Crowdsourcing for the identification of event nominals: an experiment

Rachele Sprugnoli[1 2] **and Alessandro Lenci**[3]
[1]Fondazione Bruno Kessler, Via Sommarive 18, 38123 Povo (Italy)
[2]University of Trento, Via Sommarive 5, 38123 Povo (Italy)
[3]Computational Linguistics Laboratory, University of Pisa, via Santa Maria, 36, 56126, Pisa (Italy)
sprugnoli@fbk.eu, alessandro.lenci@ling.unipi.it

Abstract

This paper presents the design and results of a crowdsourcing experiment on the recognition of Italian event nominals. The aim of the experiment was to assess the feasibility of crowdsourcing methods for a complex semantic task such as distinguishing the eventive interpretation of polysemous nominals taking into consideration various types of syntagmatic cues. Details on the theoretical background and on the experiment set up are provided together with the final results in terms of accuracy and inter-annotator agreement. These results are compared with the ones obtained by expert annotators on the same task. The low values in accuracy and Fleiss kappa of the crowdsourcing experiment demonstrate that crowdsourcing is not always optimal for complex linguistic tasks. On the other hand, the use of non-expert contributors allows to understand what are the most ambiguous patterns of polysemy and the most useful syntagmatic cues to be used to identify the eventive reading of nominals.

Keywords: Crowdsourcing, event nominals, corpus annotation

1. Introduction

This paper describes the design and presents the results of a crowdsourcing experiment on the recognition of Italian event nominals. The research question that inspired this experiment is: is it possible to assign the task of the annotation of event nominals within Italian texts to non-experts using crowdsourcing as a promising alternative solution to the employment of well-trained annotators?

This question arose after the analysis of the state of the art in Natural Language Processing (NLP) that shows an increased interest in the automatic analysis of events and temporal information as a fundamental component of a large number of applications such as question answering, information extraction and automatic summarization systems (see, among others, (Saquete et al., 2009; Daniel et al., 2003; Alonso et al., 2010)). In this context, a significant position is occupied by the manual annotation of corpora used to train automatic systems and to evaluate their performances (Pustejovsky et al., 2003a; Bittar et al., 2011; Caselli et al., 2011). However, the annotation of nominals denoting events and thus their identification in a text are not simple tasks due to semantic ambiguities they tend to exhibit. Nominals can show polysemous alternations between eventive and non-eventive readings. The process-result polysemy of deverbal nouns has been widely discussed in the literature from the syntactic (Grimshaw, 1990) and lexico-semantic points of view (Pustejovsky, 1995) but other types of regular polysemy of event nominals are also possible (Apresjan, 1974; Pustejovsky, 2005) making the identification of event denoting nominals challenging. For this reason, their annotation is traditionally carried out by expert annotators with considerable investment in terms of time and costs.

Recently the use of crowdsourcing platforms (such as Amazon Mechanical Turk[1]) to perform various tasks related to NLP has emerged[2]. In particular crowdsourcing has been exploited for the creation of language resources and the annotation of text (Finin et al., 2010; Irvine and Klementiev, 2010; Negri et al., 2012), images (Deng et al., 2009), and speech (Novotney and Callison-Burch, 2010; Sprugnoli et al., 2013).

Through these crowdsourcing platforms, a complex task is segmented into small work units that are distributed among a large pool of non-expert workers, usually via the Web. Many studies (e.g. (Callison-Burch and Dredze, 2010)) have shown that crowdsourcing can reduce time and cost with respect to conventional methods but also that the most critical point of this approach concerns the quality control, i.e. how to collect data with high quality. For example, CrowdFlower[3], a popular provider of crowdsourcing services, has an embedded quality control system based on "gold units", i.e. items for which the correct answer is known, to distinguish between trusted and untrusted contributors. The former are those who provide a correct answer for at least 70% of the gold units whereas the latter are those who fail the gold units and thus are automatically excluded from the task without being paid.

In order to assure the quality of the data collected in the experiment described in this contribution, the gold standard quality assurance mechanism of CrowdFlower has been adopted, multiple judgments from different workers were requested and a comparison with expert annotation on the same task was performed.

The final results demonstrate that the use of crowdsourcing is not always optimal or trivial for complex linguistic tasks. On the other hand, the use of non-expert contributors allowed to understand what are the most ambiguous classes of polysemy and the most useful syntagmatic cues to be

[1]https://www.mturk.com/

[2]A recent survey of crowdsourcing annotations for NLP is given in Wang et al. (2013).

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

used to identify the eventive reading of nominals.

The remainder of the paper is structured as follows: Section 2 provides an overview of related literature in the field of linguistic analysis and computational processing of event nominals. Section 3 describes the setup and the outcome of the experiment highlighting the procedure adopted to create the dataset, the accuracy and inter-annotator agreement results and the comparison with expert annotation on the same task. Conclusions and future perspectives are reported in Section 4.

2. Related Works

The analysis of the linguistic status of event nominals is highly complex given that various types of nominals can denote an event (e.g. deverbal and non-deverbal nouns) (Kiefer, 1998) and that different classes of polysemy can be identified (Grimshaw, 1990). Thus, the first works about the annotation and the automatic processing of events took in consideration only events expressed by verbs or, at most, by nominalizations (Filatova and Hovy, 2001; Katz and Arosio, 2001; Schilder and Habel, 2001). At a later stage, TimeML annotation scheme (Pustejovsky et al., 2003b) has assigned a key role to the recognition of both deverbal and simple nouns event nominals. More recently, several studies have focused on the analysis of the complex semantic concept inherent in event-denoting nominals in order to create lexicons to be used as resources for automatic systems (Bel et al., 2010; Arnulphy, 2011; Russo et al., 2011) or to define an eventivity measure for nouns (Caselli and Russo, 2009). Despite these efforts, the major problem for automatic systems is still the identification of non-verbal events (Kolya et al., 2013; Zavarella and Tanev, 2013).
As for crowdsourcing, Snow et al. (2008) report the first comprehensive survey of the use of non-expert workers for linguistic annotation. Among the experiments presented in the paper, one is inspired by TimeML specifications and concerns event temporal ordering but it takes into consideration only verbal events. Caselli and Chu-Ren (2012) describe a crowdsourcing experiment on the identification and classification of event types in Italian texts; also in this case, only verbs were selected. On the contrary, Alonso et al. (2013) carry out the annotation of regular polysemy of nouns including the dot type PROCESS • RESULT (Pustejovsky, 1995) thus taking into account a typical ambiguity of eventive nominals. The results of their crowdsourcing annotation on English data show the difficulty of the task: in particular, the identification of PROCESS • RESULT alternation obtained the worse results with Krippendorffs alpha coefficient (Krippendorff, 1980) of 0.1[4].

3. Experiment Setup and Results

The experiment focused on the identification of event denoting nominals within Italian sentences taken from newswire texts and the Web. It was built using the services of CrowdFlower and published on the Amazon Mechanical Turk marketplace.

From the theoretical point of view the classification of polysemy for Italian event nominals proposed in (Jezek, 2008) was adopted. In particular, 9 types of polysemy patterns were taken into account:

1. EVENT / STATE (e.g. *abbandono* "state of abandonment");
2. EVENT / ABSTRACT OBJECT (e.g. *accordo* "agreement");
3. EVENT / INFORMATION OBJECT (e.g. *lezione* "lecture");
4. EVENT / PHYSICAL OBJECT (e.g. *fasciatura* "bandage");
5. EVENT / FOOD (e.g. *cena* "dinner");
6. EVENT / MEAN (e.g. *illuminazione* "lighting");
7. EVENT / PERSON (e.g. *fenomeno* "phenomenon");
8. EVENT / HUMAN GROUP or ORGANIZATION (e.g. *balletto* "ballet");
9. EVENT / LOCATION or PATH (e.g. *fermata* "stop").

From the original list, the EVENT / INTERVAL pattern (e.g. *fioritura* "blooming") was removed given that it is a domain-preserving alternation between two temporal objects, thus it should be better described as vagueness than real polysemy.
Six different syntagmatic cues that typically determine the eventive reading of nominals were selected:

1. temporal adjectives: e.g. *recente* "recent";
2. aspectual verbs or nouns: e.g. *fine* "the end";
3. temporal adverbs and prepositions: e.g. *dopo* "after";
4. non aspectual verbs requiring an event as argument: e.g. *avvenire* "to occur";
5. temporal expressions: e.g. *per 12 ore* "for 12 hours";
6. light verbs constructions: e.g. *fare una scelta* "make a choice"

In addition, we decided to add to the dataset some sentences containing a combination of the cues and other sentences in which the eventive meaning of nominals is inferable only from the context (i.e. sentences with no explicit syntagmatic cues).
On the base of the classification of polysemy reported above and of the selected syntagmatic cues, a set of sentences was extracted using Sketch Engine (Kilgarriff et al., 2004) from two corpora: itWaC (Baroni and Kilgarriff, 2006), a corpus of Italian Web pages, and I-CAB (Magnini et al., 2006), a corpus of Italian news stories. At the end, 192 sentences containing 75 different nominals were selected. Thanks to the use of Sketch Engine, the dataset was balanced with the same percentage of sentences containing event and non-event nominals, but also with sentences

[4]Krippendorff (2004) states that tentative conclusions can be drawn only with $\alpha \geq 0.667$

Figure 2: One of the sentences of the experiment taken from the CrowdFlower interface. The English translation of the sentence is the following: "During the **lunch**, a direct and analytic study of what are the techniques and methodologies of living together at meal time is carried out."

taken from both corpora and showing different types of polysemy and syntagmatic cues.

The target polysemous nominal in the sentence was highlighted and 5 judgments were collected from different contributors for each test item. Regional qualifications were applied in order to reduce the risk of spam: more specifically, the geographical location of contributors was limited to Italy. Moreover, the built-in quality control system of CrowdFlower was used to select only reliable contributors. For this purpose, a gold standard of 20 sentences (that is 10% of the data set[5]) was created by an expert annotator.

Finally, particular attention was devoted to the preparation of instructions to be used in the experiment trying to provide workers with simple but complete indications, with many examples. Figure 1 shows the instructions in Italian as presented in the CrowdFlower interface: a concise definition of what is an event is given (something that happens, that can have a beginning and an end, be sudden or prolonged). Workers are suggested to check the presence of a few simple syntagmatic cues saying that, within a sentence, nominals encoding events often occur together with verbs such as "to begin" or "to end", with adjectives such as "frequent" and "recent" and prepositions such as "during" or "before". Then some sentences are listed showing that the same nominal can encode an event or not depending on the context. Finally, five sentences are shown to contributors asking whether the noun highlighted in yellow identifies an event: a yes/no answer is required and a field for comments is available (see Figure 2).

3.1. Results

The reward for judging 5 sentences was set at $0.05: the total task cost was $19.39. The task was completed in 16 days: 9 workers participated in the experiment but only 4 passed the minimum level of reliability required by CrowdFlower calculated on the gold standard sentences.

In Table 1 we report the results of accuracy and inter-annotator agreement (IAA) calculated using the Fleiss' kappa (Fleiss, 1971) on each pattern of polysemy, on only the sentences that contain eventive nominals and on those not containing eventive nominals.

The general accuracy measured after applying majority voting on the judgments of reliable workers was 74%. The different types of polysemy showed a high variability in the accuracy values: the alternations EVENT / ABSTRACT OBJECT and EVENT / FOOD are the most complex in terms of understanding and recognition with an accuracy of 58%, whereas EVENT / HUMAN GROUP clearly recorded the maximum value (92%). Taking into consideration the global distinction between sentences with or without eventive nominals, no relevant difference was registered in the accuracy (76% versus 73% respectively). The accuracy of sentences containing event nominals were further analyzed by taking into account the distinction between classes of syntagmatic cues in order to understand which of them proved to be more useful for the recognition of the eventive meaning of nominals.

[5]CrowdFlower recommends having approximately 10% of the entire dataset flagged as gold.

CLASSES OF POLYSEMY	ACCURACY	IAA
Event/Abstract Object	58%	-0.04%
Event/Food	58%	0.15%
Event/Information Object	67%	0.18%
Event/Physical object	67%	0.30%
Event/Mean	75%	0.57%
Event/Location-Path	79%	0.39%
Event/State	83%	0.31%
Event/Person	83%	0.36%
Event/Human Group	92%	0.55%
Eventive Nominals	76%	0.25%
Non Eventive Nominals	73%	0.24%
TOTAL	**74%**	**0.34%**

Table 1: Accuracy and inter-annotator agreement (Fleiss' kappa) scores.

Table 1 shows that the combination of several cues (e.g. aspectual verb + temporal expression as in *L'**assedio** era iniziato il 18 settembre* "The **siege** began on Sept. 18") led to perfect accuracy after applying majority voting to the judgments. Very high accuracy values (90%) were registered also for sentences containing temporal expressions (e.g. *la prima tappa prevede 8-9 ore di **cammino** sotto un sole cocente* "the first stage includes 8-9 hours of **walking** under a blazing sun"), temporal prepositions and adverbs (e.g. *L'azienda è stata dichiarata fallita dopo l'**arresto** del commercialista* "The company went bankrupt after the **arrest** of accountant") and aspectual verbs and nouns (e.g. *Lei si affrettò a riprendere la **spiegazione*** "She hurried to resume the **explanation**"). On the contrary, determining the eventive reading of a nominal only by the context without the support of any cue led to incorrect judgments (e.g. *istituzioni che garantiscano i diritti umani sono necessarie al **governo** dell'economia globale, sostiene l'organizzazione newyorkese* "institutions that ensure human rights are necessary for the **government** of the global economy, says the New York-based organization").

CUES	ACCURACY
Combination of cues	100%
Temporal expressions	90%
Prepositions and adverbs	90%
Aspectual verbs and nouns	90%
Adjectives	80%
Light Verb Constructions	60%
Non aspectual verbs	60%
NO cues - Only Context	20%

Table 2: Accuracy of the recognition of event nominals on the basis of the presence (or absence) of syntagmatic cues.

The total inter-coder agreement among the contributors is

Figure 1: Instructions as shown in the CrowdFlower interface.

0.34, that is a fair agreement (Landis et al., 1977), but with a high between-class variation. In particular, for the class of polysemy EVENT / ABSTRACT OBJECT the agreement was poor (-0.04) and only the alternations EVENT / HUMAN GROUP and EVENT / MEAN reached a moderate agreement (0.55 and 0.57 respectively).

Out of a total of 192 sentences, 61 (32%) showed complete agreement among the 5 contributors: on these sentences the accuracy was 88%. Among the sentences with complete agreement, 7 were incorrectly judged by all contributors while for 54 sentences the judgments were correct. A completely incorrect agreement was recorded on 6 sentences containing eventive nominals and on 1 containing a non eventive nominal belonging to the EVENT / INFORMATION OBJECT alternation: *Sono alle prese con un tediosissimo **esame** di Economia delle aziende* ("I'm struggling with a really boring **exam** of Business Economics"). As for the 6 sentences having an event-denoting nominal and showing a complete wrong agreement, one doesn't have any cue, two contained light verb constructions (i.e. *fare una **fasciatura*** "make a **bandage**" and *portare **disturbo*** "cause **disturbance**"), two non aspectual verbs requiring an event argument (i.e. *garantire l'**illuminazione*** "ensure **enlightenment**" and *ridurre l'**illuminazione*** "reduce **enlightenment**"), and one an aspectual verb (i.e. *il **pericolo** non è cessato* "the **danger** has not ended").

3.1.1. The Chi-Squared Test

The chi-squared test has been used to measure whether the difference between the accuracy values obtained in the crowdsourcing experiment was statistical significant. In particular, the test was applied to the analysis of 2–by–2 contingency tables to compare the distribution of eventive and non eventive nominals across the dataset. The null hypothesis of the test was that there is no difference between the observed results thus that it is not easier to identify event-denoting nominals than non event-denoting ones. The chi-square test revealed no significant differences between eventive and non eventive nominals ($\chi^2 = 2.06$, $df = 1, p > 0.05$).

Additional chi-squared analyses were conducted in order to detect whether the obtained accuracy values differed significantly between polysemy patterns. The results of these tests, reported in Figure 3, indicated that the difference in the number of accurate judgments was significant only between some pairs of patterns. In particular, all the comparisons between the EVENT / HUMAN GROUP pattern and the other patterns turned out to have significant chi-squared values.

3.2. Expert Annotation

In order to better understand if the low accuracy value recorded in the crowdsourcing experiment was mainly imputable to the inexperience of the contributors, the task was performed also by to two Italian native speakers expert annotators, with proven knowledge of Italian linguistics and previous experiences in the field of semantic annotation.

Figure 4: Percentages of accuracy for each pattern of polysemy achieved by expert annotators on 54 sentences.

The two annotators have been specifically trained on the task: they were given the same instructions used in CrowdFlower together with a copy of Jezek's paper (2008), then they carried out a pilot annotation on a small set of sentences and met to discuss their doubts. Once the training was completed, the annotators performed the task on 54 sentences taken from the dataset used in the crowdsourcing experiment. This sub-set of sentences was created so as to respect the balance between patterns of polysemy and between eventive and non eventive nominals. In particular, this sub-set included the sentences that were the most problematic for non-expert annotators, e.g. those with a completely wrong agreement.

At the end, the task was completed in less than 2 hours recording a high level of accuracy (93%) and agreement (0.81). Anyway, the alternation EVENT / ABSTRACT OBJECT proved to be the most challenging even for experts with an accuracy of 75%, whereas 5 patterns (i.e. EVENT / FOOD, EVENT / HUMAN, EVENT / HUMAN GROUP or ORGANIZATION, EVENT / LOCATION or PATH, and EVENT / MEAN) achieved a perfect accuracy (see Figure 4).

As for agreement, both expert annotators wrongly identified the nominal *storia* ("story") in *Ma ricapitoliamo la **storia*** ("But let's recap the **story**") as an event and not as the informational content of a narrative that can be summarized. In the crowdsourcing experiment, 4 out of 5 contributors gave the same wrong judgment on this sentence.

3.3. Discussion on Ambiguous Cases

An analysis of the sentences that proved to be more challenging for both the expert and non-expert annotators led to highlight some ambiguous cases that probably had a direct effect on the agreement. For example, in the sentence:

- *Dopo aver lanciato questo **avvertimento**, la Bce ribadisce la necessità che tutti i Paesi europei affrontino, con la dovuta rapidità ed energia, i problemi di finanza pubblica* ("After launching this **warning**, the ECB reaffirms the need for all European countries to tackle, with all due speed and energy, public finance problems").

the expression *lanciare un avvertimento* was interpreted by the expert as a light verb construction introduced by an extended light verb (*lanciare*) (Cicalese, 1999) [6] . Following this interpretation, the annotators identified *avvertimento* as

[6] Cicalese (1999) distinguishes between basic or neutral light verbs (e.g. *fare, dare* for Italian) and extended light verbs that add a semantic value to the whole construction. In the present

Figure 3: Results of the Chi-squared test between the accuracy of polysemy patterns: two stars highlight the differences that are statistically significant at $p < 0.01$ while one star indicates those that are statistically significant at $p < 0.05$.

an event-denoting nominal although it was classified as non event-denoting in the original dataset.
A case of co-predication was also detected:

- *Lo rivela un **esperimento**, pubblicato su New Scientist, condotto in Inghilterra che ha dimostrato che le foto di ragazzi e uomini riscuotevano maggior successo da parte delle donne se questi si chiamavano Ed , Elliott o Mike* ("This is revealed by an experiment, published in New Scientist, conducted in England which showed that photos of boys and men achieved resounding success among women if they were called Ed, Elliott or Mike").

The nominal *esperimento* was classified in the dataset as non-event denoting because the verbs *rivelare* "to reveal" and *pubblicare* "to publish" denote the abstract object resulting from the experimenting event. However, the verb *condurre* "to conduct" selects the eventive reading making the interpretation of this sentence not straightforward.

4. Conclusions and future works

This paper investigates the possibility of adopting crowdsourcing methods for the identification of polysemous nominals that show different types of sense alternation between eventive and non eventive readings. The general aim of the experiment was to try to push the boundaries of crowdsourcing applying it to a complex linguistic task.
The accuracy of the results obtained with the crowdsourcing experiment (74%) proved not to be comparable to that obtained by experts on the same task (93%). Data on the inter-coder agreement confirm the problematic nature of this task for non-expert contributors that obtained a kappa of 0.34 whereas experts achieved an agreement of 0.81. In other words, few expert annotators produced better results than many non-expert annotators. These results shows that the recognition of nominal events is not an intuitive task, easily accomplished using only practical instructions made available to non-expert contributors.
From the linguistic point of view, the problems recorded in the recognition of almost all the classes of polysemy and, most of all, of the ones involving abstract and informational objects seem to indicate that these alternations are not entirely well-defined.
As for future works, it would be interesting to take advantage of the "wisdom of the crowd" asking to non-experts to freely highlight in a text all the nominals with an eventive reading following only a personal interpretation: the collected data could be analyzed to find an operational definition of event nominals. Such a task was proposed to students of computational linguistics by Hatzivassiloglou and Filatova (2003) who report that nouns such as war and earthquake had never been identified as events.
Finally, we plan to replicate the task presented in this paper with languages other than Italian to make a comparison of the results in different languages. In this way it would be possible, for example, to analyze which are the most intuitive types of polysemy and the most useful syntagmatic cues across different languages. These differences may be useful to improve the performance of cross-language automatic systems.

example, *lanciare un avvertimento* can be seen as an extension of *dare un avvertimento* with *lanciare* belonging to a more formal register that *dare*.

5. Acknowledgements

This work was partially supported by the European Union's 7th Framework Programme via the NewsReader Project (ICT-316404).

6. References

Alonso, O., Berberich, K., Bedathur, S. J., and Weikum, G. (2010). NEAT: News Exploration Along Time. In *32nd European Conference on IR Research, ECIR*, volume 5993 of *LNCS*, page 667.

Alonso, H. M., Pedersen, B. S., and Bel, N. (2013). Annotation of regular polysemy and underspecification. In *51st Meeting of the Associatation for Computation Linguistics (ACL 2013)*.

Apresjan, J. D. (1974). Regular polysemy. *Linguistics*, 12(142):5–32.

Arnulphy, B. (2011). A weighted lexicon of french event names. In *RANLP Student Research Workshop*, pages 9–16.

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.

Bel, N., Coll, M., and Resnik, G. (2010). Automatic detection of non-deverbal event nouns for quick lexicon production. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 46–52. Association for Computational Linguistics.

Bittar, A., Amsili, P., Denis, P., and Danlos, L. (2011). French timebank: an iso-timeml annotated reference corpus. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 130–134. Association for Computational Linguistics.

Callison-Burch, C. and Dredze, M. (2010). Creating speech and language data with amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 1–12. Association for Computational Linguistics.

Caselli, T. and Chu-Ren, H. (2012). Sourcing the crowd for a few good ones: Event type detection. In *Proceedings of COLING 2012: Posters*, pages 1239–1248.

Caselli, T. and Russo, I. (2009). Fires and blizzards syntagmatic cues for event nouns in italian. In *5th International Conference on Generative Approaches to the Lexicon*.

Caselli, T., Lenzi, V. B., Sprugnoli, R., Pianta, E., and Prodanof, I. (2011). Annotating events, temporal expressions and relations in italian: the it-timeml experience for the ita-timebank. In *Linguistic Annotation Workshop*, pages 143–151.

Cicalese, A. (1999). Le estensioni di verbo supporto. uno studio introduttivo. *Studi italiani di linguistica teorica ed applicata*, 28(3):447–485.

Daniel, N., Radev, D., and Allison, T. (2003). Sub-event based multi-document summarization. In *HLT-NAACL Text summarization workshop*, pages 9–16. ACL.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE.

Filatova, E. and Hovy, E. (2001). Assigning time-stamps to event-clauses. In *Proceedings of the ACL-EACL 2001 Workshop for Temporal and Spatial Information Processing*.

Finin, T., Murnane, W., Karandikar, A., Keller, N., Martineau, J., and Dredze, M. (2010). Annotating named entities in twitter data with crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 80–88. Association for Computational Linguistics.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Grimshaw, J. (1990). *Argument Structure*. MIT Press, Cambridge, Massachusetts.

Hatzivassiloglou, V. and Filatova, E. (2003). Domain-independent detection, extraction, and labeling of atomic events. Proceedings of the RANLP Conference.

Irvine, A. and Klementiev, A. (2010). Using mechanical turk to annotate lexicons for less commonly used languages. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 108–113. Association for Computational Linguistics.

Jezek, E. (2008). Polysemy of italian event ominals. *Nominalisations. Numero special di Faits des Langues*, pages 251–264.

Katz, G. and Arosio, F. (2001). The annotation of temporal information in natural language senteces. In *Proceedings of the ACL-EACL 2001 Workshop for Temporal and Spatial Information Processing*.

Kiefer, F. (1998). Les substantifs déverbaux événementiels. *Langages*, pages 56–63.

Kilgarriff, A., Rychly, P., Smrz, P., and Tugwell, D. (2004). The sketch engine. *Information Technology*, 105:116.

Kolya, A. K., Kundu, A., Gupta, R., Ekbal, A., and Bandyopadhyay, S. (2013). Ju_cse: A crf based approach to annotation of temporal expression, event and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 64–72, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Krippendorff, K. (1980). *Content Analysis: An Introduction to Its Methodology (second edition)*. Sage Publications.

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, 30(3):411–433.

Landis, J. R., Koch, G. G., et al. (1977). The measurement of observer agreement for categorical data. *biometrics*, 33(1):159–174.

Magnini, B., Pianta, E., Girardi, C., Negri, M., Romano, L., Speranza, M., Bartalesi Lenzi, V., and Sprugnoli, R. (2006). I-cab: The italian content annotation bank. In *Proceedings of LREC*, pages 963–968. Citeseer.

Negri, M., Mehdad, Y., Marchetti, A., Giampiccolo, D., and Bentivogli, L. (2012). Chinese whispers: Cooperative paraphrase acquisition. In *LREC*, pages 2659–2665.

Novotney, S. and Callison-Burch, C. (2010). Cheap, fast and good enough: Automatic speech recognition with non-expert transcription. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 207–215. Association for Computational Linguistics.

Pustejovsky, J., Hanks, P., Saurí, R., See, A., Gaizauskas, R., Setzer, A., Radev, D., Sundheim, B., Day, D., Ferro, L., and Lazo, M. (2003a). The TIMEBANK Corpus. In *Corpus Linguistics*, pages 647–656.

Pustejovsky, J., Castaño, J. M., Ingria, R., Saurí, R., Gaizauskas, R., Setzer, A., and Katz, G. (2003b). TimeML: Robust Specification of Event and Temporal Expressions in Text. In *IWCS-5*.

Pustejovsky, J. (1995). *The Generative Lexicon*. MIT Press.

Pustejovsky, J. (2005). A survey of dot objects. manuscipt.

Russo, I., Caselli, T., and Rubino, F. (2011). Recognizing deverbal events in context. In *Proceedings of CICLing*.

Saquete, E., González, J. L. V., Martínez-Barco, P., Muñoz, R., and Llorens, H. (2009). Enhancing QA Systems with Complex Temporal Question Processing Capabilities. *Journal of Artificial Intelligence Reresarch (JAIR)*, 35:775–811.

Schilder, F. and Habel, C. (2001). From temporal expressions to temporal information: Semantic tagging of news messages. In *Proceedings of the ACL-EACL 2001 Workshop for Temporal and Spatial Information Processing*.

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. (2008). Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*, pages 254–263. Association for Computational Linguistics.

Sprugnoli, R., Moretti, G., Fuoli, M., Giuliani, D., Ben-

tivogli, L., Pianta, E., Gretter, R., and Brugnara, F. (2013). Comparing two methods for crowdsourcing speech transcription. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8116–8120. IEEE.

Wang, A., Hoang, C. D. V., and Kan, M.-Y. (2013). Perspectives on crowdsourcing annotations for natural language processing. *Language resources and evaluation*, 47(1):9–31.

Zavarella, V. and Tanev, H. (2013). Fss-timex for tempeval-3: Extracting temporal information from text. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 58–63, Atlanta, Georgia, USA, June. Association for Computational Linguistics.