

PARAD-it: Eliciting Italian Paradigmatic Relations with Crowdsourcing

Irene Sucameli
University of Pisa
Pisa, Italy

irenesucameli@gmail.com

Alessandro Lenci
University of Pisa
Pisa, Italy

alessandro.lenci@unipi.it

Abstract

English. In this paper, we present a new dataset of semantically related Italian word pairs. The dataset consists of nouns, adjectives and verbs together with their synonyms, antonyms and hypernyms. The data have been collected with crowdsourcing from a pool of Italian native speakers. The dataset, the first of its kind, is useful not only to evaluate computational models of Italian semantic relations, but also for linguistic and psycholinguistic investigations of the mental lexicon.

Italiano. *In questo articolo si presenta un nuovo dataset di parole italiane legate da relazioni semantiche. L'analisi si basa su una raccolta di nomi, verbi e aggettivi a cui sono stati associati sinonimi, antonimi e iperonimi. I dati sono stati raccolti da un gruppo di parlanti nativi di italiano tramite crowdsourcing. Il dataset, primo del suo tipo, è utile per valutare modelli computazionali relativi alle relazioni semantiche dell'italiano, per la ricerca linguistica teorica e psicolinguistica.*

1 Introduction

The present project aims at providing new data about the internal organization of the Italian lexicon. For this purpose, we present PARAD-it¹ a paradigmatic relation dataset elicited from Italian native speakers with crowdsourcing. This dataset consists of a set of target words selected from the Italian section of MultiWordNet paired with relation belonging to different kinds of paradigmatic

semantic relations. The data have been collected using the same method adopted by Scheible and Schulte im Walde (2014) for German and by Benotto (2015) for English, thereby making the three datasets fully comparable for crosslingual analyses. PARAD-it is a collection of hypernyms, antonyms, and synonyms for a set of Italian nouns, adjectives and verbs.

2 Related Works

Our contribution is just the latest in a series of recent works aimed at eliciting judgments about semantic relations, to develop testsets for computational models. Besides Scheible and Schulte im Walde (2014) and Benotto (2015), we can mention BLESS, realized by Baroni and Lenci (Baroni and Lenci, 2011). Bless is a dataset created for the evaluation of distributional semantic models. The BLESS dataset includes 200 English nouns, equally divided into animate and inanimate entities. Each noun is associated to multiple relation belonging to five types of relations: hyperonymy, co-hyponymy, meronymy, attributes and events.

Another relevant project is EVALution. This dataset combines data extracted from Concept-Net 5.0 (Liu and Singh, 2004) and WordNet 4.0 (Fellbaum, 1998), and then checked by native speakers. The crowdsourcing task consisted in rating the truthfulness of sentences generated from the selected word pairs, according to templates indicative of various semantic relations and to be used as a proxy for the prototypicality of the relations. PARAD-it extends this line of research to Italian for the first time.

3 Collecting PARAD-it

3.1 Target Selection

The PARAD-it targets were extracted from the Italian section of the MultiWordNet database (Pianta, Bentivogli and Girardi, 2002).

¹ PARAD-it is freely distributed and it will be available for download from:
<http://colinglab.humnet.unipi.it/resources/>

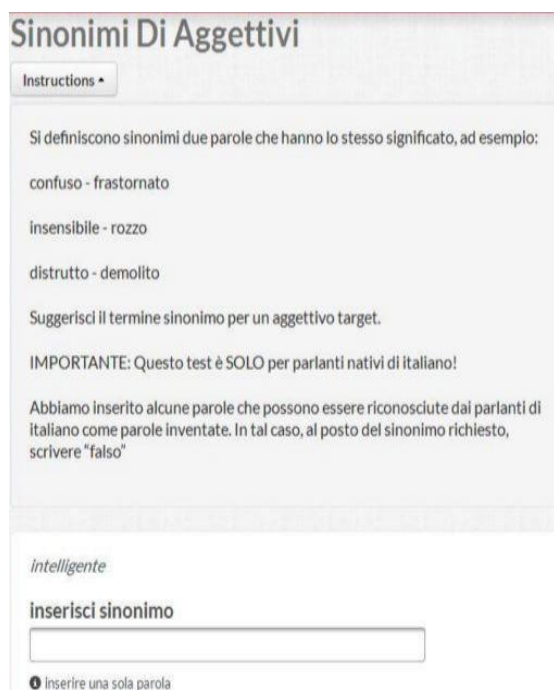
The selection of nouns, adjectives and verbs was balanced for:²

- **Frequency** - three frequency classes were identified using the itWaC corpus (Baroni et al. 2009): i.) words with frequency from 200 to 2999, ii.) words with frequency from 3,000 to 9,999, and iii.) words with frequency greater than 10,000.
- **Polysemy** - three polysemy classes were identified, according to the number of synsets in MultiWordNet: i.) words with one synset, ii.) words with two synsets, iii.) words with three or more synsets.

Then, 11 targets were randomly sampled for each class, making a total of 99 targets for each PoS.

3.2 Data Elicitation

Italian native speakers were asked to produce, for each target word, a synonym, an antonym and a hypernym. The data were collected through CrowdFlower,³ a crowdsourcing web-based platform to design various data collection tasks (i.e., sentiment analysis, data categorization, etc.) thanks to the help of external workers which are paid according to the type of task.



Sinonimi Di Aggettivi

Instructions •

Si definiscono sinonimi due parole che hanno lo stesso significato, ad esempio:

confuso - frastornato

insensibile - rozzo

distrutto - demolito

Suggerisci il termine sinonimo per un aggettivo target.

IMPORTANTE: Questo test è SOLO per parlanti nativi di italiano!

Abbiamo inserito alcune parole che possono essere riconosciute dai parlanti di italiano come parole inventate. In tal caso, al posto del sinonimo richiesto, scrivere "falso"

intelligente

inserisci sinonimo

Figure 1: Example of CrowdFlower form

In the present project, we collected data from ten subjects, for each target word, and for each semantic relation. In order to guarantee that the tasks would be completed only by Italian native speakers, the CrowdFlower form also included a test to discriminate Italian words from “pseudo words”. The responses produced by subjects that failed to pass the test were excluded. All the elicited data were then manually normalised: Typing errors were corrected and the words written in lower case and capital letters were mapped onto a single standard form.

3.3 Results

The number of responses for each PoS and each relation type is shown in Table 1. The lowest number of responses concerns mainly antonyms and then hypernyms. This is due to the fact that antonyms are characterized by a high degree of canonicity (Paradis and Willners 2011, de Weijer et al. 2012). For this very reason, it may be more difficult for a speaker to provide an antonym for a input word since he can rely only on a small group of possible answers.

Compared to antonyms and hypernyms, synonyms are more easily identified by users. In fact, 2,674 tokens have been provided for this paradigmatic relation. However, if we consider the number of types, instead of the number of tokens, the situation is different. In fact, with 1,528 types, the relation of hypernymy is the one with the highest number of types produced. This result shows that, even if for the users it is simpler to provide a synonym for a given target, words have in general a lower number of distinct synonyms. On the other hand, the users have provided less responses for the hypernyms but more differentiated. This might be due to the fact that taxonomies (typical of hypernyms) have different levels of depth (Murphy, 2010). Concerning the target PoS, verbs have elicited the highest number of responses, possibly because of their inherent higher polysemy (Murphy, 2010). These results regarding the identification of verbs and hypernyms by native speakers are in line with those obtained by Scheible and Schulte im Walde for German and with those produced by Benotto for English.

² The balancing parameters are the same used by Scheible and Schulte im Walde (2014) and by Benotto (2015).

³ <https://www.crowdfLOWER.com>

	ANT		HYP		SYN		all	
	types	tokens	types	tokens	types	tokens	types	tokens
Adj	269	805	435	706	455	853	1159	2364
Noun	306	493	570	843	453	883	1329	2219
Verb	444	849	523	915	466	938	1433	2702
all	1019	2147	1528	2464	1374	2674	3921	7285

Table 1: Number of total responses

	ANT+SYN		HYP+SYN		ANT+HYP		ANT+HYP+SYN	
	types	tokens	types	tokens	types	tokens	types	tokens
Adj	3	15	182	883	3	27	0	0
Noun	48	195	109	541	35	140	21	147
Verb	55	243	214	916	45	208	39	330
all	106	453	505	2340	83	357	60	447

Table 2: Ambiguous responses

As an additional level of analysis, we have identified the ambiguous responses (Table 2). When users have provided the same response for different paradigmatic relation, that response has been considered as ambiguous. Here, the highest number of ambiguity has been recorded in relation to the synonymy-hypernymy pair. Actually, this high number of ambiguity was expected and the result seems to be reasonable since it is similar to the one obtained by Scheible and Schulte im Walde for German (with 470 types recorded as ambiguous within the couple synonymy-hypernymy). This result may depend on the fact that in many cases the distinction between synonymy and hypernymy is blurred or not easily identifiable, especially for more abstract items. For instance, the target *mattino* ('morning') has prompted the word *giorno* ('day') both as synonym and as hypernym.

Concerning the different responses provided by subjects (Figure 2), we saw that a) speakers are mostly in agreement referring to the relation of antonymy, consistently with the trend in the parallel English and German data; b) only in few cases more than 7 different responses have been provided for the same input, while c) in most

cases between 3 and 5 different responses have been indicated for target.

This suggests that Italian native speakers do not tend to have one-to-one lexical associations. At the same time, they tend to identify a reduced group of terms that can be used with a certain relation.

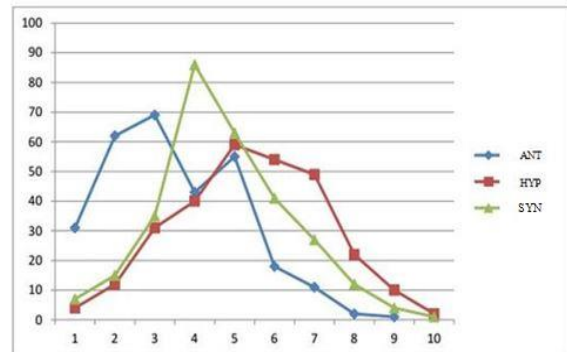


Figure 2: Targets for different responses.

The Y axis reports the number of targets provided by users while the X axis reports the number of different responses per input

Figure 3 and Figure 4 show the production of frequency distribution among classes and relations.

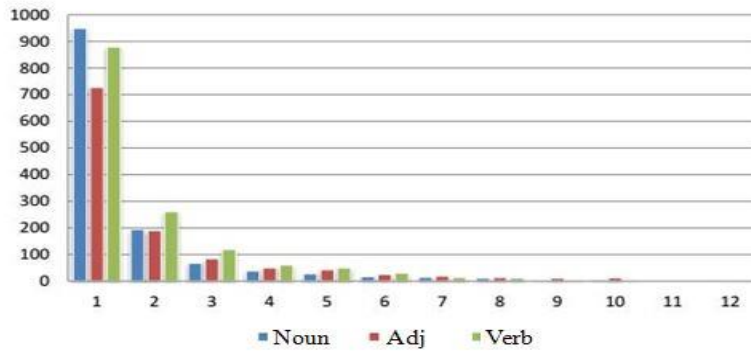


Figure 3: Distribution of production frequency among classes

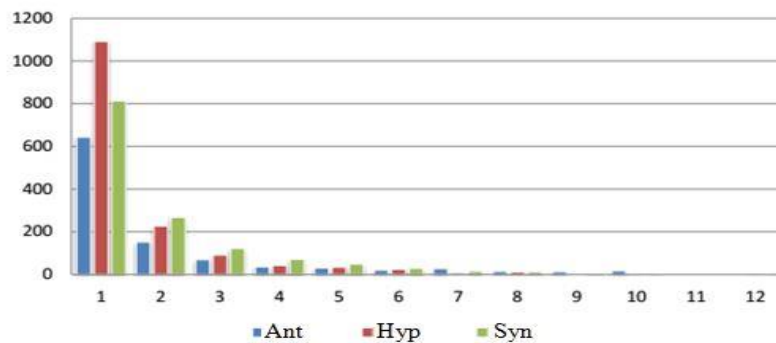


Figure 4: Distribution of production frequency among relations

Concerning the distribution among classes, 949 nouns have been produced by users only once. On the other hand, verbs have 879 hapax responses, and adjectives 727. From Figure 4, it is possible to observe that hypernyms have the highest number of hapax. In fact, for this relation there are 1,090 hapax, while synonymy has 812 hapax and antonymy only 643. This result is due to the existence of canonicity relations for antonymy, and to the notorious paucity of true synonyms.

3.4 Distributional Semantic Analysis of the Elicited Data

A distributional space has been built in order to analyse the synonyms, antonyms and hypernyms produced by subjects. Distributional Semantic Models (DSMs) use corpus co-occurrences to measure the similarity/relatedness between two words: The closer two vectors are in distributional space, the more semantically related the two words are.

We used DISSECT (DIStributIonal SEMantic Composition Toolkit) to train a standard count-based DSM on the *Repubblica* corpus, a corpus made up of newspaper articles with over 300 million tokens. Our targets and contexts include

the PARAD-it data plus all the content words in *Repubblica* with frequency greater than 200. Co-occurrences have been extracted, using a context window of 2 content words to the left and right of each target item. For each PARAD-it relatum, we measured its cosine with the target word, using PPMI (Positive Pointwise Mutual Information) as weighting scheme, and truncated SVD (Singular Value Decomposition) to 300 latent dimensions. Figure 5 and Figure 6 report the boxplot summarizing the cosine distribution by semantic relation and by PoS.

The analysis shows that there are no significant differences in the cosine median neither between different types of relations nor between different grammatical classes. As shown in Figure 5, the highest cosine values have been recorded for antonyms (over 0.90). This is due to the fact that this type of relation is characterized by a high rate of canonicity. On the other side, hypernyms show the greatest median values (0.76).

Concerning the distribution of relata cosine by PoS, nouns have the highest cosine values, while adjectives and verbs show a more reduced variability. These results are coherent with the production data. Indeed, as we saw above, high frequency values were recorded both for nouns and hypernyms while speakers' production show

a greater homogeneity in responses for the relation of antonymy.

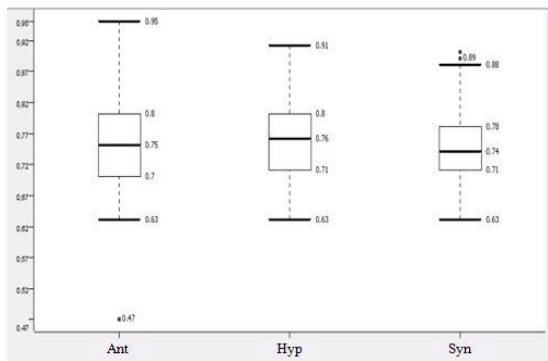


Figure 5: Distribution of relata cosines by relations

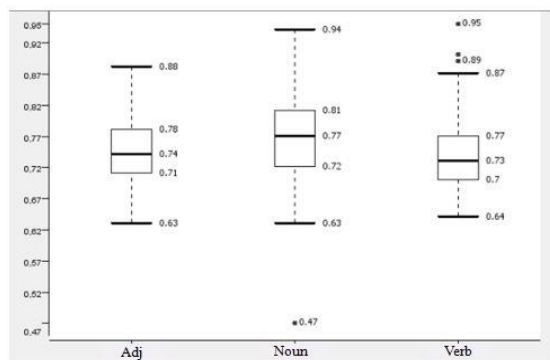


Figure 6: Distribution of relata cosines by target PoS

3 Conclusion

This project presents PARAD-it, a new collection composed by pairs of Italian nouns, verbs and adjectives related by different types of paradigmatic relations, elicited by native speakers with crowdsourcing. Starting from this new resource, a quantitative analysis was carried out to analyze the mechanisms underlying the Italian language. In particular, the analysis has shown that: i) high frequency values tend to be recorded for nouns and hypernyms while ii) Italian speakers tend to use a more uniform vocabulary to describe the relation of antonymy. This analysis has revealed some interesting differences in the response distribution both with respect to the PoS of the target, and with respect to the semantic relation. Moreover, this study confirms the differential salience of the various paradigmatic relations in organizing the mental lexicon.

To the best of our knowledge, PARAD-it is the first, freely available resource of this kind for Italian, paving the way for its use as a test set for computational models of semantic relation identification and classification. For future research,

we plan to realize an additional round of crowdsourcing in order to validate the words previously produced, checking also if there is an overlap between these words and the targets from MultiWordNet. Moreover, we plan to carry out a crosslingual comparison with the similar datasets collected for German and for English.

References

- Marco Baroni, Silvia Bernardini, Federica Comastri, Lorenzo Piccioni, Alessandra Volpi, Guy Aston, Marco Mazzoleni. 2004. Introducing the La Repubblica Corpus: A Large, Annotated, TEI(XML)-Compliant Corpus of Newspaper Italian. *LREC, European Language Resources Association*.
- Marco Baroni, Silvia Bernardini, Adriano Ferraresi, Eros Zanchetta. 2009. The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. In *Language Resources and Evaluation*, Vol. 43, Issue 3, pp.209-226.
- Marco Baroni, Alessandro Lenci. 2011. How we BLESSED distributional semantic evaluation. *Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics*, pp.1-10. Association for Computational Linguistics.
- Giulia Benotto. 2015. *Distributional Models for Semantic Relations: A study on Hyponymy and Antonymy*. PhD thesis Pisa, Italia.
- Christiane Fellbaum. 1998. *WordNet: An electronic lexical database*. MIT Press, Cambridge.
- Hugo Liu, Push Singh. 2004. ConceptNet: A Practical Commonsense Reasoning Toolkit. *BT Technology Journal* Vol. 22 pp. 211-226. Kluwer Academic Publishers.
- M. L. Murphy. 2010. *Lexical Meaning*. Cambridge University Press, Cambridge.
- Carita Paradis, Caroline Willners. 2011. Antonymy: From convention to meaning-making. *Review of cognitive linguistics*, pp.367-391.
- Emanuele Pianta, Luisa Bentivogli, Christian Girardi. 2002. MultiWordNet: developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet* pp.293-302. Mysore, India.
- Adriana Roventini, Antonietta Alonge, Francesca Bertagna, Nicoletta Calzolari, Christian Girardi, Bernardo Magnini, Rita Marinelli, and Antonio Zampolli. 2003. Italwordnet: building a large semantic database for the automatic treatment of italian. *Computational Linguistics in Pisa*, Special Issue, Vol. 18-19, 2:745-791, IEPI, Pisa-Roma.
- Entico Santus, Alessandro Lenci, Frances Young, e Chu-Ren Huang. 2015. EVALution 1.0: an Evolv-

ing Semantic Dataset for Training and Evaluation of Distributional Semantic Models. In *Proceedings of the 4th Workshop on Linked Data in Linguistics*, pp. 64-69. Beijing.

Silke Scheible, Sabine Schulte im Walde. 2014. A Database of Paradigmatic Semantic Relation Pairs for German Nouns, Verbs, and Adjectives. In *Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing*, pp. 111-119. Dublin, Ireland.

Joost van de Weijer, Carita Paradis, Caroline Willners & Magnus Lindgren. 2012. *As lexical as it gets: the role of co-occurrence of antonyms in a visual lexical decision experiment*. In D. Divjak & St. Th. Gries (Eds.). *Frequency effects in language: linguistic representation*. De Gruyter Mouton. pp. 255-279.