Towards an Italian Constructicon



Francesca Masini[*], **Beatrice Bernasconi**[^], **Claudia Borghetti**[*], **Lucia Busso**[§], **Maria Pina De Rosa**[°], **Claudio Iacobini**[#], **M. Silvia Micheli**, **Ludovica Pannitto**[*], **Flavio Pisciotta**[°], **Fabio Tamburini**[*]

[*]Università di Bologna, [^]Università di Torino, [#]Università Roma Tre, [§]Aston University, [°]Università di Salerno

ICCG13, Göteborg, 26-28 August 2024

Why an Italian Constructicon?

- Marginal position of Italian in the international constructionist landscape
  - CxG applied to Italian too but to a lesser extent than other languages
  - No digital resource published
    - Several attempts towards an Italian FrameNet (Tonelli et al. 2009; Lenci et al. 2010)
    - ***IFrameNet*** (Basili et al. 2017), under development



First steps towards an Italian Constructicon

- Preliminary features and functionalities
- External challenges
  - A shared standard is still lacking
  - The project is still unfunded…
- Internal challenges
  - how to develop a **machine-readable formalization of constructions** flexible enough to represent all relevant units/facts
  - how to make the resource as **interoperable** as possible while extending it to represent **non-standard features** or **language-specific** facts
  - how to represent **links** among constructions
  - how to make the resource **useful** for several purposes





Outline [adapted from Lyngfelt 2024]

General issues

- Purpose
- Autonomy
- Scope and focus
- Heuristics
- Openness

Design issues

- Annotation scheme
- CoNLL format
- Availability and population

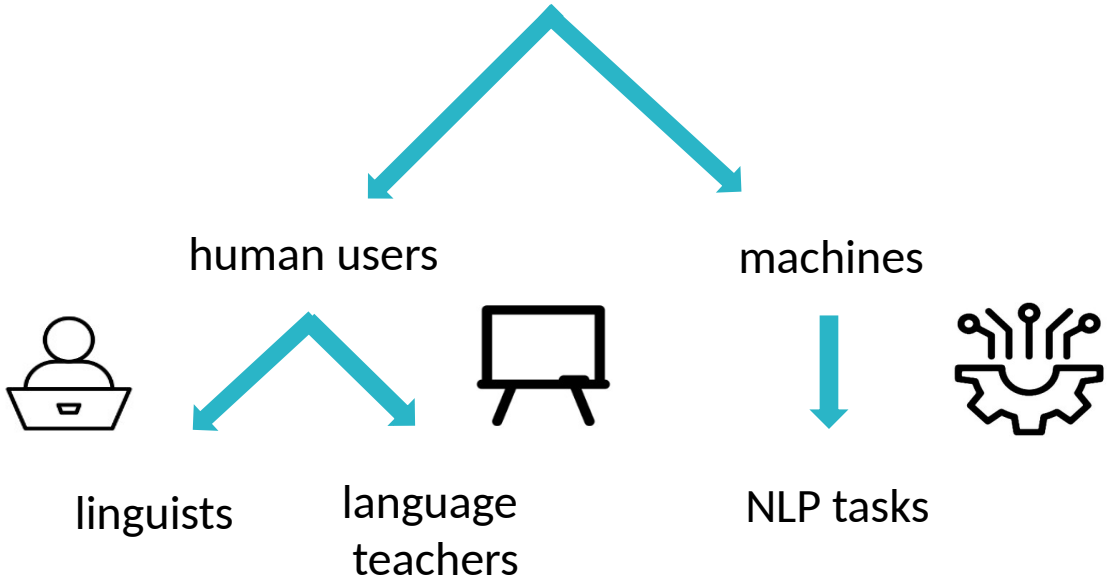
General issues



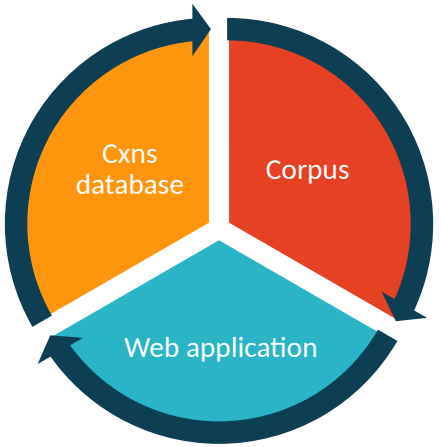
Purpose
Autonomy
Scope and focus
Heuristics
Openness

General issues: purpose

- The Italian Constructicon has **theoretical**, **applied**, and **computational purposes**
- Therefore, it is designed to be addressed to:

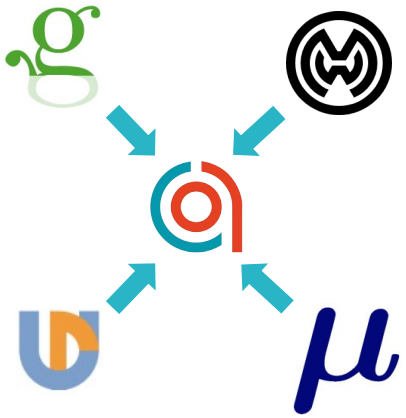


- It is a **tripartite resource** made of:
  - a git-managed **database of constructions**
  - a **corpus/treebank of annotated example sentences in UD-format,** linked to as many constructions as possible
  - a human-user friendly **web application**



General issues: autonomy

- Our theoretical and design choices make the Italian Constructicon **autonomous** and not directly dependent on other resources
- However, we still want it to be as «**connected**» as possible with other existing resources:
  - **Universal Dependencies** → constructions and examples annotated in UD (see Nivre later today, Weissweiler et al. 2024)
  - **Grew** → integration with Grew formalism (Graph Rewriting for NLP)
  - **Open Multilingual Wordnet** and **Universal Morphology** → ontological classes (adapted)
  - **Unified Verb Index** → semantic roles (adapted)
  - Other constructicons (**MoCCA**, Lorenzi et al. 2024) → reference to Croft's (2022) comparative concepts whenever possible
  - **IFrameNet** → unavailable but under development (future intergration)



General issues: scope and focus

- Differently from other constructicons, Italian Constructicon aims to reach a **greater coverage** in terms on construction types, in compliance with the **lexicon-syntax continuum** promoted by CxG, by giving space to **morphological constructions**
- The resource is **not biased on predicative constructions** since it is not based on FrameNet
- The main focus falls on (**semi**-)**schematic constructions**:
  - both <u>syntactic</u> (argument structures, formal idioms, etc.) and <u>morphological</u> (word formation)



  - Cognate Object Constructions
    dormire$_{V}$ *sonni*$_{N}$ *tranquilli* ‘sleep soundly’
  - Multiword expressions
    fare una N-*ata* ‘do an action in the style of N’



  - Verb-Noun compounds
    porta$_{V}$*foglio*$_{N}$ ‘wallet’
  - Prefixation, e.g. *semi*-Adj ‘semi-Adj’
    seminascosto$_{ADJ}$ ‘half hidden’

General issues: heuristics

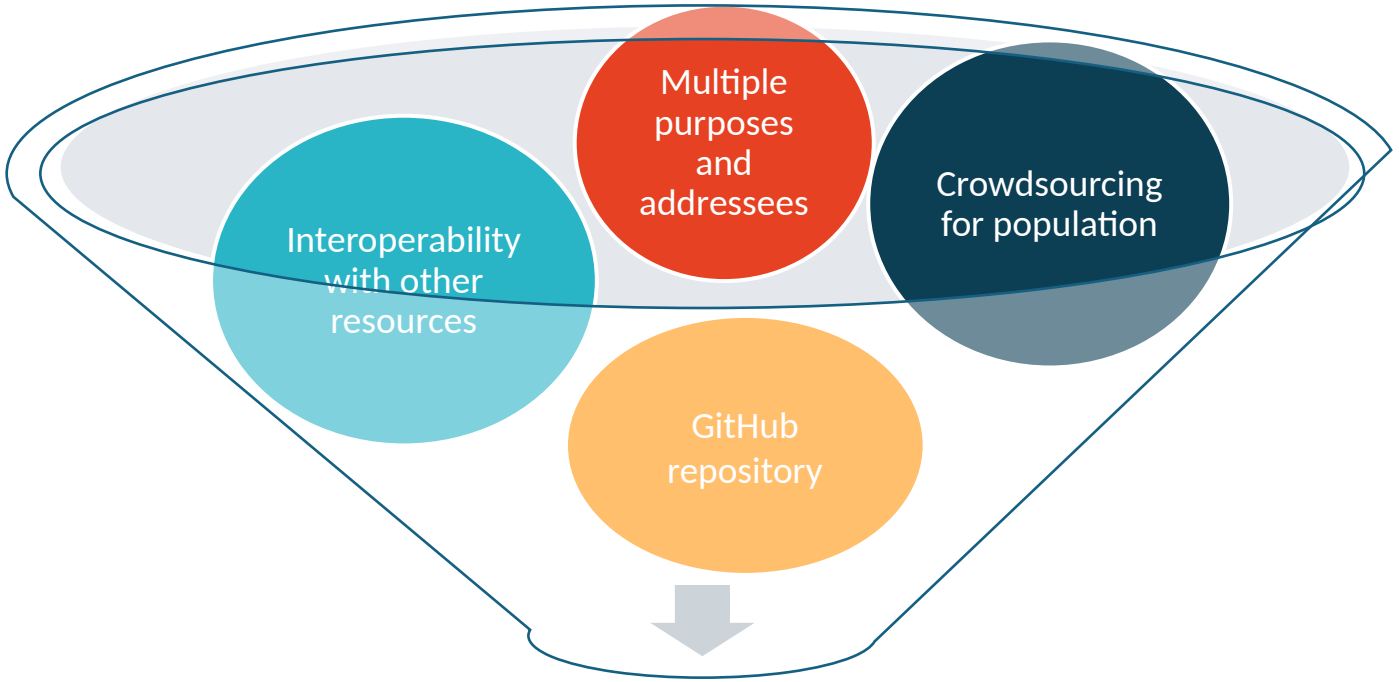


- The Constructicon currently hosts 14 cxns (pilot)
- The first population phase of the resource relies on the following strategies
  - Extraction of constructions already described in the literature
  - Extraction of constructions from already existing lists (e.g., CombiNet, MWEs)
  - Extraction of constructions from Italian L2 textbooks and learner corpora
- In the future we plan to implement:
  - Construction mining
  - Crowdsourcing population



General issues: openness

- The characteristics of the Italian Constructicon outlined so far contribute to make it an **open and collaborative resource**



Design issues



Annotation scheme
CoNLL format
Availability and population

Annotation scheme

Create construction

Name

Definition

Formattazioni

Restrictions

Collocational preferences

Formattazioni

CoNLL format

Usage

Select...

Formal tags

Functional tags

Complexity level tags

Category tags

Schematicity Level

Select...

CEFR level

Select...

Horizontal links

Vertical links

Examples 1

Formattazioni

Examples 5

UD examples

Notes

References

Data Collector

Status

submitted

Annotation scheme

cxn-id: 68

unique and stable identifier for the constructional entry

name: salta fuori che V

“human-intelligible” name for the cxn, chosen through a decision tree

cxn-machine-readable: cxn_68.conllc

definition:
A new piece of information comes to the speaker's knowledge from an external source. The information acquired is often unexpected or contradicts the speaker's expectations on the state of affairs, thus generating surprise in the speaker.
However, since the moment of acquisition and the moment of enunciation are distinct, this construction does not convey that the speaker is currently surprised, but it is used to convey or generate surprise in the audience.

restrictions:
The main verb saltare fuori is always impersonal, so it has no subject and it is always found in the 3rd person singular.
The verb in the complement clause is always in a finite form.

coll-preferences:

"Textual" fields
(dictionary-like descriptions)

- *definition*: description of the construction's **function**
- *restrictions*: **soft and hard constraints** on the cxn slots/fillers
- *collocational preferences*: **preferred fillers** for the open slots

Annotation scheme

usage:

form-tags:
- cc:cxn:complement-clause-construction
- impersonal construction

function-tags:
- cc:sem:evidentiality
- cc:sem:mirative

complexity-level:
- clause

category-tags:
- not applicable

schematicity: partially filled/schematic

Other constructional properties
(coded as tags)

- Usage: preferred **contexts of occurrence** of the cxn (wrt Style, Variety, Modality)
- Complexity and schematicity of the cxn
- Part of speech (when dealing with phrases/words)

https://comparative-concepts.github.io/cc-database/#sem:evidentiality

evidentiality (*sem*)

Id	sem:evidentiality
Type	meaning
Alias(es)	evidentiality
Taxonomy	modality (*sem*) **evidentiality (*sem*)**
Definition	a category related to epistemic modality which indicates the epistemic justification for believing a proposition. *Example*: English *I hear you're going to Stanford this fall* is an evidential construction, using the first singular simple present form *I hear...*, which indicates that my evidence for your going to Stanford in the fall is hearsay. (Section 12.3.4)

Annotation scheme

horizontal-links:
- 167

vertical-links:

IDs of horizontally and vertically linked cxns

```
name: "viene fuori che V"

cxn-machine-readable: cxn_167.conllc

definition:
    A new piece of information comes to the speaker's knowledge
from an external source, ...
```

examples:
- 1_Paisà_FP06072024
- 2_Paisà_FP06072024
- 3_Paisà_FP06072024
- 4_Paisà_FP06072024
- 5_Paisà_FP06072024

cefr-level:

Level of proficiency of the cxn for L2 learners of Italian

references:
- Pisciotta2023confini

collector: Flavio

note:
This construction can have both an evidential and a mirative reading, depending on the surrounding context.
It is often found in adversative (example1, example5), temporal, or more generally, coordinate clauses (example3), which favour a mirative 'counterexpectation' reading (i.e., the event or state in the complement clause is in contrast with the speaker's expectations). More rarely, the external source of information is specified in the context, triggering an evidential interpretation (i.e., the speaker gets to know something from a source).

Further comments on usage, function and behavior of the cxn

Annotation scheme

cxn-id: 68

name: salta fuori che V

cxn-machine-readable: cxn_68.conllc

CoNLL format

definition:
A new piece of information comes to the speaker's knowledge from an external source. The information acquired is often unexpected or contradicts the speaker's expectations on the state of affairs, thus generating surprise in the speaker.
However, since the moment of acquisition and the moment of enunciation are distinct, this construction does not convey that the speaker is currently surprised, but it is used to convey or generate surprise in the audience.

restrictions:
The main verb saltare fuori is always impersonal, so it has no subject and it is always found in the 3rd person singular.
The verb in the complement clause is always in a finite form.

coll-preferences:

CoNLL format

Metadata

```
# cxn_id = 68
# name = salta fuori che V
# function = ref:D is found out unexpectedly

# horizontal_links = 167
# vertical_links =

# fields = ID UD.FORM  LEMMA UPOS FEATS HEAD DEPREL REQUIRED WITHOUT SEM_FEATS SEM_ROLES ADJACENCY IDENTITY
```

A	_	saltare	VERB	Number=Sing|Person=3	0	root	1	CHILDREN:DEPREL=nsubj	_	_	_	_
B	fuori	fuori	ADV	_	A	advmod	1	_	_	_	_	_
C	che	che	SCONJ	_	D	mark	1	_	_	_	_	_
D	_	_	VERB,NOUN,ADJ	VerbForm=Fin	A	csubj	1	_	_	Eventuality	_	_

Fields

CoNLL format: fields

```
pattern { A [lemma="saltare", Number=Sing, Person=3];
          B [form="fuori", upos=ADV];
          C [form="che", upos=SCONJ];
          D [upos=NOUN|VERB|ADJ];
          A -[advmod]-> B;
          D -[mark]-> C;
          A -[csubj]-> D;}

without {A -[nsubj]-> S}
```

ID	UD.FORM	LEMMA	UPOS	FEATS	HEAD	DEPREL	REQUIRED	WITHOUT	SEM_FEATS	SEM_ROLES	ADJACENCY	IDENTITY
A	_	saltare	VERB	Number=Sing|Person=3	0	root	1	CHILDREN:DEPREL=nsubj	_	_	_	_
B	fuori	fuori	ADV	_	A	advmod	1	_	_	_	_	_
C	che	che	SCONJ	_	D	mark	1	_	_	_	_	_
D	_	_	VERB,NOUN,ADJ	VerbForm=Fin	A	csubj	1	_	_	Eventuality	_	_

CoNLL compatible fields

Constraining matching process in UD-annotated sentences

Further semantic, formal and mophosyntactic constraints

Connection to other resources

CoNLL format: fields



ID	UD.FORM	LEMMA	UPOS	FEATS	HEAD	DEPREL
A	_	saltare	VERB	Number=Sing|Person=3	0	root
B	fuori	fuori	ADV	_	A	advmod
C	che	che	SCONJ	_	D	mark
D	_	_	VERB,NOUN,ADJ	VerbForm=Fin	A	csubj

CoNLL compatible fields

Constraining matching process
in UD-annotated sentences

How to handle morphological constructions?

How to handle morphological constructions?

semi-
Adj

semi-autonomo
'semi-autonomous'

UPOS

- **free morphemes** = UD UPOS
- **BMORPH** (Bound MORPHeme) = bound forms, including affixes, affixoids, and combining forms

name = semiADJ
function = not fully or not properly ref:A-2
horizontal_links = 170
vertical_links =

# fields = ID	UD.FORM	LEMMA	UPOS	FEATS	HEAD	DEPREL	...
A	semi.+	semi.+	ADJ, NOUN	_	_	root	…
A.1	_	semi	BMORPH	_	_	der/m	...
A.2	_	_	ADJ	_	_	root/m	...

How to handle morphological constructions?

DEPREL (/m notation inspired by Guillaume et al. 2024)

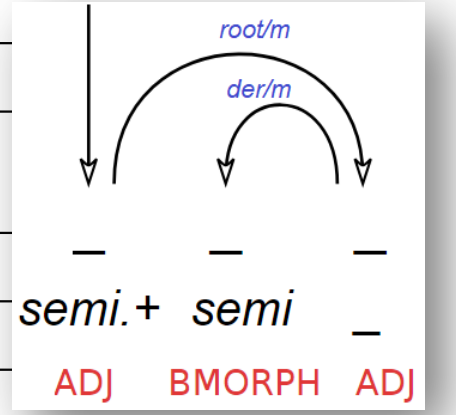
- **root/m** = the root inside the morphological construction
- **der/m** = links the *derivational* BMORPHs to the root/stem
- **case/m** = link among constituents in *subordinate compounds*
- **mod/m** = link among constituents in *attributive compounds*
- **conj/m** = link among constituents in *coordinate compounds*

semi-
Adj

semi-autonomo
'semi-autonomous'

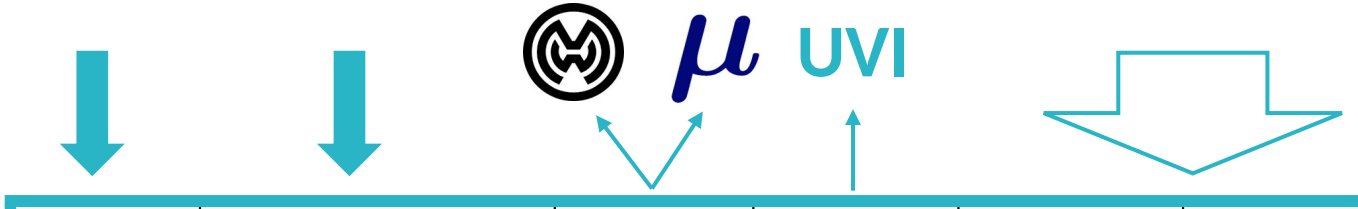
name = semiADJ
function = not fully or not properly ref:A-2
horizontal_links = 170
vertical_links =

# fields = ID	UD.FORM	LEMMA	UPOS	FEATS	HEAD	DEPREL
A	semi.+	semi.+	ADJ, NOUN	_	_	root
A.1	_	semi	BMORPH	_	_	der/m
A.2	_	_	ADJ	_	_	root/m



CoNLL format: fields

What about the additional fields?



REQUIRED	WITHOUT	SEM_FEATS	SEM_ROLES	ADJACENCY	IDENTITY
1	CHILDREN:DEPREL=nsubj	_	_	_	_
1	_	_	_	_	_
1	_	_	_	_	_
1	_	_	Eventuality	_	_

Further semantic, formal and mophosyntactic constraints

Connection to other resources

CoNLL format: fields

```
pattern { A [upos="NOUN"];
          B [form ="dopo"];
          C [upos="NOUN"];
          A < B; B < C;
          A -[nmod]-> C;
          C -[case]-> B;
          A.form = C.form
        }
```



N_i *dopo* 'after' N_i

ID	UD.FORM	LEMMA	UPOS	FEATS	HEAD	DEPREL	REQUIRED	WITHOUT	SEM_FEATS	SEM_ROLES	ADJACENCY	IDENTITY
A	_	_	NOUN	Number=Sing	0	root:obl	1	_	_	_	_	UD.FORM=C
B	dopo	dopo	ADP	_	C	case	1	_	_	_	A	_
C	_	_	NOUN	Number=Sing	A	nmod	1	_	_	_	B	UD.FORM=A

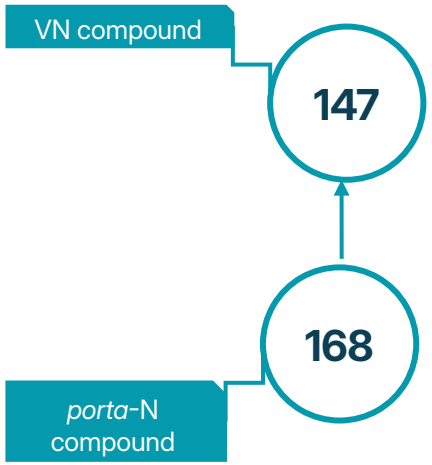


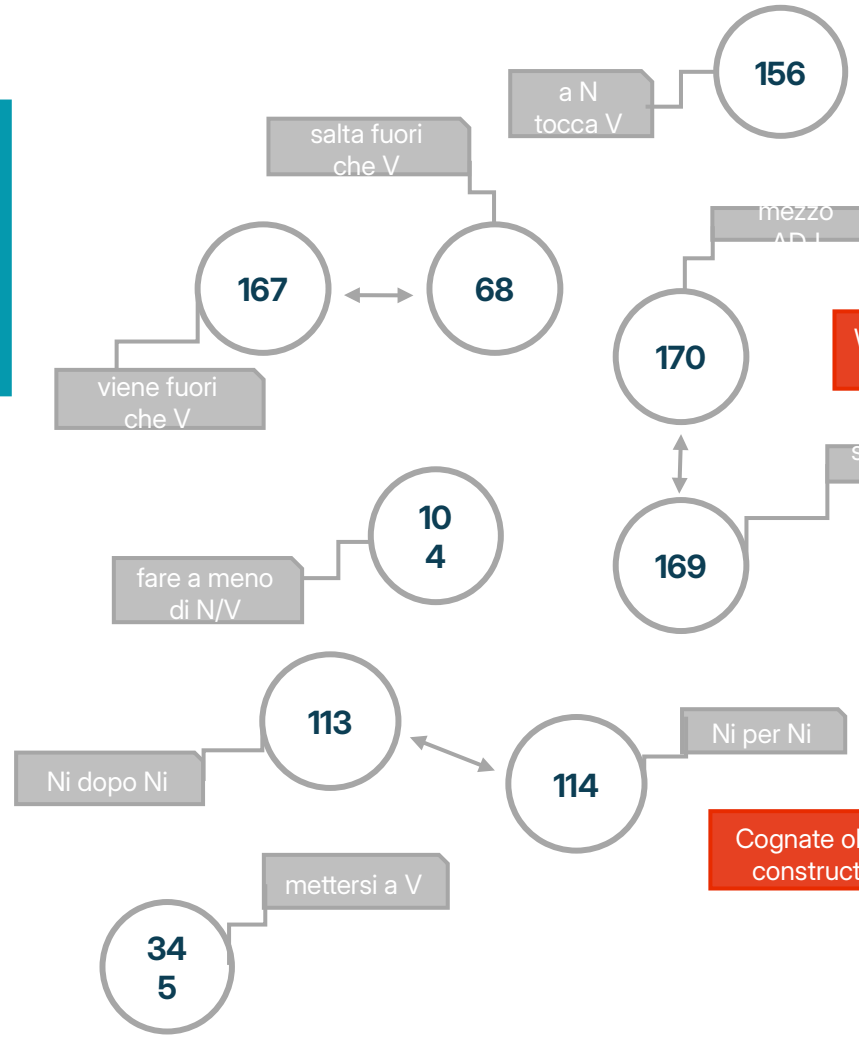
cxn_id = 113
name = discontinuous reduplication Ni dopo Ni
function = modifier conveying a succession/iteration/distribution of ref:A
horizontal_links = 114
vertical_links =

giorno dopo giorno
'day after day'

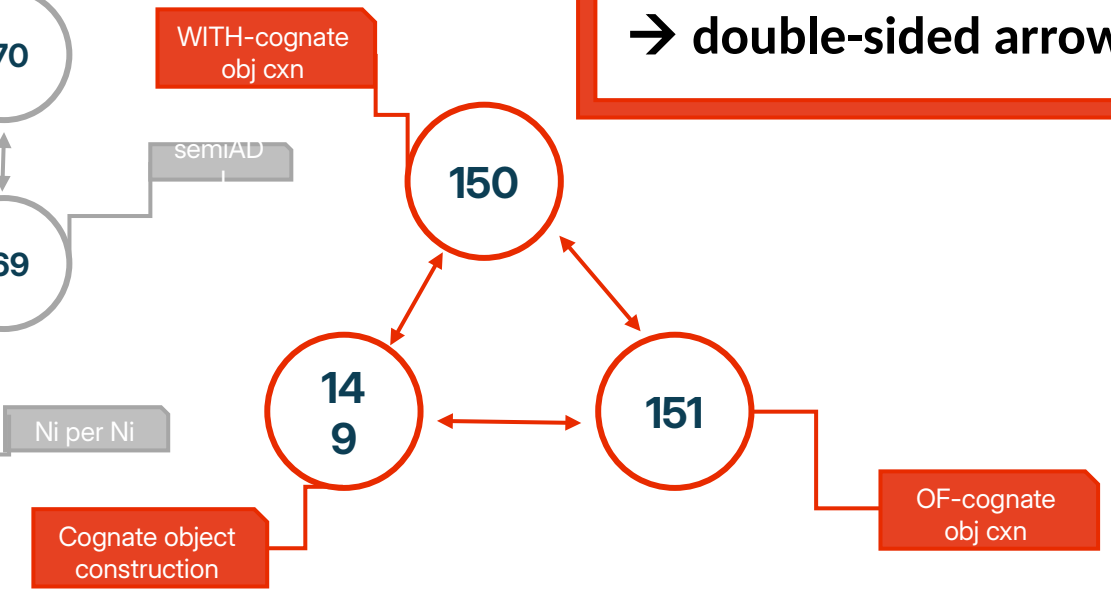
How to represent links between constructions?

VERTICAL LINKS:
one-way connection
→ **one-sided arrow**





HORIZONTAL LINKS:
mutual connection
→ **double-sided arrows**



How to account for formal variability?

(*a*) NP *tocca* V ‘lit. (to) NP touch VP’

- Deontic modality construction (cxn 156)
  - *Definition*: The subject (the Experiencer expressed by the NP) has to carry out an action unwillingly. Such an obligation is the consequence of the State of Affairs.
  - *Restrictions*: The verb *toccare* (lit. ‘touch’) is in the 3rd person singular form and V is in the infinitive form.

156a: with full complement

- *credo che* ***a Davide*** ***toccherà portare l'acqua*** […]

  ‘I think David will have to carry water […]’

156b: with clitic pronoun

- *il lunedì* ***mi*** ***toccava digiunare*** *tutto il giorno* […]

  ‘on Mondays I had to fast all day […]’

→ We include the forms as two different **construction subtypes** in the CoNLL formalization and include examples that account for such a variability

How to account for formal variability?

(*a*) NP *tocca* V

cxn_id = 156a
name = (a) NP tocca V-inf
function = ref:B has to ref:D
horizontal_links =
vertical_links =

ID	UD.FORM	LEMMA	UPOS	FEATS	HEAD	DEPREL	REQUIRED	WITHOUT	SEM_FEATS	SEM_ROLES	ADJACENCY	IDENTITY
A	_	a	ADP	_	B	case	0	_	_	_	_	_
B	_	_	PRON, NOUN, PROPN	Prontype=Prs	C	iobj	0	_	per	Experiencer	A	_
C	_	toccare	VERB	Number=Sing|Person=3	0	root	1	_	contact	_	B	_
D	_	_	VERB	VerbForm=Inf	C	csubj	1	_	_	_	_	_

cxn_id = 156b
name = (a) NP tocca V-inf
function = ref:A has to ref:D
horizontal_links =
vertical_links =

ID	UD.FORM	LEMMA	UPOS	FEATS	HEAD	DEPREL	REQUIRED	WITHOUT	SEM_FEATS	SEM_ROLES	ADJACENCY	IDENTITY
A	_	_	PRON	Clitic=yes	B	iobj	0	_	per	Experiencer	_	_
B	_	toccare	VERB	Number=Sing|Person=3	0	root	1	_	contact	_	A	_
C	_	_	VERB	VerbForm=Inf	B	csubj	1	_	_	_	_	_

From theory to usage...

Create construction

Name

Definition

Formattazioni

Restrictions

Collocational preferences

Formattazioni

CoNLL format

Usage

Select...

Formal tags

Functional tags

Complexity level tags

Category tags

Schematicity Level

Select...

CEFR level

Select...

Horizontal links

Vertical links

Examples 1

Formattazioni

Examples 5

UD examples

Notes

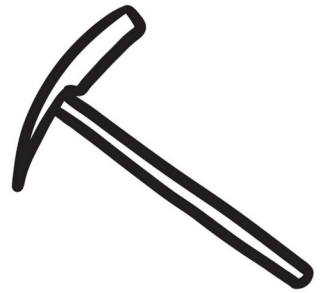
References

Data Collector

Status

submitted

How to mine for constructions (in the wild)?



sent_id = tut-3121

source = http://hdl.handle.net/11234/1-5502/UD_Italian-ISDT/it_isdt-ud-train/tut-3121

text = Vediamo allora qualche criterio per ricomporre o confermare il proprio portafoglio.

1	Vediamo	vedere	VERB	V	Mood=Ind|Number=Plur|Person=1|Tense=Pres|VerbForm=Fin	0	root	0:root	_
2	allora	allora	ADV	B	_	1	advmod	1:advmod	_
3	qualche	qualche	DET	DI	Number=Sing|PronType=Ind	4	det	4:det	_
4	criterio	criterio	NOUN	S	Gender=Masc|Number=Sing	1	obj	1:obj	_
5	per	per	ADP	E	_	6	mark	6:mark	_
6	ricomporre	ricomporre	VERB	V	VerbForm=Inf	4	acl	4:acl:per	_
7	o	o	CCONJ	CC	_	8	cc	8:cc	_
8	confermare	confermare	VERB	V	VerbForm=Inf	6	conj	4:acl:per|6:conj:o	_
9	il	il	DET	RD	Definite=Def|Gender=Masc|Number=Sing|PronType=Art	11	det	11:det	_
10	proprio	proprio	DET	AP	Gender=Masc|Number=Sing|Poss=Yes|PronType=Prs	11	det:poss	11:det:poss	_
11	**portafoglio**	**portafoglio**	**NOUN**	S	Gender=Masc|Number=Sing	8	obj	8:obj	SpaceAfter=No|**CXN=147:A|CXN=168:A**
11.1	**porta**	**portare**	**VERB**	_	Mood=Imp|Number=Sing|Person=2	11	root/m	_	**CXN=147:A.1|CXN=168:A.1**
11.2	**foglio**	**foglio**	**NOUN**	_	Numer=Sing	11.1	case/m	_	**CXN=147:A.2|CXN=168:A.2**
12	.	.	PUNCT	FS	_	1	punct	1:punct	_

Availability and population

- How to access the resource
  - **GitHub** (https://github.com/LaboratorioSperimentale/adoc)
    - § Public repository and wiki
    - § Consultation by experts (linguists, NLP community)
  - **Web application** (not public yet)
    - Future crowdsourcing
    - Consultation by laymen and teachers

- Future population of the resource: several phases and sources
  - Crowdsourcing and citizen science: **Adopt a Construction** (**AdoC**)
    - People (linguists, students, teachers, etc.) will be invited to adopt constructions, which will be reviewed by the AdoC team

Adopt a Construction



The AdoC team



Adopt a Construction

Thank you for your attention!

Comments, suggestions, and **adoptions** are welcome!