Acquiring Opposition Relations among Italian Verb Senses using Crowdsourcing

Anna Feltracco[1,2], Simone Magnolini[1,3], Elisabetta Jezek[2], Bernardo Magnini[1]
[1]Fondazione Bruno Kessler, Via Sommarive 18, 38100 Povo-Trento, Italy
[2]University of Pavia, Strada Nuova 65, 27100 Pavia, Italy
[3]University of Brescia, Piazza del Mercato 15, 25121 Brescia, Italy
feltracco@fbk.eu, magnolini@fbk.eu, jezek@unipv.it, magnini@fbk.eu

Abstract

We describe an experiment for the acquisition of opposition relations among Italian verb senses, based on a crowdsourcing methodology. The goal of the experiment is to discuss whether the types of opposition we distinguish (i.e. *complementarity*, *antonymy*, *converseness* and *reversiveness*) are actually perceived by the crowd. In particular, we collect data for Italian by using the crowdsourcing platform CrowdFlower. We ask annotators to judge the type of opposition existing among pairs of sentences -previously judged as opposite- that differ only for a verb: the verb in the first sentence is opposite of the verb in second sentence. Data corroborate the hypothesis that some opposition relations exclude each other, while others interact, being recognized as compatible by the contributors.

Keywords: opposition relations, verb sense, crowdsourcing, T-PAS resource

1. Introduction

Traditionally, the linguistic study of semantic relations among verbs or verb senses has focused on the manner relation (also known as troponymy, e.g. *move* and *walk*), the cause relation (*kill* and *die*), and, more generally the relation of lexical entailment (*snore* and *sleep*) - which, according to the classification in Fellbaum (1998), subsumes several relation types (such as cause). In the computational field, several initiatives have proposed schemes for the annotation of both the internal structure of the events encoded in verbs (see, for instance, Aguilar et al. (2014), Fokkens et al. (2013)) and the relations among events, including temporal relations as proposed in the TimeML scheme (Pustejovsky et al., 2003). Less works have systematically addressed the relation of opposition between verbs or verb senses (*increase* and *decrease*) and its annotation in existing lexicons or sense repertoires.

In this paper we describe an experiment we run to acquire opposition relations among Italian verb senses, based on a crowdsourcing methodology. In the experiment, we assume a four-output classification of opposition types, namely *complementarity*, *antonymy*, *converseness* and *reversiveness*.

We rely on an existing repository of verb frames for Italian (see Section 5.1), and are interested to verify how the different opposition types are perceived by the crowd and whether it is possible for a pair of verb senses to have characteristics that belong to more than one type. The potential output of the work is a gold-standard of verb frame pairs annotated with opposition relations.

The paper is organized as follows. Section 2 introduces our notion of opposition and the classification we use in our work and Section 3 presents examples of related works using the crowdsourcing methodology. In Section 4 the methodology we adopt is described, followed by the experimental settings in Section 5. Finally, in Section 6 the results are discussed while Section 7 provides some conclusions and directions for future work.

2. Opposition Relations

Our notion of opposition is based on the study of Lyons (1977) and Cruse (1986; 2002; 2011) - as synthesized in Jezek (2016)- , and includes pairs of terms that "typically differ along only one dimension of meaning: in respect of all other statements they are identical" (Cruse, 1986, p.197). Examples include the following pairs: *to open / to close, to rise / to fall*. Opposites cannot be true simultaneously of the same entity, for example a price cannot be said to rise and fall at exactly the same point in time.

Among the various types of oppositions that can be said to exist among verbs, we focus on complementarity, antonymy, converseness and reversiveness, which have been discussed at length in the literature, with some points of divergence regarding the latter.

Complementaries are opposites that "divide some conceptual domain in two mutual exclusive counterparts, so that what does not fall into one of the compartments must necessarily fall into the other" (Cruse, 1986, p.198). The distinction between them is binary, thus there is not an intermediate degree between them: e.g. *to pass / to fail* (an examination).

Antonyms have the characteristic of being gradual from a conceptual point of view. Two antonyms, therefore, oppose each other in relation to a scale of values for a given property, of which they may specify the two poles (or bounds), e.g. *to like / to dislike* (a person).

Converses (or relational opposites) describe the same action from a different perspective: e.g. *to give / to receive* (a present).

Finally, *reversives* denote a change (literal movement or abstract change) in opposite direction between two states: one term indicates a change from a state to another and viceversa: e.g. *to build / to destroy* a building or *to wrap / to unwrap* an object. This category includes the group of *restitutives* (Cruse, 1986), cases in which, according to Chklovski and Pantel (2004), the opposition relation systematically interacts with the happens-before relation, as for the pairs *to*

damage / to repair. Also Fellbaum (1998) has noted that the relation between the verbs in these pairs seems one of entailment (Fellbaum, 1998, p.75); for example one can only unwrap something which has been previously wrapped.
To distinguish different types of opposition is relevant for several reasons. Consider for example the following sentences, where according to our definition *to remain - to leave* are *complementaries* and *to increase - to decrease* are *antonyms*:

(1) (a) John did not remain at home;
(b) John did not leave home;
(c) The price did not increase;
(d) The price did not decrease;

In these examples, (a) and (b) contradict each other, but this is not the case for (c) and (d). In fact, while (a) implicitly deny the truth of (b) (i.e. John actually left his home as he did not remain in it), (c) does not coincide with the negation of (d) and the two sentence can be jointly stated (i.e. The price did neither increase nor decrease). To know whether a pair of verb senses is holding an opposition relation and to know which type of opposition relation they hold is thus relevant to understand the relation between the events they describe.
A point which, to our knowledge, has not been explored systematically so far is whether opposition relations are exclusive, that is, whether it is possible for a pair of verb senses to have characteristics that belong to more than one type of opposition. We expect our experiment to provide us with valuable insights in this regard.

3. Crowdsourcing: Related Works

As mentioned in Section 1, for the acquisition of opposition relations, we collected data using crowdsourcing, a methodology already used to acquire information on large scale which can be used to enrich lexical resources.
In order to design properly the task we have tried to follow the best practices suggested in Sabou et al. (2014)[1], in particular for what concerns, e.g.:

- project definition: main task need to be decomposed in simple tasks, suitable task setting has to be selected, rewarding need to be determined, tasks should be simple and intuitive;
- data preparation: the interface and the instruction should be clear, task need to be design in order to prevent and reduce cheating (crowdsourcing platforms offer some functionality that can help in this sense);
- project execution: it is important to attract and retain contributors and to filter cheating workers in order to improve quality (it is possible e.g. to embed gold standard question to determine the general quality of data provided by each worker).

Related works which use crowdsourcing in order to collect information on opposition relation include the contribution of Mohammed et al. (2013). The authors use crowdsourcing to determine the level of human agreement on considering terms in a pair (adjectives, adverbs, nouns and verbs) as contrasting (i.e. word pairs that have some non-zero degree of binary incompatibility and/or have some non-zero difference across a dimension of meaning) or opposites (i.e. word pairs that have a strong binary incompatibility relation with each other and/or are saliently different across a dimension of meaning) and on classifying the pairs into one of the different type of oppositions the authors distinguish (i.e. antipodals, complementaries, disjoint, gradable opposites, reversibles). Specifically, observing the collected data the authors find that annotators agree markedly on identifying contrasting word pairs while there is a variation in the agreement in the questions that aim at identifying the different kinds of opposites. They also verify that contrasting pairs can be classified into more than one kinds.
More recently, Takabatake et al. (2015) use crowdsourcing in order to collect contradictory event pairs and create a large-scale database of Japanese contradictory event pairs by asking the crowd to write and evaluate in-domain contradictory sentences. In their study, the authors classify contradictory event pairs in a taxonomy which includes also "binary event pairs", i.e. events that contradict each other for presenting e.g."mutually exclusive antonyms", such as *single* and *married*.
Other related works include experiments that use crowdsourcing not to directly investigate opposition relations but that are connected to other aspects of our work. For example, in the study of Fossati et al. (2013), authors apply crowdsourcing to perform FrameNet annotation. They compare two approaches: in the first they ask the crowd to select the frame evoked by a given predicate in a sentence and then to identify the frame elements typically involved in the identified frame (2-steps approach); in the second approach they start from the frame elements annotation to identify the frame expressed in a sentence (1-step approach). Among other results, they notice that crowdsourcing can produce usable results for the annotation of frame element. Feizabadi and Padó (2014) also report on a study on the annotation of frame-semantic roles using crowdsourcing, confirming the usefulness of crowdsourcing methodology, provided that the task is carefully designed. For what concerns lexical substitution -a relevant aspect of our experiment, as described in Section 4-, Kremer et al. (2014) adopt crowdsourcing to collect information (i.e. they asked annotators to provide a synonym for a word in a sentence that would not change the meaning of the sentence) and create a "all-words" lexical substitution corpus for English.
Similarly to these works, we expect to acquire reliable information on opposition relations among verb senses at large scale, inexpensively and over a relative short period of time. This information will be possibly used to enrich lexical resources with opposition relations.

4. Acquiring Verb Oppositions

In order to acquire information on opposition relations among verb senses, we have applied a methodology based on three steps. In the first step (*opposition identification*),

[1] Authors also provide a number of examples in which paid-crowdsourcing has been used to create corpora that support a broad range of NLP problems.

Consider the following sentences:
S1: The appeal must be **rejected**.
S2: The appeal must be **approved**.

Task: Consider the two verbs **to reject** and **to approve** in S1 and S2. Which of the following statements are true?
1: A situation in which both events take place simultaneously and with the same participants is possible.
2: Given a situation, if one of the two events does not take place, the other occurs.
3: Given a situation, a "neutral" alternative in which none of the events occurs is possible.
4: It is possible to use gradable modifiers as "a bit", "moderately", "a lot", etc..
5: It is possible that the two events occur repeatedly several times one after the other.
6: One event is possible only if the other has occured.
7: If one event occurs, and then the other takes place, the situation returns to the initial state.
8: The two verbs describe the same situation by two different points of view.

Table 1: An English version of the original type-of-opposition identification step. In the Italian version, each statement was provided with an example.

we want to determine whether there is an opposition relation between a certain sense of a verb (the *source verb*) and another verb (the *target verb*); in the second step (*sense disambiguation*), we focus on understanding which senses of the target verb are involved in the opposition -if previously identified. In the third step (*type-of-opposition identification*), we aim at determining which type of opposition the two verb senses hold.

Specifically, for the *opposition identification* step, we show annotators a pair of sentences, S1 and S2: S1 contains the *source verb* and S2 is identical to S1, with the exception of the *source verb*, which is substituted with the *target verb* in S2. This substitution may generate a S2 sentence that does not make sense. For instance, in Example 2, where "ridare" is the *source verb* and "trattenere" the *target verb*, the relation between the *target verb* and the direct object argument produces a "non sense" sentence.

(2) S1: Posso **ridare** un esame già sostenuto come opzionale?
Eng.: Can I take an exam that I have already taken again?
S2: Posso **trattenere** un esame già sostenuto come opzionale?
Eng.: Can I keep an exam that I have already taken again?

Annotators are asked to compare the two sentences and choose if: (i) S2 makes sense and holds an opposition relation with S1, or (ii) S2 makes sense but it does not hold an opposition relation with S1, or (iii) S2 does not make sense. If a relation of opposition is identified (i), the crowdsourcing task follows with the other two steps, otherwise a new pair S1-S2 is shown.

In the *sense disambiguation* step we ask annotators to disambiguate the sense of the *target verb* in S2, in order to understand which senses of the *target verb* hold the opposition relation with the *source verb*. To perform this disambiguation, we show annotators a list of sentences each one containing the *target verb* in one of its meaning. We ask annotators to mark the sentences in which the *target verb* has the "same" meaning as in S2.

It can be notice that both for the *opposition identification* step and for the *sense identification* step we shown the *source verb* and the *target verb* in context. This is motivated by the fact that we want to collect information on opposition relations at verb senses level; thus showing verbs out of context will not help in clarify which sense of the verbs we are referring to. Also, we think this choice helps in keeping the task more simple and intuitive for annotators than to provide a definition of the senses of the verbs.

These first two steps - *opposition identification* and *sense identification*- have been presented in details in (Feltracco et al., 2015).

In this paper we are discussing the *type-of-opposition identification* step, which consists in gathering information about the type of opposition that the two verbs in the sentences S1 and S2 hold. We also discuss how clustering the human judgments can be used to validate the initial four-classes hypothesis described in Section 2.

4.1. Data Acquisition

In order to acquire information on the type of oppositions, we focused on the types of relation described in Section 2: complementarity, antonymy, converseness and reversiveness. Particularly, we based our experiment on the characteristics of these relations as identified in the literature- Section 2. We represent these characteristics as testing statements (see Table 1) that are given to the annotators. We ask annotators to evaluate which of these statements can be correctly apply to the pairs of verbs they are asked to consider each time. In particular: statement 2 characterizes complementarity, statements 3 and 4 antonymy, statements 5, 6, 7 reversiveness and statements 8 converseness. More in details, the task is defined as follows. Annotators are asked to consider S1 (with the *source verb*) and S2 (with the *target verb*) and to select "which (of the following) statements are true?". Then the list of statements is provided. Statement 1 has been introduced as a control statement as it is meant not to be a valid option in case of opposition relation.

These statements (and in general, each step of the task) were formulated by the authors taking into account that the task was proposed to non expert annotators "in the crowd" with no preparation or instruction on opposition relations and whose background was not known by the authors. Therefore, in order for the task to be more intuitive, we avoid definitions, and provide simple statements with an example. Notice that there is no restriction on the maxi-

mum number of statements the annotator can select. As an example, the crowd annotator #33221216, for the example in Table 1, has selected the options 2 e 7.[2]
The final design of our task was tuned in an off-line pilot experiment we proposed to five annotators of different background, age and experience in the NLP field.

4.2. Data Evaluation

In order to evaluate if there is a correspondence between the four categories of opposition relation described in Section 2 and the judgements by the crowd, we clustered the judgments in four groups and used the resulting clusters as evidence of how humans perceive the four opposition relations. We use K-means, a well known unsupervised algorithm (MacQueen, 1967). This algorithm, given a certain number of clusters (K) and a way to measure the distance between the objects to be clustered, identifies K centroids, and splits the data in K clusters where the mean points are the centroids.
In our case, every annotation judgments is represented by a 8-tuple (an array with eight value) where the i-th coordinate is equal to 1 if the i-th answer is selected, e.g. if the annotator selects the second and the fourth statements the array would be (0,1,0,1,0,0,0,0). To compute the distance between arrays we used their Euclidean distance. We set K = 4 because our hypothesis is that there are 4 different types of opposition. A possible drawback of the algorithm is that it may converge in a local minimum. To avoid this problem, we used several techniques: first we used K-means++ (Arthur and Vassilvitskii, 2007) to initialize our centroids and then we defined an optimization parameter, i.e. the average of the distances between every array in the cluster and the centroid. We run the algorithm 1000 times (an arbitrary number) minimizing the optimization parameter; even if this approach does not guarantee the total optimum, it significantly reduces the probability of a not significant local optimum.

5. Experimental Settings

As previously described, we ask annotators to provide judgments on the types of opposition existing between a *source verb* and a *target verb* by considering two sentences S1 and S2. In particular, S1 is a sentence that contains the *source verb* and is extracted from the annotated corpus of the T-PAS resource, while S2 is created by automatically substituting the *source verb* with an opposite *target verb* (see Table 2 for an example).

[2]An alternative setting of this question would have been to present each statement separately as a yes/no question in order to reduce the amount of information in one question as in Mohammed et al. (2013). We decided to show all the 8 statements as answers of a unique question to enhance the selection of the best choice among some possible answers and reduce the cases of "second thought" of previous answers (i.e. seeing a more adequate option, the annotator would want to change his/her judgments on previous answer. This would require the annotator to go back in the task, find the judgment s/he want to change, change it, return to present page). Also, we want to avoid the annotators to be tempted of answering repeatedly "yes" or repeatedly "no".

S1 contains the *source verb* and is extracted from the T-PAS resource Ex: *source verb* = **to reject** → S1: The appeal must be **rejected**. S2 = S1, but *source verb* is automatically substituted with *target verb* Ex: *target verb* = **to approve** → S2: The appeal must be **approved**. (i.e. a "contrary" of the *source verb*)

Table 2: Description and example for S1 and S2.

This Section will briefly introduce the T-PAS resource from which we extracted S1, will explain the verb selection, sentences extraction and verb substitution process and finally will provide description of the crowsourcing plaftorm setting.

5.1. The T-PAS Resource

We extracted S1 from the annotated corpus of T-PAS resource[3] for the *source verb*.
The T-PAS resource is an inventory of Typed Predicate Argument Structures for Italian manually acquired from corpora following the Corpus Pattern Analysis (CPA) methodology (Hanks, 2004). T-PASs are semantically motivated and are identified through inspection and annotation of actual uses of the analyzed verbs in a corpus of sentences extracted from a a reduced version of the ItWAC corpus (Baroni and Kilgarriff, 2006). An example of T-PAS for the Italian verb *divorare* (Engl. to devour) is given in (3):

(3) T-PAS#2 of the verb *divorare* (Eng. to devour):
[[Human]] divorare [[Document]]

Each T-PAS corresponds to a distinct verb sense, for example in (3) the sense is "read eagerly". After analyzing a sample of 250 concordances of the verb in the corpus, the lexicographer defines each T-PAS recognizing its relevant structure and identifying the Semantic Types (STs) for each argument slots by generalizing over the lexical sets observed in the concordances. For instance, in (3) [[Document]] generalizes over *libro, romanzo, saggio* etc. (Eng. book, newspaper, essay). Then, the lexicographer associates the instances in the corpus to the corresponding T-PAS (see Figure 1). These sentences in the corpus correspond to a list of examples of the particular sense of the verb and are used in our work to extract the S1 sentences.

e lo consiglio a chi ha voglia di	**divorare**	un romanzo, e sottolineo romanzo,
sono chiusa in casa, mangio e studio.	**Divoro**	libri, trascrivo appunti, le mani nei
sfigato "quattrocchi" sempre preso a	**divorare**	romanzi e saggi ormai sia roba da
poi gli avrei reso la cortesia! Mentre	**divoravamo**	libri-game e provavamo tutti i giochi
a chi ancora non lo ha letto, è di non	**divorare**	questo libro in poche ore come

Figure 1: Annotated corpus for T-PAS#2 for the verb *divorare*.

5.2. Task Implementation

We selected the *source verb* for S1 and the *target verb* for S2 according to three conditions: (i) both verbs are present in the T-PAS resource; (ii) both verbs appear in the

[3]tpas.fbk.eu

Dizionario dei Sinonimi e dei Contrari - Rizzoli Editore[4] as lemmas; (iii) the *target verb* is annotated as "contrary" for the *source verb* and viceversa in the Dizionario dei Sinonimi e dei Contrari. The total number of verb pairs extracted according to these criteria is 436. Notice that we have no information about the distribution of the types of opposition among these extracted verb pair.

Since our aim is to acquire information on opposition among verb senses, we implemented the *opposition identification* step for each of the T-PASs of the extracted source verbs (i.e. for T-PAS#1 of the source verb *abbattere*, for T-PAS#2, for T-PAS#3, ..), for a total of 2263 T-PASs.

We extracted up to three sentences for each sense (T-PAS) of the verbs from the T-PAS resource, according to their availability in the resource. We discarded metonymical uses and, to simplify the task, we selected the shortest sentences, composed by at least 5 tokens. These are the S1 sentences.

Finally, we generated S2 from S1 substituting the *source verb* with the *target verb* automatically conjugated accordingly, using the library: italian-nlp-library [5].

5.3. Crowdsourcing Settings

For crowdsourcing we used the Crowdflower platform[6], with the following parameter setting. We initially set the payment to 0.04 USD, then to 0.05 USD for each page and the number of sentence pairs for page to 5. In order to reduce the likelihood of unreliable users to participate in the task, we include in each page a Test Question (TQ): a question for which we already know the answer.[7] If an annotator misses many TQs s/he is not permitted to continue the annotation and his/her judgments are rejected: we set the threshold of this accuracy to 71%. We selected the TQs among the total sentence pairs and we annotated them before lunching the task. We also set parameters in order to have annotators with Italian Language skills.[8]

6. Result and Discussion

In almost a month, we collected a total of 502 judgments by 24 annotators who were considered trusted.[9] In this Section we report the clustering output, we evaluate the agreement between annotators and we discuss the crowdsourcing experience.

[4]http://dizionari.corriere.it/dizionario_sinonimi_contrari/

[5]https://github.com/jacopofar/italian-nlp-library

[6]http://www.crowdflower.com

[7]20% of gold data per task is the recommended amount by the Clowdflower Platform.

[8]The Clowdflower Platform enables the task manager to set a number of parameters to filter workers prior to the task and control its quality.

[9]Not all the sentence pairs created with the methodology described in Section 5.2 have been annotated in this period of time -see (Feltracco et al., 2015). Furthermore, results for the *type-of-opposition identification* step are calculated only for the pairs which collected a minimum of two (out of the three required judgements per pairs) answers "S2 makes sense and holds an opposition relation with S1" in the *opposition identification* step. Later, judgments of two annotators were removed as they were considered unreliable: over 70% of their answers includes the selection of statement 1.

Relation	Statement	cluster1	cluster2	cluster3	cluster4
No Opposition	Statement 1	0	2.16	2.05	0
Complementarity	Statement 2	**100**	0	21.23	**100**
Antonymy	Statement 3	0	**86.58**	**74.66**	**100**
	Statement 4	5.61	17.31	26.03	27.78
Reversiveness	Statement 5	4.67	16.01	**91.78**	0
	Statement 6	6.54	9.09	**68.49**	11.11
	Statement 7	7.48	10.82	**87.67**	5.56
Converseness	Statement 8	0.09	1.73	1.37	0
Total Judgments		107	231	146	18

Table 3: Clustering output. Percentage of times a statement is selected in the judgments of a cluster.

6.1. Clustering Output

The output of the clustering process in Section 4.2 is shown in Table 3 and is the following. Statement 2, related to the complementarity relation, is clearly predominant in one out of the four clusters (Cluster 1, composed by 107 judgments); on the other hand, one of the statements related to the antonymy relation (statement 3) is the center of a second cluster (i.e. Cluster 2 with 231 judgments). A third cluster (Cluster 3 composed by 146 judgments) presents the statements that characterize antonyms (statement 3) and reversives (statement 5, 6, 7) as predominant. Finally, a fourth cluster is created. Analysing the different iterations we performed, we observe that while the first three clusters remain stable, the Cluster 4 tends to include judgments that do not clearly fit in the previous groups and it is always the less populated. In the iteration that we took as our final result (the 1000th times - an arbitrary number), cluster four is constituted by 18 judgments (in which both statement 2 and 3 were marked together). Statement 8 related to converseness is not prevailing in any cluster and is selected in 7 cases.

By comparing the first two clusters, results seem to suggest that annotators recognize a distinction between the complementarity relation and the antonymy relation. In fact, in Cluster 1, statement 2 is selected in all the judgments, while statement 3 is never selected, and statement 4 is selected in 5.6% of the cases. On the contrary, in Cluster 2 where statement 3 is prevalent and statement 4 is selected in 17,3% of the cases, statement 2 is never chosen. According to our definition in Section 2, these two categories are actually very different with respect to the "scalar dimension": while antonyms are gradable, complementaries are not.

A deep analysis of the results in this direction shows that statements 2 and 3 are selected in most of cases (455/502) but they are selected together only in 28 cases. If we include the other statement related to antonymy (statement 4), cases of complementary and antonymy statements overlapping are 39 over 502 judgments. This seems to demonstrate that for the verb sense pairs annotated by the contributors, the two categories are frequently chosen but in general not confused.

Results for the third cluster seems to support the hypothesis that reversiveness, interacting with a temporal relation (a dimension that is not captured by the other opposition relations) is not an exclusive relation, but in some cases coexists with other opposition types, particularly antonymy.

6.2. Annotator's Judgments

The majority of the annotators marked different statements for different verb pairs and their judgments are in general distributed over different clusters.

A further analysis of the collected data, which is required for the annotation of the relations in the T-PAS resource, concerns the observation of annotators' agreement at sense pairs level (e.g. for all the pairs S1-S2 of the T-PAS#7 of the source verb *caricare* and the target verb *scaricare*).[10] Given that in our task opposition relations are associated to statements (e.g. for reversiveness, statements 5, 6, and 7), we calculate this agreement on single statements in each sense pair (e.g. joining together all judgments for statement 2 for the pair *caricare*#7- *scaricare*).[11] Value are in average over 95% for statements 1 and 8 (annotators mainly agree in not selecting these statement, being both marked only in 7 pairs) and not inferior to 56% (statement 3) for the other statements (Figure 2).

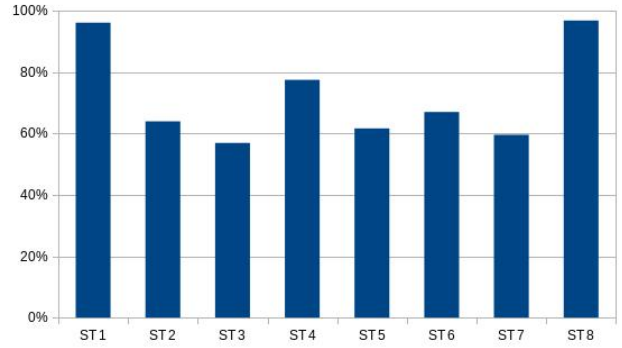

Figure 2: Average interannotator agreement for each statement (ST).

6.3. Discussion on Crowdsourcing Methodology

As regards the crowdsourcing methodology, although the use of examples in place of sense definition simplifies the annotation, the task has been considered rather difficult by many annotators. Furthermore, annotators were not as many as expected (in fact, we expected to collect a greater number of judgments in a less period of time) and most of them were discarded for low accuracy in the initial page which has only TQs. To try to attract more annotators we increased the rewarding but this did not help in rising the number of contributors. On the contrary, we decided not to decrease the threshold of accuracy in the TQs filter in order not to affect quality of the data.

[10]The 502 collected judgments refer to 138 sense pairs in T-PAS.

[11]We first calculate the interannotator observed agreement (IAA) on a single statement in each pair, considering a match when annotators agree both on selecting and not selecting the statement. Then, for each statement separately, we sum the 135 observed agreement (3 pairs have just one judgment, the others being from unreliable annotators - see footnote 8) and we calculate the average.

7. Conclusion

In this paper we have presented a crowdsourcing experiment for the acquisition of opposition relations among verb senses based on the Italian resource T-PAS.

We indirectly collected judgments on four opposition relations -*complementarity, antonymy, conversesness, reversivesness*- by asking annotators in the crowd to consider a pair of verbs in context and mark, among a list of eight statements, which one are valid for that pair of verbs. These statements express single characteristics of the four opposition relations which we assumed from the literature. Clustering annotators judgments, three main groups can be distinguished: one cluster includes judgments where the statement for complementarity is predominant, another one includes judgements in which one of the antonymy statement is prevailing, and a third one includes judgements were both statements for antonymy and reversiveness have been marked by annotators. Collected data allow us to draw interesting conclusions about the categories of oppositions and their relatedness. In fact, results seem to confirm a main distinction between complementarity and antonymy, and suggest that the relation of reversiveness is not an exclusive relation, but it tends to add to other opposition relations, particularly antonymy.

Further work includes the annotation of opposition relations in lexical resources, such as the T-PAS resource. This entails the design of an annotation scheme that considers the results of the crowdsourcing experiment, e.g. should reversiveness be considered as a type or sub-type of opposition? Also, the analysis of judgments at sense pairs level (e.g. for all the pairs S1-S2 of T-PAS#7 of the source verb *caricare* - target verb *scaricare*) is needed in order to tag a relation to a verb pair at sense level.

In addition, the collected judgments can be seen as a first step for the creation of a more complete gold-standard of pairs of opposite sentences in which the types of opposition are identified and annotated.

Moreover, provided the annotation being extended for all the opposite Italian verb pairs, important insights on the distribution of the different types of oppositions will be collected.

For what concerns the use of crowdsourcing, the experiment demonstrates the feasibility of using a methodology based on showing annotators verbs in context in order to collect information on opposition relations among Italian verb senses. Space of improvement is present, as less effort on the acquisition of the data was initially expected (i.e. continuous monitoring was required and more data were expected).

Acknowledgments

We are grateful to Lorenzo Gatti, Manuela Speranza, and Rachele Sprugnoli for their contribution in setting the final design of the task.

8. Bibliographical References

Aguilar, J., Beller, C., McNamee, P., Van Durme, B., Strassel, S., Song, Z., and Ellis, J. (2014). A comparison of the events and relations across ace, ere, tackbp, and framenet annotation standards. In *Proceedings*

of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation, pages 45–53, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Arthur, D. and Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics.

Baroni, M. and Kilgarriff, A. (2006). Large linguistically-processed web corpora for multiple languages. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations*, pages 87–90. Association for Computational Linguistics.

Chklovski, T. and Pantel, P. (2004). Verbocean: Mining the web for fine-grained semantic verb relations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-04)*, volume 2004, pages 33–40, Barcelona, Spain, July.

Cruse, D. A. (1986). *Lexical semantics*. Cambridge University Press.

Cruse, D. A. (2002). Paradigmatic relations of exclusion and opposition ii: Reversivity. *Lexikologie: Ein internationales Handbuch zur Natur und Struktur von Wörtern und Wortschätzen: Lexicology: An international handbook on the nature and structure of words and vocabularies*, 1:507–510.

Cruse, D. A. (2011). *Meaning In Language: An Introduction To Semantics And Pragmatics*. Oxford University Press, USA.

Feizabadi, P. S. and Padó, S. (2014). Crowdsourcing annotation of non-local semantic roles. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 226–230, Gothenburg, Sweden, April. Association for Computational Linguistics.

Fellbaum, C. (1998). *WordNet*. Wiley Online Library.

Feltracco, A., Jezek, E., Magnini, B., and Magnolini, S. (2015). Annotating opposition among verb senses: a crowdsourcing experiment. In *Proceedings of the Second Italian Conference on Computational Linguistics (CLIC-it 2015)*, Trento, Italy, December.

Fokkens, A., Van Erp, M., Vossen, P., Tonelli, S., Van Hage, W. R., SynerScope, B., Serafini, L., Sprugnoli, R., and Hoeksema, J. (2013). Gaf: A grounded annotation framework for events. In *Proceedings of the The 1st Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, volume 2013, pages 11–20, Atlanta, Georgiad, USA, June. Association for Computational Linguistics.

Fossati, M., Giuliano, C., and Tonelli, S. (2013). Outsourcing FrameNet to the crowd. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 742–747, Sofia, Bulgaria, August. Association for Computational Linguistics.

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

Jezek, E. (2016). *The Lexicon. An Introduction.* Oxford: Oxford University Press.

Kremer, G., Erk, K., Padó, S., and Thater, S. (2014). What substitutes tell us - analysis of an "all-words" lexical substitution corpus. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 540–549, Gothenburg, Sweden, April. Association for Computational Linguistics.

Lyons, J. (1977). *Semantics, Vol. I*. Cambridge: Cambridge.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

Mohammad, S. M., Dorr, B. J., Hirst, G., and Turney, P. D. (2013). Computing lexical contrast. *Computational Linguistics*, 39(3):555–590.

Pustejovsky, J., Castaño, J., Ingria, R., Saurí, R., Gaizauskas, R., Setzer, A., Katz, G., and Radev, D. (2003). TimeML: A specification language for temporal and event expressions. In *Proceedings of the International Workshop on Computational Semantics*, page 193.

Sabou, M., Bontcheva, K., Derczynski, L., and Scharl, A. (2014). Corpus annotation through crowdsourcing: Towards best practice guidelines. In *LREC*, pages 859–866.

Takabatake, Y., Morita, H., Kawahara, D., Kurohashi, S., Higashinaka, R., and Matsuo, Y. (2015). Classification and acquisition of contradictory event pairs using crowdsourcing. In *Proceedings of The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 99–107, Denver, Colorado, June. Association for Computational Linguistics.