Constraining constructions with WordNet: pros and cons for the semantic annotation of fillers in the Italian Constructicon

Flavio Pisciotta
University of Salerno
fpisciotta@unisa.it

Ludovica Pannitto
University of Bologna
ludovica.pannitto@unibo.it

Lucia Busso
Aston University
l.busso@aston.ac.uk

Beatrice Bernasconi
University of Turin
beatrice.bernasconi@unito.it

Francesca Masini
University of Bologna
francesca.masini@unibo.it

Abstract

The paper discusses the role of WordNet-based semantic classification in the formalization of constructions, and more specifically in the semantic annotation of schematic fillers, in the Italian Constructicon. We outline how the Italian Constructicon project uses Open Multilingual WordNet topics to represent semantic features and constraints of constructions.

1 Introduction

In Construction Grammar (CxG, Hoffmann and Trousdale, 2013), the basic units of linguistic description are constructions (cxns), which are conventionalized pairings of form and function (Goldberg, 1995, 2006). Crucially, cxns can vary in complexity and schematicity, including not only words, but also more complex and/or abstract units such as predicative structures, idioms, and word formation processes. CxG, as other usage-based approaches, assume that cxns are not stored as a mere list, but as a structured network (the *Construct-i-con*) in which cxns are linked by different kinds of relationships (Diessel, 2019, 2023).

Despite traditional research in CxG not focusing much on language as a system, recent years have seen a growing interest in *Constructicography*, a blend of "Practical Lexicography" and CxG (Boas et al., 2019). That is, the notion of "Constructicon" has acquired an additional meaning. Beside relating to the structured inventory of all constructions in a language, it has come to indicate a linguistic resource that aims at representing and formalising the network of constructions in a given language (Lyngfelt et al., 2018b).

Constructicography, therefore, is the research field that aims to build Constructicons, that is, to develop repositories of cxns that consistently and coherently describe the grammar (and thus the constructional network) of a specific language. Constructicons already exist for a number of languages (e.g., Janda et al. 2018; Lyngfelt et al. 2018a; Torrent et al. 2018), and are often linked to the FrameNet enterprise (Baker et al., 1998). In fact, Frame Semantics is considered a "sister" framework to CxG - as both theories stem from Fillmore's work on semantic roles (Fillmore, 1968) - and is typically used to represent semantic aspects of constructions (Borin and Lyngfelt, forthcoming).

Despite Constructicography being a fast-growing research area in many languages, Italian has so far been at the periphery of it. Not only there is no Constructicon, but there is also no published Italian FrameNet, although there have been several attempts at developing such a resource (Tonelli et al., 2009; Lenci et al., 2010; Basili et al., 2017). The present contribution introduces the Italian Constructicon (ItCon) project (Masini et al., 2024). This project aims to bridge this gap by building an open and collaborative resource that is designed to be interoperable with existing resources for Italian (treebanks, lexical databases, corpora). Crucially, we outline how we use WordNet-based semantic classification to represent the semantic layer of Italian constructions.

As it stands, the resource is still in its infancy. Therefore, the primary goal so far is to develop a solid theoretical and operational background for the project. In this contribution, we focus specifically on how to constrain the generative power of cxns with respect to the semantic productivity of the open slots of semi-specified cxns (Suttle and Goldberg, 2011; Perek, 2016), and how this problem can be addressed operationally through the integration of data from WordNet(s) available for Italian (Roventini et al., 2000; Pianta et al., 2002) in our annotation format. We will proceed by briefly describing the architecture of ItCon and the annotation format of constructional entries (Section 2), and then we discuss how our annotation scheme can benefit from the connection with WordNet, as well as the possible limitations of such proposals

```
#cxn-id = 171
#cxn = fare Npsych
#function = cause to feel ref:B

ID  UD.FORM  LEMMA  UPOS  FEATS          HEAD  DEPREL
A   _        fare   VERB  _              0     root
B   _        _      NOUN  Number=Sing    A     obj

REQUIRED  WITHOUT                SEM.FEATS                ADJACENCY  IDENTITY
1         _                      _                        _          _
1         CHILDREN:DEPREL=det    OntoClass=feeling        _          _
```

Listing 1: Example of CoNLL-C annotation for the light verb cxn *fare* $N_{feeling}$ 'make feel $N_{feeling}$' (lit. do $N_{feeling}$) (Pisciotta and Masini, forthcoming). Since this construction only occurs with a psychological noun in the singular form, the features of the noun are specified with "number=sing", and the semantic layer uses the topic of "feeling" to constraint the nouns that can occurr in the second slot of the construction.

(Sections 3 and 4).

2 Architecture of the Italian Constructicon

ItCon consists of three linked structures:

- a **database of cxns**;
- the **graph of cxns**, where each node represents a cxn in terms of the set of constraints that it expresses and edges represent horizontal and vertical links holding between cxns;
- a body of **annotated examples** in CoNLL-U format (Nivre et al., 2016), incrementally built by annotating instances of a specific cxn (i.e., constructs) in texts by means of a specific feature in the `MISC` field.

In the **database of cxns**, each entry describes a cxn through a number of text fields and tags. They serve the purpose of specifying information about the properties and behavior of the constructs, as well as linking the database entry to a node in the **graph of cxns** and to a subset of the **annotated examples**.

Each node in the graph of cxns consists of a columnar formalization customized for cxns representation, based on CoNLL-X format (Buchholz and Marsi, 2006) and therefore named CoNLL-C (Masini et al., 2024), that can be converted into a Grew query (Guillaume, 2021) in order to match occurrences of the cxn in CoNLL-U annotated corpora, i.e., Universal Dependencies (UD, Nivre et al. 2016) treebanks. The generative power of the cxn gets constrained at this level, as it is necessary to narrow down the possible set of matched occurrences. This is done through a set of fields specifying formal and functional constraints, which we briefly describe.

2.1 The CoNLL-C format

The CoNLL-C format is a UD compatible format[1]. As shown in Listing 1[2], each formalized cxn is described by a set of metadata (i.e., the lines prefixed by #) that specify holistic properties of the cxn (such as its *semantic function*), and by a number of fields, containing a token-by-token description of the cxn components. The first 7 fields (`ID`, `UD.FORM`, `LEMMA`, `UPOS`, `FEATS`, `HEAD`, `DEPREL`) can be mapped on the matching fields in CoNLL-U format. Since one of the aims of such formalization is to match the relevant constructs in UD-annotated corpora, some other fields were added to formally constrain the queried pattern. They include information such as whether a token is necessarily expressed (`REQUIRED`), the possibility of intervening material within the cxn (`ADJACENCY`), any excluded values (`WITHOUT`), as well as the need for sharing of some features between two tokens (`IDENTITY`).

Taking into account the aforementioned fields, the formalization in Listing 1 can be rewritten in Grew query language (Guillaume, 2021) as follows:

```
pattern {X1 [lemma='fare'];
         X2 [upos=NOUN, Number=Sing];
         X1 < X2;
         X1 -[obj]-> X2}
without {X2 -[det]-> X3}
```

However, such formalization can only partially constrain the set of matching patterns. For instance, by searching the PoSTWITA-UD treebank (Sanguinetti et al., 2018) applying such a query, we obtain both patterns corresponding to *fare* $N_{feeling}$

[1] For a comprehensive description of the format and the relevant fields, see (Pannitto et al., 2024).

[2] The columnar format was split in two lines for space reasons.

'make feel $N_{feeling}$' cxn (1), as well as false positives (2):

(1) *fare schifo* 'to disgust', *fare paura* 'to frighten', *fare piacere* 'to please'

(2) *fare demagogia* 'to be demagogic', *fare parte* 'to be part', *fare cassa* 'to make profit'

Patterns in (2) are not instances of the cxn we want to match: they do not express a causative nor a psychological semantics (since they do not involve nouns expressing psychological states). For such reasons, we added the `SEM.FEATS` field, where semantic features of the tokens filling the empty slots can be specified.

As for now, the semantic features include the semantic class (`OntoClass`) for nouns and verbs, and Aktionsart (`Aktionsart`) for verbs only. Given the need for interoperability with other resources, however, cross-linguistically and cross-resource shared annotation schemes are needed for such features. In the following section, we show how we intend to employ WordNet data to annotate the `OntoClass` semantic feature in our cxns, discussing the advantages and limitations of such an approach.

3 WordNet for semantic classification

As mentioned, one of the semantic features we included in the formalization is `OntoClass`. In this category, we annotate the semantic classes of slots in our cxns using Open Multilingual WordNet (OMW) topics (Bond and Foster, 2013), as currently mapped onto Italian MultiWordNet (Pianta et al., 2002). These topics are the Lexicographer files used by Princeton WordNet (Fellbaum, 1998), and correspond to the top nodes used to build the hierarchy of the four WordNet categories: noun, verb, adjective, and adverb (Miller et al., 1990). Currently, we decided to employ the tagset only for nouns (26 classes) and verbs (15 classes).

We chose to employ OMW topics over developing an original classification for several reasons. Firstly, using OMW topics provides ItCon with an annotation scheme that is cross-linguistically interoperable, and a shared standard. Even though at the moment of writing (January 2025) no other Constructicon annotates semantic constraints on fillers of cxns, we hope that in the future using OMW will provide an easy and theoretically-grounded way to link constructicons.

Secondly, OMW topics have been already used as a semantic classification in sense-tagged corpora[3], which potentially makes ItCon interoperable with other, not CxG-related sense-tagged or WordNet-related resources.

Another advantage of using OMW's ontology is that it includes the hierarchy of synsets, which allows for flexibility in determining the level of granularity needed in tagging semantic constraints case by case, while still relying on a relatively small number of tags[4].

As for now, we found ourselves resorting to such a semantic classification in constraining the matching process of our cxns, although a more systematic testing is necessary to prove its usefulness. For instance, by tagging the noun slot in the *fare* $N_{feeling}$ cxn with the class `noun.feeling` (Listing 1), we are able to exclude most of the false positives in the matching process (cf. 1-2):

(3) Instances of *fare* $N_{feeling}$:

a. *fare* ***schifo***
do.INF disgust.SG
noun.feeling

b. *fare* ***paura***
do.INF fear.SG
noun.feeling

c. *fare* ***piacere***
do.INF pleasure.SG
noun.feeling

(4) False positives:

a. *fare* ***demagogia***
do.INF demagogy.SG
noun.communication

b. *fare* ***parte***
do.INF part.SG
noun.group

c. *fare* ***cassa***
do.INF cash.SG
noun.quantity

3.1 Coverage of Italian Treebanks lexicon

Since the primary aim of our formalization is to map cxns in ItCon to UD-annotated corpora, as a preliminary evaluation of our tagset we checked how many lemmas and how many forms in Italian UD treebanks are associated to at least one synset (and thus, at least one OMW topic) in Italian MultiWordNet. We extracted the frequency lists for

[3] See, for instance, SemCor (Miller et al., 1994) and the subsequent work on multilingually aligning sense-tagged corpora (Bentivogli and Pianta, 2005; Attardi et al., 2010).

[4] The lower number of tags is the reason why we chose OMW topics over EuroWordNet (Rodríguez et al., 1998) top nodes (n = 63).

noun and verb lemmas from Italian treebanks[5], and selected the lemmas with frequency higher than 5 (n = 5273). We then extracted all the synsets and the associated *lexnames* (OMW topics) for each lemma, using `NLTK`[6] WordNet interface in Python to access data from `omw-it 1.4`.

Though not all the lemmas in Italian Treebanks have a corresponding OMW topic, the results are encouraging (Appendix A). Only 10% of the noun lemmas (n = 394) and 12.7% of the verb lemmas (n = 173) are not assigned any semantic tags (Table 3). Moreover, the percentage of untagged nouns and verbs in Italian Treebank gets lower if we look at the forms count (obtained by adding together the frequencies of the lemmas). Namely, only 3.5% of the forms (both for the verbs and for the nouns) is not associated to any topics (Table 4).

Although a broader coverage of the Italian treebanks would be desirable, also considering that we set a strict frequency threshold, these results are promising. In fact, they suggest that a substantial number of constructs can be identified using our semantic annotation.

3.2 Limitations

Nonetheless, using OMW topics as semantic tags can bear some limitations. Firstly, a pre-defined classification does not necessarily include all needed semantic classes, as opposed to a bottom-up classification[7]: using an existing widely used ontology makes adding new, *ad hoc* semantic tags impossible, as it would hinder interoperability with other WordNet-connected resources. Secondly, the semantic classification is only available for nouns and verbs, since there are no top nodes for adverbs and only three top nodes for adjectives (`all`, `participial`, `pertainyms`). Currently, the choice of the semantic tagset for adjectives and adverbs stands as an open challenge: while at least for adjectives some classifications exist (e.g., Dixon 2004), also in the context of some WordNets (e.g., GermaNet, Hamp and Feldweg 1997), they are not mapped onto Italian resources. Thus, while they could be used for descriptive purposes, it would be difficult to employ them consistently in the matching process.

[5] `https://universaldependencies.org/#italian-treebanks` with the exception of Italian-Old (Corbetta et al., 2023), as it is actually a treebank of old Italian, containing Dante Alighieri's *Divine Comedy*.

[6] `https://www.nltk.org/`

[7] See for instance the approach taken in Jezek et al. (2014).

4 Future steps: annotation of inter-slot semantic relations

A challenge for our formalization is represented by the cases in which constraining the fillers of a single slot is not enough in order to match the instances of a cxn. As a matter of fact, the idiosyncratic behaviour of some syntactic and multiword cxns consists in the semantic interdependence of their slots (Desagulier, 2016). Some examples include:

(5) Oxymorons (La Pietra and Masini, 2020)

a. *l'* *ingiustizia* *della* *giustizia*
DET.F.SG injustice.SG of.DET.F.SG justice.SG
noun.attribute noun.attribute
'the injustice of justice'

b. *allegria* *triste*
joy.SG sad.SG
noun.feeling adj.all
'sad joy'

(6) Cognate cxns (Melloni and Masini, 2017; Busso et al., 2020)

a. *vivere* *la* *vita*
live.INF DET.F.SG life.SG
verb.stative noun.state
'to live life'

b. *danzare* *una* *danza*
dance.INF DET.F.SG dance.SG
verb.motion noun.act
'to dance a dance'

For instance, in (5) the two slots are filled by antonymic words, while in (6) the verb and the object are derivationally or semantically related. In such cxns, acknowledging the paradigmatic or semantic relationship between the fillers is necessary in order to define the cxns and to distinguish such instances from other formally similar cxns. Such relations can take place between same-POS fillers (5a) but also between different-POS fillers (5b, 6).

A possible solution could be to use the network structure of WordNet. As a matter of fact, OMW topics are only taken as a semantic classification, since they are top nodes of the hierarchy and no semantic relation is specified among them (let alone cross-POS relations). It should therefore be quite straightforward to constrain the possible fillers by checking if a specific semantic relation between two fillers exists in WordNet's database. This can be implemented through the `IDENTITY` field in the CoNLL-C formalization.

Normally, we employ the IDENTITY field to specify if two or more fillers' fields should have the same value for a given feature. For instance, Table 1 shows how this field is employed in the case of discontinuous reduplication cxns N_i *non* N_i 'N not N' (e.g. *sapone non sapone*, lit. soap not soap, meaning 'soap-free detergent') (Masini and Di Donato, 2023).

ID	UD.FORM	LEMMA	UPOS	...	IDENTITY
A	_	_	NOUN	...	_
B	non	non	ADV	...	_
C	_	_	NOUN	...	UD.FORM=A

Table 1: Partial CoNLL-C formalization of N_i *non* N_i 'N not N' cxn.

However, IDENTITY can easily be adapted to represent same-POS relations by making reference to WordNet relations between synsets: for instance, in a oxymoronic N_1 Prep N_2 cxn such as (5a), the synsets of the two nouns are linked by an antonym relation. This relation could be formalized, so as to remain queryable in WordNet, as:

LEMMA=antonym:N_1

Problems arise in case of different-POS fillers, such as in *Cognate Object cxns*, where the verb and the object are semantically (and often derivationally) related, the object being a shadow argument of the verb. While MultiWordNet does not encompass cross-POS relations, ItalWordNet (Roventini et al., 2000) includes a number of cross-POS relations, inherited from EuroWordNet (Vossen, 1998)[8].

However, by consulting the most recent OMW-compliant version of ItalWordNet[9] (Quochi et al., to appear), such relations seem to be employed only partially. Nonetheless, the behaviour of the constructs in (6) is captured in ItalWordNet: *danzare* 'to dance' is in a similar relation with *danza* 'dance', pointing that the two synsets express similar meanings[10], and the same holds for *vivere* 'to live' and *vita* 'life'.

While being ideally very powerful for our formalization, such an approach needs a wide and consistent coverage of the Italian lexicon and its relations in WordNet. This is needed in order to avoid filtering out possible instances of the cxns if a semantic relation is absent in WordNet. For instance, quite common examples of Cognate cxns (7-8) would be filtered out since the verb and the object bear no relation in ItalWordNet:

(7) *Sara* ***ha*** ***dormito*** *un* ***sonno*** *di piombo*
Sara AUX.3SG sleep.PST DET.M.SG sleep.SG of plumber.SG
'Sara slept a deep sleep.' (Busso et al., 2020)

(8) ***Ho*** ***sognato*** *un* *bel* ***sogno*** *stanotte.*
AUX.1SG dream.PST DET.M.SG beautiful.M.SG dream.SG last_night
'Last night, I dreamed a beautiful dream.'
(adapted from Melloni and Masini 2017)

Nonetheless, for the moment this annotation can still be useful at the descriptive level, since it provides us with a consistent way to annotate constructional properties of our entries in a fine-grained fashion, and will be hopefully exploited in the future for the matching process.

5 Conclusions

The present contribution has outlined how the Italian Constructicon project aims at making lexical and constructional resources interoperable in a fruitful manner. We have shown how WordNet's network structure can be employed to flexibly describe the idiosyncratic behaviour of constructions. The biggest limitations of this approach are practical in nature. In fact, this protocol would work properly only with a greater coverage of OMW semantic classification, together with Italian corpora annotated with (super)senses. Moreover, as ItCon is committed to include morphological (i.e., word-formation) cxns, an unanswered question is whether this semantic classification will prove to be adequate for the annotation of semantic constraints in morphological cxns as well (as for now it has been employed for multiword and syntactic constructions). Despite these open questions, and despite the ItCon project still being in its infancy, we have shown how using Open Multilingual WordNet to represent cxns' semantic features is a fruitful way to link different types of language resources, making them interoperable cross-linguistically.

[8] See Alonge et al. (1998) and Roventini et al. (2000) for a description of the relations.

[9] https://github.com/valeq/IWN-OMW/

[10] Actually, the similar relation was not defined in EuroWordNet, but is part of the Princeton WordNet relations (https://globalwordnet.github.io/schemas/#rdf).

References

Antonietta Alonge, Nicoletta Calzolari, Piek Vossen, Laura Bloksma, Irene Castellon, Maria Antonia Marti, and Wim Peters. 1998. The linguistic design of the EuroWordNet Database. In Piek Vossen,

editor, *EuroWordNet: A multilingual database with lexical semantic networks*, pages 19–43. Springer Netherlands, Dordrecht.

Giuseppe Attardi, Stefano Dei Rossi, Giulia Di Pietro, Alessandro Lenci, Simonetta Montemagni, and Maria Simi. 2010. A resource and tool for Super-sense Tagging of Italian Texts. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

Roberto Basili, Silvia Brambilla, Danilo Croce, and Fabio Tamburini. 2017. Developing a large scale FrameNet for Italian: the IFrameNet experience. In Roberto Basili, Malvina Nissim, and Giorgio Satta, editors, *Proceedings of the Fourth Italian Conference on Computational Linguistics CLiC-it 2017*, page 59–64. Accademia University Press.

Luisa Bentivogli and Emanuele Pianta. 2005. Exploiting parallel texts in the creation of multilingual semantically annotated resources: the MultiSemCor corpus. *Natural Language Engineering*, 11(3):247–261.

Hans C. Boas, Benjamin Lyngfelt, and Tiago Timponi Torrent. 2019. Framing constructicography. *Lexicographica*, 35(2019):41–85.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual WordNet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362, Sofia, Bulgaria. Association for Computational Linguistics.

Lars Borin and Benjamin Lyngfelt. forthcoming. FrameNets and ConstructiCons. *The Cambridge Handbook of Construction Grammar*.

Sabine Buchholz and Erwin Marsi. 2006. Conll-X shared task on multilingual dependency parsing. In *Proceedings of the tenth conference on computational natural language learning (CoNLL-X)*, pages 149–164.

Lucia Busso, Alessandro Lenci, and Florent Perek. 2020. Valency coercion in Italian: An exploratory study. *Constructions and Frames*, 12(2):171–205.

Claudia Corbetta, Marco Passarotti, Flavio Massimiliano Cecchini, and Giovanni Moretti. 2023. Highway to Hell. Towards a Universal Dependencies Treebank for Dante Alighieri's Comedy. In *Proceedings of CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy*.

Guillaume Desagulier. 2016. A lesson from associative learning: asymmetry and productivity in multiple-slot constructions. *Corpus Linguistics and Linguistic Theory*, 12(2):173–219.

Holger Diessel. 2019. *The Grammar Network: How Linguistic Structure Is Shaped by Language Use*. Cambridge University Press.

Holger Diessel. 2023. *The Constructicon: Taxonomies and Networks*. Elements in Construction Grammar. Cambridge University Press.

Robert M. W. Dixon. 2004. Adjective Classes in Typological Perspective. In Robert M. W. Dixon and Alexandra Y. Aikhenvald, editors, *Adjective Classes: A Cross-Linguistic Typology*. Oxford University Press.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press.

Charles J. Fillmore. 1968. The case for case. *Universals in Linguistic Theory*.

Adele Goldberg. 1995. *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

Adele Goldberg. 2006. *Constructions at Work: The Nature of Generalization in Language*. Oxford University Press.

Bruno Guillaume. 2021. Graph matching and graph rewriting: GREW tools for corpus exploration, maintenance and conversion. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 168–175, Online. Association for Computational Linguistics.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet - a lexical-semantic net for German. In *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*.

Thomas Hoffmann and Graeme Trousdale, editors. 2013. *The Oxford Handbook of Construction Grammar*. Oxford University Press.

Laura A. Janda, Olga Lyashevskaya, Tore Nesset, Ekaterina Rakhilina, and Francis M. Tyers. 2018. A constructicon for Russian: Filling in the gaps. In Benjamin Lyngfelt, Lars Borin, Kyoko Ohara, and Tiago Timponi Torrent, editors, *Constructicography: Constructicon development across languages*, page 165–182. John Benjamins Publishing Company.

Elisabetta Jezek, Bernardo Magnini, Anna Feltracco, Alessia Bianchini, and Octavian Popescu. 2014. T-PAS; a resource of typed predicate argument structures for linguistic analysis and semantic processing. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 890–895, Reykjavik, Iceland. European Language Resources Association (ELRA).

Marta La Pietra and Francesca Masini. 2020. Oxymorons: a preliminary corpus investigation. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 176–185, Online. Association for Computational Linguistics.

Alessandro Lenci, Martina Johnson, and Gabriella Lapesa. 2010. Building an Italian FrameNet through semi-automatic corpus analysis. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Benjamin Lyngfelt, Linnéa Bäckström, Lars Borin, Anna Ehrlemark, and Rudolf Rydstedt. 2018a. Constructicography at work: Theory meets practice in the Swedish constructicon. In Benjamin Lyngfelt, Lars Borin, Kyoko Ohara, and Tiago Timponi Torrent, editors, *Constructicography: Constructicon development across languages*, pages 41–106. John Benjamins Publishing Company.

Benjamin Lyngfelt, Lars Borin, Kyoko Ohara, and Tiago Timponi Torrent, editors. 2018b. *Constructicography: Constructicon development across languages*. John Benjamins Publishing Company.

Francesca Masini, Beatrice Bernasconi, Claudia Borghetti, Lucia Busso, Maria Pina De Rosa, Claudio Iacobini, M. Silvia Micheli, Ludovica Pannitto, Flavio Pisciotta, and Fabio Tamburini. 2024. Towards an Italian Constructicon. In *The 13th International Conference on Construction Grammar (ICCG13)*, Göteborg (Sweden), 26–28 August 2024.

Francesca Masini and Jacopo Di Donato. 2023. Non-prototypicality by (discontinuous) reduplication: the N-*non*-N construction in italian. *Zeitschrift für Wortbildung / Journal of Word Formation*, 7(1):130–155.

Chiara Melloni and Francesca Masini. 2017. Cognate constructions in Italian and beyond: A lexical semantic approach. In *Contrastive Studies in Verbal Valency*, page 220–250. John Benjamins Publishing Company.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to WordNet: An On-line Lexical Database*. *International Journal of Lexicography*, 3(4):235–244.

George A. Miller, Martin Chodorow, Shari Landes, Claudia Leacock, and Robert G. Thomas. 1994. Using a semantic concordance for Sense Identification. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Ludovica Pannitto, Beatrice Bernasconi, Lucia Busso, Flavio Pisciotta, Giulia Rambelli, and Francesca Masini. 2024. Annotating Constructions with UD: the experience of the Italian Constructicon. In *UniDive 3rd general meeting*, Hungarian Research Centre for Linguistics, Budapest (Hungary), 29–30 January 2025.

Florent Perek. 2016. Using distributional semantics to study syntactic productivity in diachrony: A case study. *Linguistics*, 54(1):149–188.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. MultiWordNet: developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, pages 293–302, Mysore, India. Global WordNet Association.

Flavio Pisciotta and Francesca Masini. forthcoming. A paradigm of psych-predicates: unraveling the constructional competition between light verb constructions and derived verbs in italian. In Anna Riccio and Jens Fleischhauer, editors, *Light verbs: synchronic and diachronic studies*. Düsseldorf University Press.

Valeria Quochi, Roberto Bartolini, and Monica Monachini. to appear. ItalWordNet goes open. In *LiLT Special Issues on Open Multilingual WordNets*. CSLI Publications.

Horacio Rodríguez, Salvador Climent, Piek Vossen, Laura Bloksma, Wim Peters, Antonietta Alonge, Francesca Bertagna, and Adriana Roventini. 1998. The Top-Down strategy for building EuroWordNet: vocabulary coverage, base concepts and top ontology. In Piek Vossen, editor, *EuroWordNet: A multilingual database with lexical semantic networks*, pages 45–80. Springer Netherlands, Dordrecht.

Adriana Roventini, Antonietta Alonge, Nicoletta Calzolari, Bernardo Magnini, and Francesca Bertagna. 2000. ItalWordNet: a large semantic database for Italian. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece. European Language Resources Association (ELRA).

Manuela Sanguinetti, Cristina Bosco, Alberto Lavelli, Alessandro Mazzei, Oronzo Antonelli, and Fabio Tamburini. 2018. PoSTWITA-UD: an Italian Twitter treebank in Universal Dependencies. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Laura Suttle and Adele Goldberg. 2011. The partial productivity of constructions as induction. *Linguistics*, 49(6):1237–1269.

Sara Tonelli, Daniele Pighin, Claudio Giuliano, and Emanuele Pianta. 2009. Semiautomatic development of FrameNet for italian. In *Proceedings of the FrameNet Workshop and Masterclass, Co-located with the Seventh International Workshop on Treebanks and Linguistic Theories (TLT8)*, Milan, Italy. EDUcatt.

Tiago Timponi Torrent, Ely Edison da Silva Matos, Ludmila Meireles Lage, Adrieli Laviola, Tatiane da Silva Tavares, Vânia Gomes de Almeida, and Natália Sathler Sigiliano. 2018. Towards continuity between the lexicon and the constructicon in FrameNet Brasil. In Benjamin Lyngfelt, Lars Borin, Kyoko Ohara, and Tiago Timponi Torrent, editors, *Constructicography: Constructicon development across languages*, page 107–140. John Benjamins Publishing Company.

Piek Vossen, editor. 1998. *EuroWordNet: A multilingual database with lexical semantic networks*. Springer, Dordrecht, Netherlands.

A Coverage

class	n. lemmas	n. forms
noun.Tops	58	9153
noun.artifact	1744	31035
noun.act	1566	45486
noun.person	1338	22933
noun.communication	1211	40686
noun.attribute	862	25920
noun.cognition	805	35952
noun.state	714	24911
noun.group	525	30015
noun.event	366	9797
noun.substance	279	3809
noun.location	267	13602
noun.possession	261	12188
noun.animal	251	3457
noun.object	237	5300
noun.feeling	235	4024
noun.body	231	6398
noun.quantity	215	7920
noun.food	208	1651
noun.time	203	14581
noun.plant	200	1559
noun.phenomenon	141	5869
noun.relation	116	5561
noun.process	112	3314
noun.shape	89	2408
noun.motive	20	1329
verb.change	552	16761
verb.communication	550	20294
verb.contact	477	11470
verb.social	414	15671
verb.motion	291	10265
verb.cognition	273	12995
verb.possession	260	11749
verb.stative	259	15682
verb.creation	210	8864
verb.body	168	4437
verb.emotion	158	3547
verb.competition	119	4594
verb.consumption	80	4099
verb.perception	27	6782
verb.weather	27	397

Table 2: Count of lemmas and forms (nouns and verbs only) for each OMW topic (a lemma can belong to more than one topic).

POS	0	1	2	3	4	5	6	7	8	9	11	Total
noun	394 (10,1%)	1882 (48,2%)	913 (23,4%)	429 (11,0%)	172 (4,4%)	83 (2,1%)	18 (0,5%)	9 (0,2%)	4 (0,1%)	3 (0,1%)	0 (0%)	3907 (100,0%)
verb	173 (12,7%)	503 (36,8%)	369 (27,0%)	159 (11,6%)	89 (6,5%)	43 (3,2%)	20 (1,5%)	4 (0,3%)	3 (0,2%)	1 (0,1%)	2 (0,2%)	1366 (100,0%)
Total	567 (10,8%)	2385 (45,2%)	1282 (24,3%)	588 (11,2%)	261 (5,0%)	126 (2,4%)	38 (0,7%)	13 (0,3%)	7 (0,1%)	4 (0,1%)	2 (0,0%)	5273 (100,0%)

Table 3: Counts and percentages and of noun and verb lemmas by number of OMW topics in Italian Treebanks.

POS	0	1	2	3	4	5	6	7	8	9	11	Total
noun	5388 (3,5%)	49610 (31,9%)	40024 (25,8%)	23757 (15,3%)	18327 (11,8%)	12270 (7,9%)	2263 (1,5%)	1340 (0,9%)	850 (0,6%)	1483 (1,0%)	0 (0%)	155312 (100,0%)
verb	2449 (3,5%)	14439 (20,4%)	15196 (21,4%)	12276 (17,3%)	8663 (12,2%)	4161 (5,9%)	3529 (5,0%)	1197 (1,7%)	2995 (4,2%)	416 (0,6%)	5595 (7,9%)	70916 (100,0%)
Total	7837 (3,5%)	64049 (28,3%)	55220 (24,4%)	36033 (15,9%)	26990 (11,9%)	16431 (7,3%)	5792 (2,6%)	2537 (1,1%)	3845 (1,7%)	1899 (0,8%)	5595 (2,5%)	226228 (100,0%)

Table 4: Counts and percentages of noun and verb forms by number of OMW topics in Italian Treebanks.