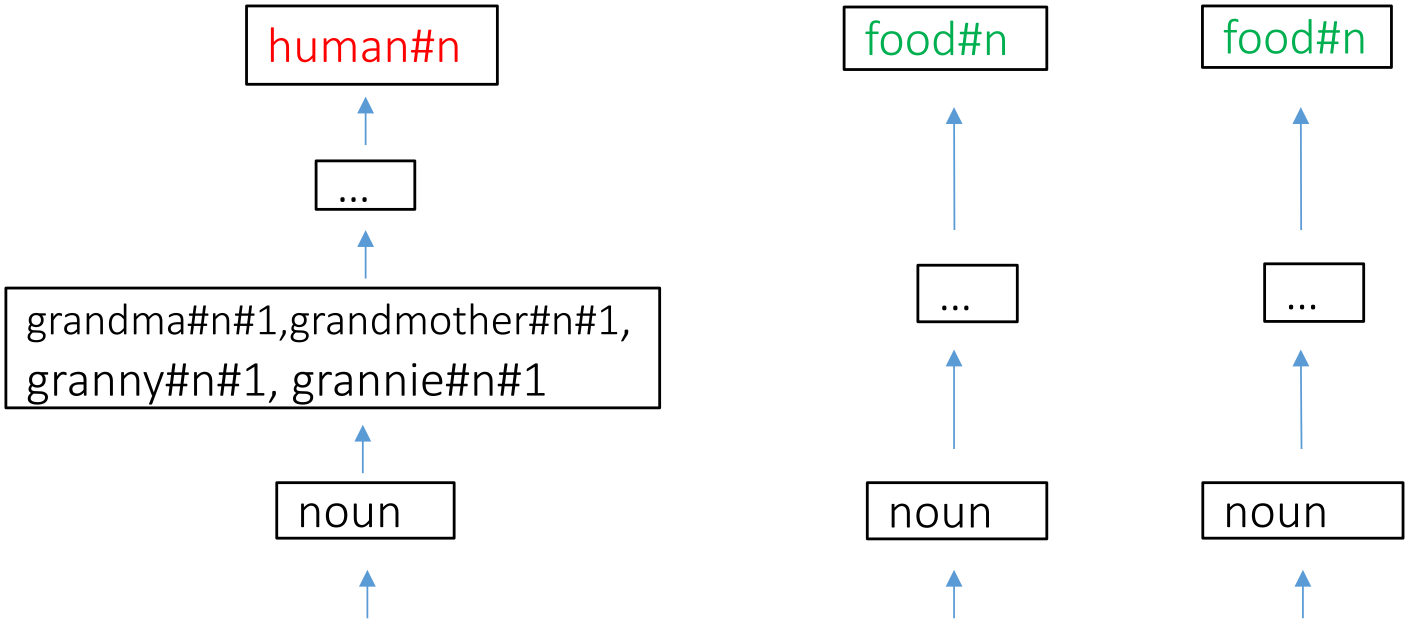
BASELINE

T-PAS#2 of *preparare*

[[Human]] **prepara** [[Food | Drug]]

Eng.: [[Human]] **prepare** [[Food | Drug]]

[[Human]] – subj = ? [[Food]] – obj = ? [[Drug]] – obj = ?



human#n

…

grandma#n#1,grandmother#n#1, granny#n#1, grannie#n#1

noun

food#n

…

noun

food#n

…

noun

"La nonna, prima di infornare le patate, **prepara** una torta"

Eng. "the grandmother, before baking the potatoes, **prepares** a cake"

LEA:
THE LEXICAL SET EXTRACTION ALGORITHM

to identify possible candidate members:

[[Human]] – subj = ?   [[Food]] – obj = ?   [[Drug]] – obj = ?

Baseline +

- uses dependency tree of the sentence
- recognizes named entities with TextPro 2.0
- checks for multiword expressions in MWN

-> we expect a higher Precision

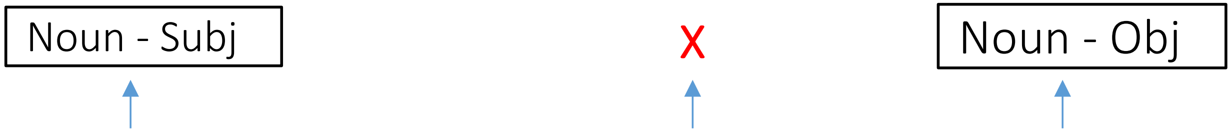
LEA: syntactic information

T-PAS#2 of *preparare*

[[Human]] **prepara** [[Food | Drug]]

Eng.: [[Human]] **prepare** [[Food | Drug]]

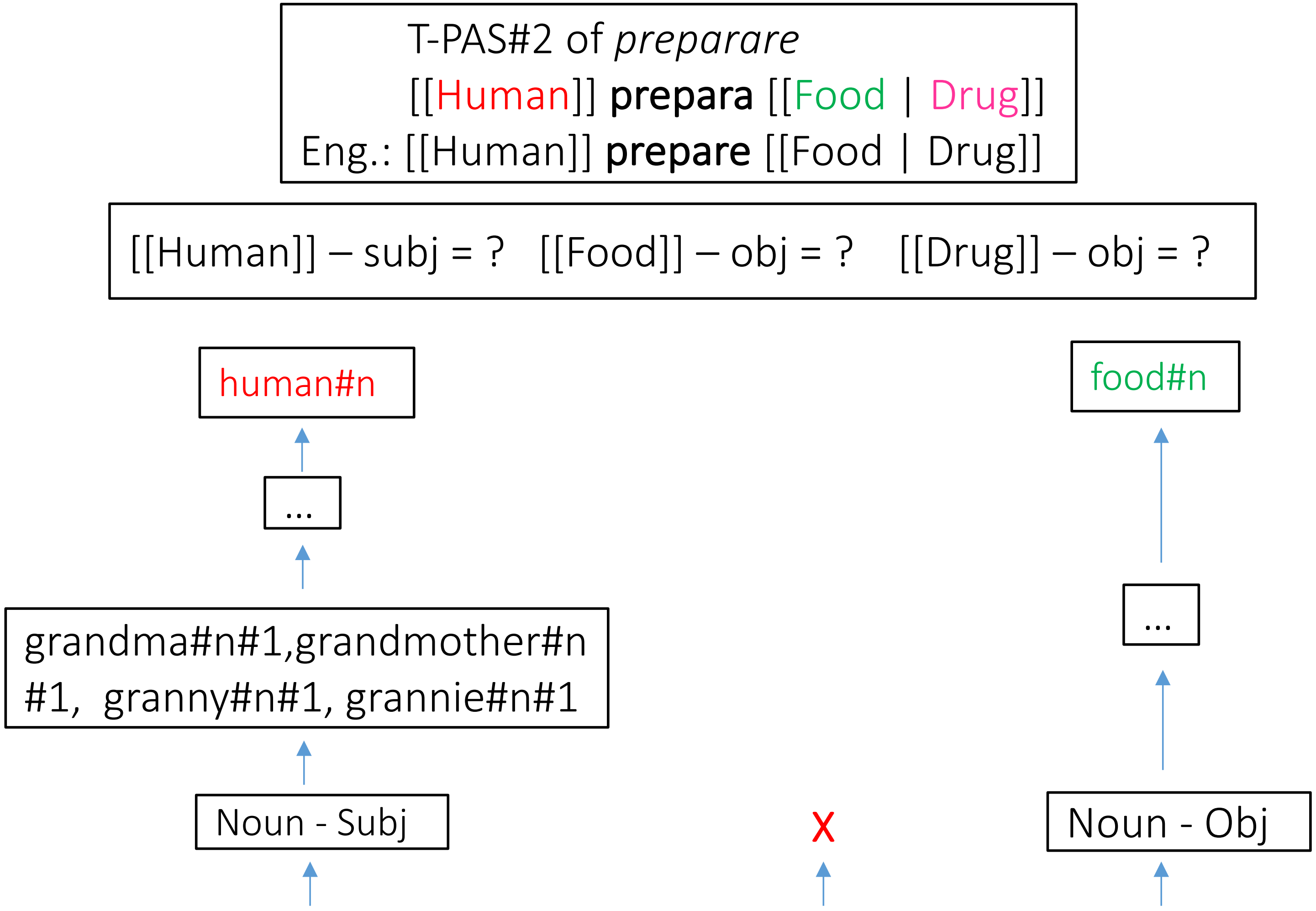
[[Human]] – subj = ?   [[Food]] – obj = ?   [[Drug]] – obj = ?



"La nonna, prima di infornare le patate, **prepara** una torta"

Eng. "the grandmother, before baking the potatoes, **prepares** a cake"

LEA: syntactic information



T-PAS#2 of *preparare*

[[Human]] **prepara** [[Food | Drug]]

Eng.: [[Human]] **prepare** [[Food | Drug]]

[[Human]] – subj = ? [[Food]] – obj = ? [[Drug]] – obj = ?

human#n

…

grandma#n#1,grandmother#n#1, granny#n#1, grannie#n#1

Noun - Subj

X

food#n

…

Noun - Obj

"La nonna, prima di infornare le patate, **prepara** una torta"

Eng. "the grandmother, before baking the potatoes, **prepares** a cake"

LEA: NER and MWE

T-PAS#2 of *preparare*
[[Human]] **prepara** [[Food | Drug]]
Eng.:[[Human]] **prepara** [[Food | Drug]]

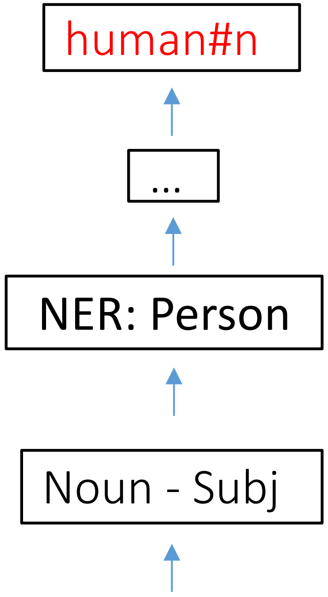
[[Human]] – subj = ? [[Food]] – obj = ? [[Drug]] – obj = ?

"Maria Rossi **prepara** la conserva di frutta"
Eng.: "Maria Rossi **prepares** the fruit conserve"

LEA: NER and MWE

T-PAS#2 of *preparare*
[[Human]] **prepara** [[Food | Drug]]
Eng.:[[Human]] **prepara** [[Food | Drug]]

[[Human]] – subj = ?   [[Food]] – obj = ?   [[Drug]] – obj = ?

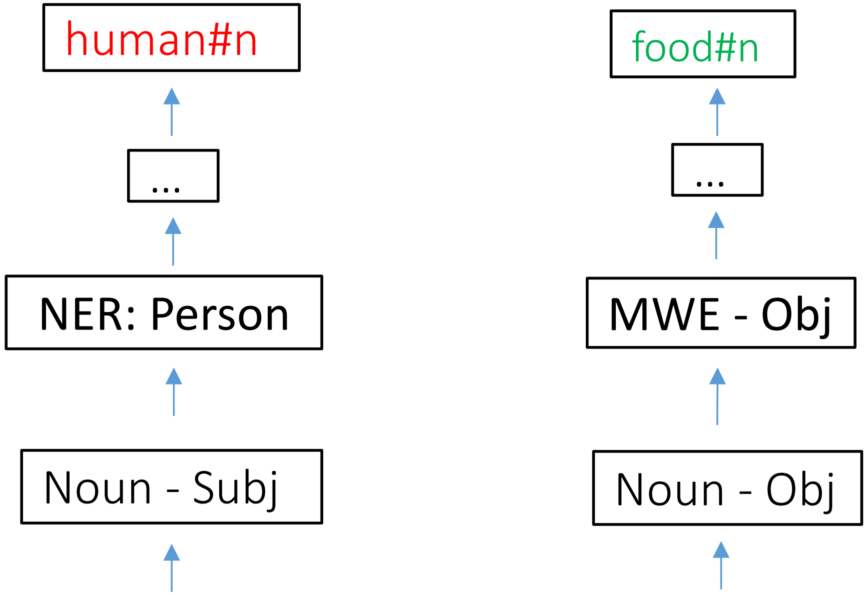


"Maria Rossi **prepara** la conserva di frutta"
Eng.: "Maria Rossi **prepares** the fruit conserve"

LEA: NER and MWE

T-PAS#2 of *preparare*
[[Human]] **prepara** [[Food | Drug]]
Eng.: [[Human]] **prepare** [[Food | Drug]]

[[Human]] – subj = ? [[Food]] – obj = ? [[Drug]] – obj = ?



human#n

...

NER: Person

Noun - Subj

food#n

...

MWE - Obj

Noun - Obj

"Maria Rossi **prepara** la conserva di frutta"
Eng.: "Maria Rossi **prepares** the fruit conserve"

GOLD STANDARD

- 3 annotators manually marked the lexical items or the multiword expressions that correspond to the T-PAS Semantic Types (no pronouns, no relative clauses)

- 500 examples
(10 sentences x a selection of 10 different STs x 5 different T-PASs;
e.g. 10 sentences x [[Food]] x 5 T-PASs)

- 981 annotated tokens out of 15090

RESULTS: SENTENCE ANNOTATION

Results for sentence annotation for Baseline and LEA

Automatic mapping			
	Precision	Recall	F1
Baseline	0.28	0.42	0.34
LEA	0.70	0.25	0.37

RESULTS: SENTENCE ANNOTATION

Results for sentence annotation for Baseline and LEA

Automatic mapping			
	Precision	Recall	F1
Baseline	0.28	0.42	0.34
LEA	0.70	0.25	0.37

Evaluation.

Inaccuracies are due to:

- recognition of proper names
  (Baseline 10 /185 , Lea 26/185)
- PoS tagging step
- dependency parsing step

- automatic mapping of STs - synsets
- different structure of the two resources
  (e.g. in T-PAS [[Machine]] is a hypernym of [[Vehicle]], the same is not true for machine#n in MWN)

RESULTS: SENTENCE ANNOTATION

Results for sentence annotation for Baseline and LEA

Automatic mapping			
	Precision	Recall	F1
Baseline	0.28	0.42	0.34
LEA	0.70	0.25	0.37

Results after manual revision of the Semantic Type - synsets mapping

Mapping with manual revision of 11 ST			
Baseline	0.30	0.52	0.38
LEA	0.72	0.32	0.44

Evaluation.

Inaccuracies are due to:

•recognition of proper names
(Baseline 10 /185 , Lea 26/185)

•PoS tagging step

•dependency parsing step

•automatic mapping of STs - synsets

•different structure of the two resources
(e.g. in T-PAS [[Machine]] is a hypernym of [[Vehicle]], the same is not true for machine#n in MWN)

RESULTS: LEXICAL SET

Similarity between Gold Standard lexical set and lexical set annotated with Baseline and LEA (Dice's coefficient)

5 most populated lexical sets	Baseline	LEA
Cuocere#2-SBJ-[[Food]] *{pasta, pesce, sugo, carciofo,..}*	0.54	0.57
Crollare#1-SBJ-[[Building]]	0.71	0.60
Dirottare#1-OBJ-[[Vehicle]]	0.83	0.66
Prescrivere#2-OBJ-[[Drug]]	0.42	0.46
Togliere#4-OBJ-[[Garment]]	0.72	0.61

Baseline -> low precision causes major differences with the gold standard sets

LEA -> low recall penalizes the amount of detected items given few sentences to annotate

CONSIDERATIONS

Final considerations:

•on large scale acquisition, the higher precision for LEA is more promising than the Baseline

•first step on automatic acquisition of lexical sets trhough lexical resources

Further work:

•extension of the sentence annotation and lexical set population for all T-PAS

•comparison of lexical sets in different T-PASs with the same Semantic type

THE "SHIMMERING" NATURE OF LEXICAL SETS

- Hanks and Pustejovsky (2005) and Hanks and Jezek (2008) propose an ontology where fillers are clustered into semantic types
- These categories are problematic, as lexical sets tend to "shimmer" (Jezek and Hanks 2010): their membership tends to change according to the verb they associate with

- [[Human]] *wash* [[**BodyPart**]]: *{hand, hair, face, foot, mouth ...}*
- **[[Human]] *amputate* [[BodyPart]] :** *{leg, limb, arm, hand, finger, ..}*

THE “SHIMMERING” NATURE OF LEXICAL SETS

- Hanks and Pustejovsky (2005) and Hanks and Jezek (2008) propose an ontology where fillers are clustered into semantic types
- These categories are problematic, as lexical sets tend to “shimmer” (Jezek and Hanks 2010): their membership tends to change according to the verb they associate with

- [[Human]] *wash* [[**BodyPart**]]: *{hand, hair, face, foot, mouth ...}*
- **[[Human]] *amputate* [[BodyPart]] :** *{leg, limb, arm, hand, finger, ..}*

GOAL: grounding “SHIMMERING” on empirical evidence exploiting methodologies offered by distributional semantics.

INTERNAL STRUCTURE OF LEXICAL SETS

Prototypes and Centroids

Linguistic Categories are radial continua with prototypical and peripheral members (Rosch 1973).
Lexical sets mapped to vectors are points in a multi-dimensional space. Their centroid (Euclidean mean) is here equated to a prototype. We calculated the centroid of S and O for every verb.

Cosine Distance

Cosine distance is a measure of closeness between vector pairs. Its values span from 0 (overlap) to 1 (maximum distance). We evaluated it for each word wrt the centroid of its lexical set.

DATA AND METHOD [1]

- Data are sourced from a sample of ItWac, (Baroni et al. 2009).
- This sample was further enriched with morpho-syntactic information through the MATE-tools parser (Bohnet 2010) and filtered by sentence length (< 100).
- Eventually, sentences in the sample amounted to 2,029,454 items.
- A target group of 20 causative-inchoative Italian verbs was taken from Haspelmath et al. (2014)
- Argument fillers are automatically extracted for three “macro-roles” (Dixon 1994):
  - subjects of transitive verbs (A),
  - subjects of intransitive verbs (S)
  - objects (O)

[1] Edoardo Maria Ponti, Elisabetta Jezek and Bernardo Magnini. Grounding the Lexical Sets of Causative-Inchoative Verbs with Word Embedding. In: Proceedings of the Third Italian Conference on Computational Linguistics (CLIC-it). 5-6 December 2016, Napoli (Italy).

DATA AND METHOD

Plinio il Vecchio non cita più il Po di Adria perche' l' Adige aveva subito una rotta ed era confluito nella Filistina.

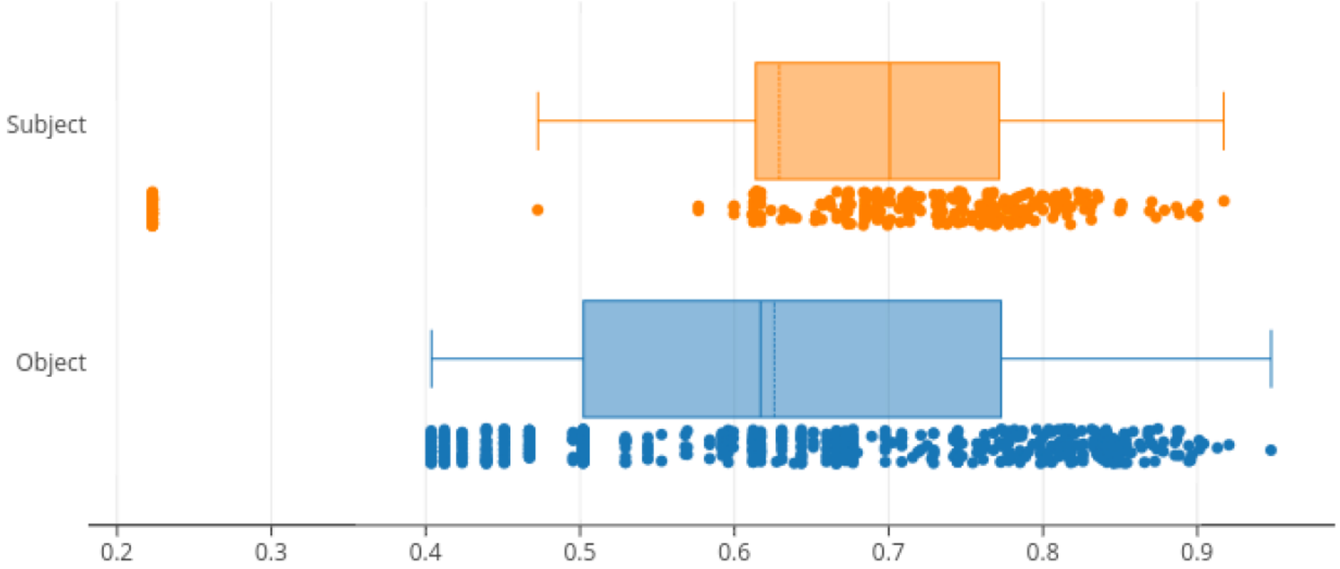
Verb	A	S	O
citare	Vecchio	--	Po
subire	Adige	--	rotta

- Entries collapsed by verb lemma so that each became associated to three sets of fillers (one per macro-role).
- Each of the argument fillers was mapped to a vector relying on a space model pre-trained **throughWord2Vec** (Dinu, et al 2015)
- Fillers weighted by absolute **frequency**; cosine similarity
- A lexical set is represented by the **centroid** of the filler vectors

DISTRIBUTIONAL APPROACH

Distance of Set Members from Centroid

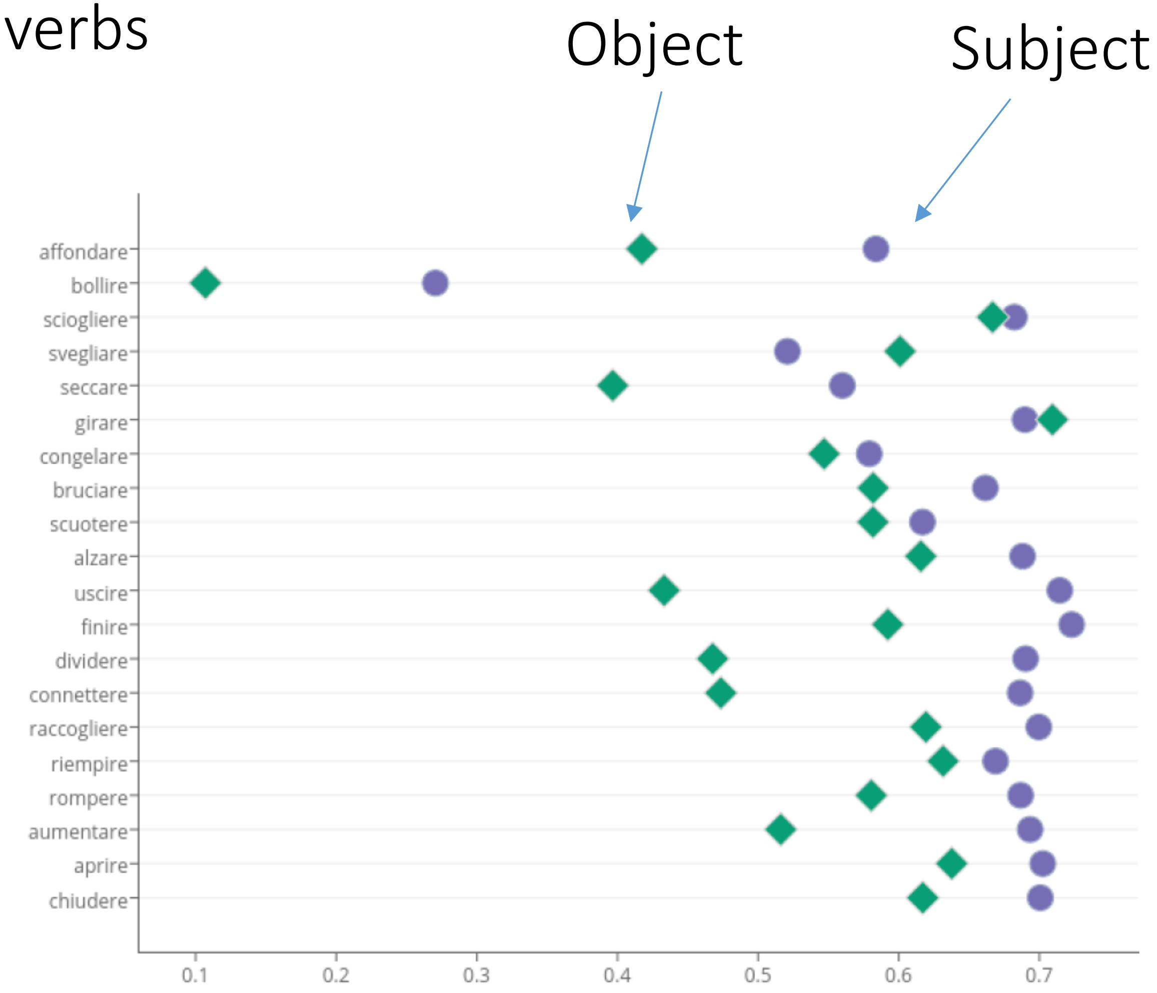
In-depth analysis: S set lies in a more compact range of distances, whereas O is more scattered. On the other hand, the vectors of S tend to be farther from the centroid.



Cosine distances of the members of S and O of the verb *dividere*

INTERNAL STRUCTURE OF LEXICAL SETS

Plot for all verbs



Median value of cosine distances of the members of S (blue circles) and the members of O (green diamonds) for each verb.

CONCLUSION

- Initial questions:
  - How are lexical set structured?
  - Do their elements distribute uniformly in the space, or rather gather near or far the prototype?

- Results: the Subject lexical set lies in a more compact range of distances, whereas Object is more scattered.
- On the other hand, the vectors of Subject tend to be farther from the centroid.
- This implies that Object behaves more similarly to a radial category, whereas S just populates the periphery

Further Directions

- Ontology-based and distributional methods

- Distributional representation may benefit from lexical resources (e.g. “retrofitting – Faruqui et al. 2015)
- Lexical resources may be grounded on distributional representations

- A unified model?