Raising Interest and Collecting Suggestions on the EVALITA Evaluation Campaign

Rachele Sprugnoli
FBK and University of Trento
Via Sommarive, 38123 Trento, Italy
sprugnoli@fbk.eu

Viviana Patti
University of Turin
c.so Svizzera 185, I-10149 Torino, Italy
patti@di.unito.it

Franco Cutugno
University of Naples Federico II
Via Claudio 21, 80126 Naples, Italy
cutugno@unina.it

Abstract

This paper describes the design and reports the results of two questionnaires. The first of these questionnaires was created to collect information about the interest of industrial companies in the field of Italian text/speech analytics towards the evaluation campaign EVALITA; the second to gather comments and suggestions for the future of the evaluation and of its final workshop from the participants and the organizers of the campaign on the last two editions (2011 and 2014). Novelties introduced in the organization of EVALITA 2016 on the basis of the questionnaires results are also reported.

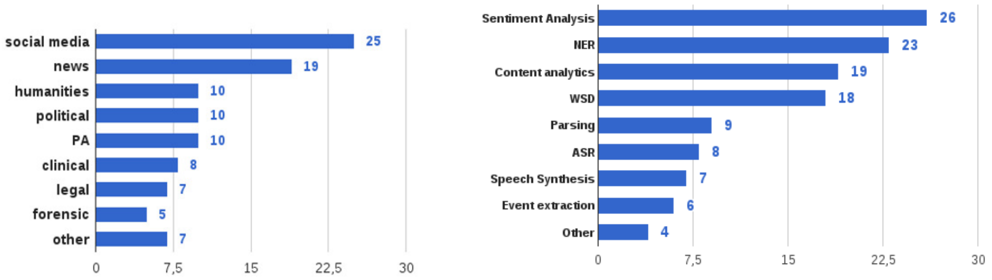
1 Introduction

EVALITA is a periodic evaluation campaign of Natural Language Processing (NLP) and speech tools for the Italian language that has been organized around a set of shared tasks since 2007[1]. Examples of tasks organized in the past EVALITA campaigns are: Named Entities Recognition (NER), Automatic Speech Recognition (ASR), and Sentiment Analysis (Attardi et al., 2015). At the end of the evaluation, a final workshop is organized so to disseminate the results providing participants and organizers with the opportunity to discuss emerging and traditional issues in NLP and Speech technologies for Italian. Over four editions (i.e. 2007, 2009, 2011 and 2014), EVALITA organized more than 30 tasks receiving almost 150 submissions from 90 different organizations: among them 31 (34.4%) were not located in Italy and 10 (11.1%) were not academic. This latter number highlights the limited contribution of enterprises in the campaign, especially in its 2014 edition in which no industrial company was involved as participant. Starting from this observation, in 2015 we designed an online questionnaire to collect information about the interest of industrial companies in the field of text/speech analytics towards EVALITA, with the main aim of understanding how the involvement of companies in the campaign can be fostered.

After four editions we also thought it was time to gather the views of all those who have contributed, until that moment, to the success of EVALITA in order to continuously improve the campaign confirming it as a reference point of the entire NLP and Speech community working on Italian. To this end we prepared another questionnaire for participants and organizers of past EVALITA evaluations to collect comments on the last two editions and receive suggestions for the future of the campaign and of its final workshop.

Questionnaires have been used for different purposes in the NLP community. For example, to carry out user requirements studies before planning long-term investments in the field (Allen and Choukri, 2000; Group, 2010) or in view of the development of a linguistic resource (Oostdijk and Boves, 2006). Moreover, online questionnaires have been adopted to discover trends in the use of a specific technique, e.g. active learning (Tomanek and Olsson, 2009). Similarly to what we propose in this paper, Gonzalo et al. (2002) designed two questionnaires, one for technology developers and one for technology deployers, to acquire suggestions about how to organize Cross-Language Evaluation Forum (CLEF) tasks. As for the feedback received from private companies, the authors report “not very satisfactory results”. On the contrary we registered a good number of responses from enterprises in Italy and abroad.

[1] http://www.evalita.it



(a)

(b)

Figure 1: (a) Domains and (b) tasks of interest for the companies that answered the questionnaire.

2 Questionnaire for Industrial Companies

The EUROMAP analysis, dated back to 2003, detected structural limits in the Italian situation regarding the Human Language Technology (HLT) market (Joscelyne and Lockwood, 2003). Within that study, 26 Italian HLT suppliers are listed: in 2015, at the time of questionnaire development, only 13 of them were still active. The dynamism of the HLT market in Italy is confirmed by a more recent survey where the activities of 35 Italian enterprises are reported (Di Carlo and Paoloni, 2009): only 18 were still operative in 2015.

Starting from the active private companies present in the aforementioned surveys, we created a repository of enterprises working on Italian text and speech technologies in Italy and abroad. In order to find new enterprises not listed in previous surveys, we took advantage of online repositories (e.g. AngelList[2] and CrunchBase[3]) and of extensive searches on the Web. Our final list included 115 companies among which 57 are not based in Italy. This high number of enterprises dealing with Italian also outside the national borders, reinforces one of the findings of the 2014 Alta Plana survey (Grimes, 2014)[4] that provides a detailed analysis of text analytics market thanks to the answers given to a questionnaire dedicated to technology and solution providers. No Italian company took part in that investigation but Italian resulted as the fourth most analyzed language other than English (after Spanish, German and French) and registered an estimated growth of +11% in two years.

All the companies in our repository were directly contacted via email and asked to fill in the online questionnaire. After an introductory description and the privacy statement, the questionnaire was divided into three sections and included 18 questions. In the first section, we collected information about the company such as its size and nationality; the second had the aim of assessing the interest towards evaluation campaigns in general and towards a possible future participation in EVALITA. Finally, in the third section we collected suggestions for the next edition of EVALITA.

We collected responses from 39 private companies (response rate of 33.9%)[5]: 25 based in Italy (especially in north and central regions) and the rest in other 9 countries[6]. 27 companies work on text technologies, 2 on speech technologies and the remaining declares to do business in both sectors. The great majority of companies (84.6%) has less than 50 employees and, more specifically, 43.6% of them are start-up.

Around 80% of respondents thinks that initiatives for the evaluation of NLP and speech tools are useful for companies and expresses the interest in participating in EVALITA in the future. Motivations behind the negative responses to this last point are related to the fact that the participation to a campaign is considered very time-consuming and also a reputation risk in case of bad results. In addition, EVALITA is perceived as too academically oriented, too focused on general (i.e. non application-oriented) tasks

[2] https://angel.co/

[3] https://www.crunchbase.com/

[4] http://altaplana.com/TA2014

[5] This response rate is in line with the rates reported in the literature on surveys distributed through emails, see (Kaplowitz et al., 2004; Baruch and Holtom, 2008) among others, and with the ones reported in the papers cited in Section 1.

[6] Belgium, Finland, France, Netherlands, Russia, Spain, USA, Sweden, and Switzerland.

and with a limited impact on media. This last problem seems to be confirmed by the percentage of respondents who were not aware of the existence of EVALITA before starting the questionnaire, i.e. 38.5% with 24.1% among Italian companies.

For each of the questions regarding the suggestions for the next campaign (third section), we provided a list of pre-defined options, so to speed up the questionnaire completion, together with a open field for optional additional feedback. Participants could select more than one option. First of all we asked what would encourage and what would prevent the company from participating in the next EVALITA campaign. The possibility of using training and test data also for commercial purposes and the presence of tasks related to the domains of interest for the company have been the most voted options followed by the possibility of advertising for the company during the final workshop (for example by means of exhibition stands or round tables) and the anonymisation of the results so avoiding negative effects on the company image. On the contrary, the lack of time and/or funds is seen as the major obstacle.

Favorite domains and tasks for companies participating in the questionnaire are shown in Figure 1. Social media and news resulted to be the most popular among the domains of interest, followed by humanities, politics and public administration. Domains included in the "Other" category are survey analysis, financial but also public transport and information technology. For what concerns the tasks of interest, sentiment analysis and named entity recognition were the top voted tasks, but a significant interest has been expressed also about content analytics and Word Sense Disambiguation (WSD). In the "Other" category, respondents suggested new tasks such as dialogue analysis, social-network analysis, speaker verification and text classification.

3 Questionnaire for EVALITA Participants and Organizers

The questionnaire for participants and organizers of past EVALITA campaigns was divided into 3 parts. In the first part respondents were required to provide general information such as job position and type of affiliation. In the second part we collected comments about the tasks of past editions asking to rate the level of satisfaction related to four dimensions: (i) the clarity of the guidelines,; (ii) the amount of training data; (iii) the data format; and (iv) the adopted evaluation methods. An open field was also available to add supplementary feedback. Finally, the third section aimed at gathering suggestions for the future of EVALITA posing questions on different aspects, e.g. application domains, type of tasks, structure of the final workshop, evaluation methodologies, dissemination of the results.

The link to the questionnaire was sent to 90 persons who participated in or organized a task in at least one of the last two EVALITA editions. After two weeks we received 39 answers (43.3% response rate) from researchers, Phd candidates and technologists belonging to universities (61.54%) but also to public (25.64%) and private (12.82%) research institutes. No answer from former participants affiliated to private companies was received.

Fifteen out of seventeen tasks of the past have been commented. All the four dimensions taken into consideration obtained positive rates of satisfaction: in particular, 81% of respondents declared to be very o somewhat satisfied by the guidelines and 76% by the format of distributed data. A small percentage of unsatisfied responses (about 13%) were registered on the quantity of training data and on the evaluation. In the open field, the most recurring concern was about the low number of participants in some tasks, sometimes just one or two, especially in the speech ones.

Respondents expressed the will to see some of the old tasks proposed again in the next EVALITA campaign: sentiment polarity classification (Basile et al., 2014), parsing (Bosco et al., 2014), frame labeling (Basili et al., 2013), emotion recognition in speech (Origlia and Galatà, 2014), temporal information processing (Caselli et al., 2014), and speaker identity verification (Aversano et al., 2009). As for the domains of interest, the choices made by participants and organizers are in line with the ones made by industrial companies showing a clear preference for social media (27), news (15), and humanities (13).

The diverging stacked bar chart (Heiberger and Robbins, 2014) in Figure 2, shows how the respondents ranked their level of agreement with a set of statements related to the organization of the final workshop, the performed evaluation and the campaign in general. The majority of respondents agree with almost all statements: in particular, there is a strong consensus about having a demo session during the work-

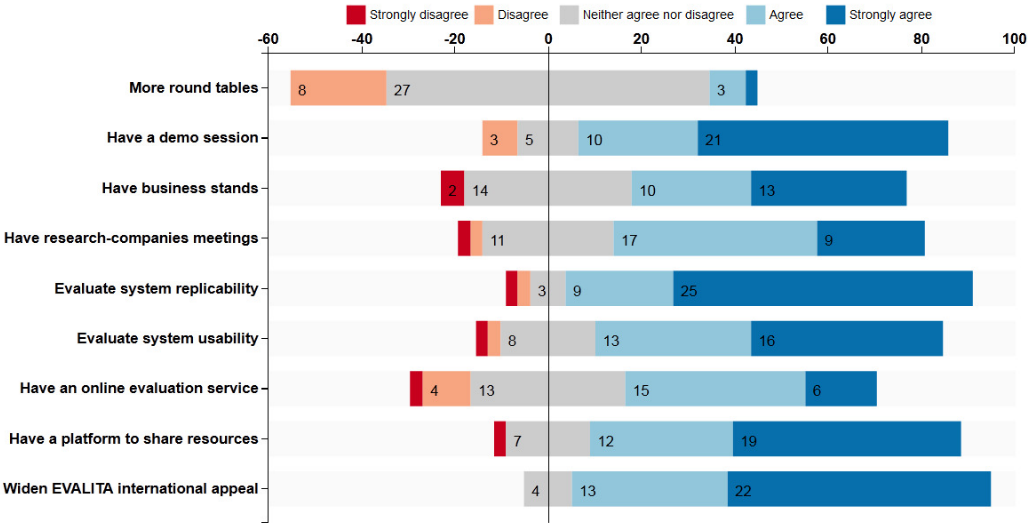


Figure 2: Questionnaire for participants and organizers. Statements assessed on a Likert scale: respondents who agree are on the right side, the ones who disagree on the left and neutral answers are split down the middle. Both percentages and counts are displayed.

shop and also about taking into consideration, during the evaluation, not only systems' effectiveness but also their replicability. Providing the community with a web-based platform to share publicly available resources and systems seems to be another important need as well as enhancing the international visibility of EVALITA. A more neutral, or even negative, feedback was given regarding the possibility of organizing more round tables during the workshop.

4 Lessons Learnt and Impact on EVALITA 2016

Both questionnaires provided us with useful information for the future of EVALITA: they allowed us to acquire input on different aspects of the campaign and also to raise interest towards the initiative engaging two different sectors, the research community and the enterprise community.

Thanks to the questionnaire for industrial companies, we had the possibility to reach and get in touch with a segment of potential participants who weren't aware about the existence of EVALITA or had little knowledge about it. Some of the suggestions coming from enterprises are actually feasible, for example by proposing more application-oriented tasks and by covering domains that are important for them. As for this last point, it is worth noting that the preferred domains are the same for both enterprises and former participants and organizers: this facilitate the design of future tasks based on the collected suggestions. Another issue emerged from both questionnaires is the need of improving the dissemination of EVALITA results in Italy and abroad, in particular outside the boarders of the research community.

The questionnaire for former participants and organizers gave us insights also on practical aspects related to the organization of the final workshop and ideas on how to change the systems evaluation approach taking into consideration different aspects such as replicability and usability.

The results of the questionnaires were presented and discussed during the panel "Raising Interest and Collecting Suggestions on the EVALITA Evaluation Campaign"[7] organized in the context of the second Italian Computational Linguistics Conference[8] (CLiC-it 2015). The panel has sparked an interesting debate on the participation of industrial companies to the campaign, which led to the decision of exploring new avenues for involving industrial stakeholders in EVALITA, as the possibility to call for tasks of industrial interest, that are proposed, and financially supported by the proponent companies. At the same time, the need for a greater internationalization of the campaign, looking for tasks linked to the ones proposed outside Italy, was highlighted. Panelists also wished for an effort in future campaigns towards

[7] http://www.evalita.it/towards2016

[8] https://clic2015.fbk.eu/

the development of shared datasets. Being the manual annotation of data a cost-consuming activity, the monetary contribution of the Italian Association for Computational Linguistics[9] (AILC) was solicited.

The chairs of EVALITA 2016[10] introduced in the organization of the new campaign novel elements, aimed at addressing most of the issues raised by both the questionnaires and the panel (Basile et al., 2016b).

EVALITA 2016 has an application-oriented task (i.e., QA4FAQ) in which representatives of three companies[11] are involved as organizers (Caputo et al., 2016). Another industrial company[12] is part of the organization of another task, i.e., PoSTWITA (Tamburini et al., 2016). Moreover, IBM Italy runs, for the first time in the history of the campaign, a challenge for the development of an app providing monetary awards for the best submissions: the evaluation follows various criteria, not only systems' effectiveness but also other aspects such as intuitiveness and creativity[13]. Given the widespread interest in social media, a particular effort has been put in providing tasks dealing with texts in that domain. Three tasks focus on the processing of tweets (i.e., NEEL-it, PoSTWITA, and SENTIPOLC) and part of the test set is shared among 4 different tasks (i.e., FacTA, NEEL-it, PoSTWITA, and SENTIPOLC) (Minard et al., 2016; Basile et al., 2016a; Barbieri et al., 2016). Part of the SENTIPOLC data was annotated via Crowdflower[14] thanks to funds allocated by AILC.

For what concerns the internationalization issue, in the 2016 edition we had two EVALITA tasks having an explicit link to other shared tasks proposed for English in the context of other evaluation campaigns: the re-run of SENTIPOLC, with an explicit link to the *Sentiment analysis in Twitter* task at SEMEVAL[15], and the new NEEL-it, which is linked to the *Named Entity rEcognition and Linking (NEEL) Challenge* proposed for English tweets at the 6th Making Sense of Microposts Workshop (#Microposts2016, co-located with WWW 2016)[16]. Both tasks have been proposed with the aim to establish a reference evaluation framework in the context of Italian tweets.

We also used social media such as Twitter and Facebook, in order to improve dissemination of information on EVALITA, with the twofold aim to reach a wider audience and to ensure timely communication about various stages of the evaluation campaign.

As for the organization of the final workshop, a demo session is scheduled for the systems participating to the IBM challenge, as a first try to address the request from the community to have new participatory modalities of interacting with systems and teams during the workshop.

Acknowledgments

We are thankful to the panelists and to the audience of the panel 'Raising Interest and Collecting Suggestions on the EVALITA Evaluation campaign' at CLiC-it 2015, for the inspiring and passionate debate. We are also very grateful to Malvina Nissim and Pierpaolo Basile, who accepted with us the challenge to rethink EVALITA and to co-organize the edition 2016 of the evaluation campaign.

References

Jeffrey Allen and Khalid Choukri. 2000. Survey of language engineering needs: a language resources perspective. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000).*

Giuseppe Attardi, Valerio Basile, Cristina Bosco, Tommaso Caselli, Felice Dell'Orletta, Simonetta Montemagni, Viviana Patti, Maria Simi, and Rachele Sprugnoli. 2015. State of the Art Language Technologies for Italian: The EVALITA 2014 Perspective. *Intelligenza Artificiale*, 9(1):43–61.

[9] http://www.ai-lc.it/

[10] They include Malvina Nissim and Pierpaolo Basile, in addition to the authors of this paper.

[11] QuestionCube:http://www.questioncube.com; AQP:www.aqp.it;
SudSistemi: http://www.sudsistemi.eu

[12] CELI: https://www.celi.it/

[13] http://www.evalita.it/2016/tasks/ibm-challenge

[14] https://www.crowdflower.com/

[15] http://alt.qcri.org/semeval2016/task4/

[16] http://microposts2016.seas.upenn.edu/challenge.html

Guido Aversano, Niko Brümmer, and Mauro Falcone. 2009. EVALITA 2009 Speaker Identity Verification Application Track - Organizer's Report. *Proceedings of EVALITA, Reggio Emilia, Italy*.

Francesco Barbieri, Valerio Basile, Danilo Croce, Malvina Nissim, Nicole Novielli, and Viviana Patti. 2016. Overview of the EVALITA 2016 SENTiment POLarity Classification Task. In Pierpaolo Basile, Anna Corazza, Franco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, and Rachele Sprugnoli, editors, *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*. Associazione Italiana di Linguistica Computazionale (AILC).

Yehuda Baruch and Brooks C Holtom. 2008. Survey response rate levels and trends in organizational research. *Human Relations*, 61(8):1139–1160.

Valerio Basile, Andrea Bolioli, Malvina Nissim, Viviana Patti, and Paolo Rosso. 2014. Overview of the Evalita 2014 sentiment polarity classification task. In *Proceedings of the 4th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'14). Pisa, Italy*, pages 50–57. Pisa University Press.

Pierpaolo Basile, Annalina Caputo, Anna Lisa Gentile, and Giuseppe Rizzo. 2016a. Overview of the EVALITA 2016 Named Entity rEcognition and Linking in Italian Tweets (NEEL-IT) Task. In Pierpaolo Basile, Anna Corazza, Franco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, and Rachele Sprugnoli, editors, *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*. Associazione Italiana di Linguistica Computazionale (AILC).

Pierpaolo Basile, Franco Cutugno, Malvina Nissim, Viviana Patti, and Rachele Sprugnoli. 2016b. EVALITA 2016: Overview of the 5th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Associazione Italiana di Linguistica Computazionale (AILC).

Roberto Basili, Diego De Cao, Alessandro Lenci, Alessandro Moschitti, and Giulia Venturi. 2013. Evalita 2011: the frame labeling over Italian texts task. In *Evaluation of Natural Language and Speech Tools for Italian*, pages 195–204. Springer.

Cristina Bosco, Felice Dell'Orletta, Simonetta Montemagni, Manuela Sanguinetti, and Maria Simi. 2014. The EVALITA 2014 dependency parsing task. *Proceedings of EVALITA*.

Annalina Caputo, Marco de Gemmis, Pasquale Lops, Franco Lovecchio, and Vito Manzari. 2016. Overview of the EVALITA 2016 Question Answering for Frequently Asked Questions (QA4FAQ) Task. In Pierpaolo Basile, Anna Corazza, Franco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, and Rachele Sprugnoli, editors, *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*. Associazione Italiana di Linguistica Computazionale (AILC).

Tommaso Caselli, Rachele Sprugnoli, Manuela Speranza, and Monica Monachini. 2014. EVENTI: EValuation of Events and Temporal INformation at Evalita 2014. In *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & and of the Fourth International Workshop EVALITA 2014*, pages 27–34. Pisa University Press.

Andrea Di Carlo and Andrea Paoloni. 2009. *Libro Bianco sul Trattamento Automatico della Lingua.*, volume 1. Fondazione Ugo Bordoni, Roma.

Julio Gonzalo, Felisa Verdejo, Anselmo Peñas, Carol Peters, Khalid Choukri, and Michael Kluck. 2002. Cross Language Evaluation Forum - User Needs: Deliverable 1.1.1. Technical report.

Seth Grimes. 2014. Text analytics 2014: User perspectives on solutions and providers. *Alta Plana*.

FLaReNet Working Group. 2010. Results of the questionnaire on the priorities in the field of language resources. Technical report, Department of Computer Science, Michigan State University, September.

Richard M Heiberger and Naomi B Robbins. 2014. Design of diverging stacked bar charts for likert scales and other applications. *Journal of Statistical Software*, 57(5):1–32.

Andrew Joscelyne and Rose Lockwood. 2003. *Benchmarking HLT progress in Europe.*, volume 1. The EUROMAP Study, Copenhagen.

Michael D Kaplowitz, Timothy D Hadlock, and Ralph Levine. 2004. A comparison of web and mail survey response rates. *Public opinion quarterly*, 68(1):94–101.

Anne-Lyse Minard, Manuela Speranza, and Tommaso Caselli. 2016. The EVALITA 2016 Event Factuality Annotation Task (FactA). In Pierpaolo Basile, Anna Corazza, Franco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, and Rachele Sprugnoli, editors, *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*. Associazione Italiana di Linguistica Computazionale (AILC).

Nelleke Oostdijk and Lou Boves. 2006. User requirements analysis for the design of a reference corpus of written dutch. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation(LREC 2006)*.

Antonio Origlia and Vincenzo Galatà. 2014. EVALITA 2014: Emotion Recognition Task (ERT). In *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & and of the Fourth International Workshop EVALITA 2014*, pages 112–115. Pisa University Press.

Fabio Tamburini, Cristina Bosco, Alessandro Mazzei, and Andrea Bolioli. 2016. Overview of the EVALITA 2016 Part Of Speech on TWitter for ITAlian Task. In Pierpaolo Basile, Anna Corazza, Franco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, and Rachele Sprugnoli, editors, *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*. Associazione Italiana di Linguistica Computazionale (AILC).

Katrin Tomanek and Fredrik Olsson. 2009. A web survey on the use of active learning to support annotation of text data. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 45–48. Association for Computational Linguistics.